RS-SAM: Integrating Multi-Scale Information for Enhanced Remote Sensing Image Segmentation

Enkai Zhang[1], Jingjing Liu[1], Anda Cao[1], Zhen Sun[1], Haofei Zhang[1], *Huiqiong Wang[2], Li Sun[2], and Mingli Song[1]

[1] Zhejiang University, China
[2] Ningbo Innovation Center, Zhejiang University, China
huiqiong_wang@zju.edu.cn

Abstract. The introduction of the Segment Anything Model (SAM) provides a powerful pre-trained model for image segmentation tasks. However, its utilization in remote sensing image segmentation encounters notable challenges. First, SAM is primarily trained on large-scale natural images as a general visual model, which hinders its direct application to remote sensing field. Second, due to the diversity of spatial objects in remote sensing images, the naïve columnar ViT structure of SAM leads to poor segmentation performance. Finally, SAM is designed primarily to distinguish between foreground and background, resulting in a simple structure that struggles with precise semantic segmentation. To address the above issues, we introduce RS-SAM, a prompt-free adaptation of SAM in the realm of remote sensing, with multi-scale ViT backbone. More specifically, we start by crafting an adapter for the SAM encoder to transferring SAM to the domain of remote sensing. Next, we addressed the encoder's limitations by integrating a Multi-scale Neck for capturing objects in different sizes. Finally, to enhance the segmentation results, we propose a Multi-scale Progressive Refinement Module to aggregate multi-scale and low-level features. Through experiments conducted on three public remote sensing datasets, our model outperforms the baseline by 0.8% to 6.2% on the Dice metric, which fully proves the effectiveness of our method.

Keywords: Deep learning · Segment Anything Model · Remote sensing image

1 Introduction

The precise segmentation of remote sensing images, a fundamental and critical task in geospatial analysis and environmental monitoring, is dedicated to identifying and separating surface features from remote sensing data, holding irreplaceable value for fields such as ecological research, resource management, and disaster assessment [27,14,6]. However, the tremendous cost of high-quality labeling in remote sensing makes it difficult to obtain large-scale datasets, presenting one major challenge for training deep models from scratch. Moreover, the characteristics of multi-scale and distribution drift in remote sensing images pose

further challenges of the segmentation model in terms of multi-scale perception and generalizability.

The segmentation models based on CNNs and ViTs [26,10,12,29,28,11,13] are difficult to adapt to the complex remote sensing field due to the lack of sufficient context information and the need for a large amount of training data. Recently, Segment Anything Model (SAM) [9] has been extensively trained on large-scale natural image dataset, showing extremely powerful feature extraction and generalization capabilities. This breakthrough has opened a path for major progress in remote sensing semantic segmentation (RSS). However, SAM is trained predominantly on natural images and relies on manual prompts, which limit its segmentation performance on RSS. Further, as the target objects are of diverse scale, the naïve columnar ViT structure of SAM leads to poor segmentation performance [10]. Finally, SAM is designed only to distinguish foreground from background. Its structure is relatively simple, so it is difficult to obtain refined segmentation results.

Therefore, exploring a more advanced SAM architecture to address the aforementioned challenges in the remote sensing field remains an open problem. In this paper, we strive to propose a prompt-free adaptation of SAM in the realm of remote sensing, named RS-SAM, with multi-scale ViT backbone. Specifically, we design an adapter utilizing LoRA [5,8] (A type of Parameter-Efficient Fine-Tuning (PEFT), the method can reduce training and storage requirements while maintaining high performance through the fine-tuning of a select few parameters) techniques for transferring SAM to the domain of remote sensing. Moreover, inspired by feature pyramid structure [16], we propose a *Multi-scale Neck* integrated into the ViT backbone, for capturing objects in different sizes. Finally, in order to solve the problem that SAM is difficult to predict the fine segmentation results, we propose a *Multi-scale Progressive Refinement Module* (MPRM) to aggregate multi-scale and low-level features. It can optimize the segmentation results and obtain a more fine-grained segmentation result map. The effectiveness of our method is verified by experiments on three publicly available remote sensing datasets. Our model outperforms the baseline by 0.8% to 6.2% on the Dice metric, which fully proves the effectiveness of our method.

Our contributions are summarized as follows:

- We introduce RS-SAM, a novel SAM-based RSS framework, by transferring a pre-trained SAM on natural images to the domain of remote sensing with LoRA adapter.
- We propose a Multi-scale Neck to solve the naïve columnar ViT structure of SAM for capturing objects with multiple sizes and provide guidance for further refined predictions.
- We propose MPRM that aggregates multi-scale image and mask features from lightweight decoder for introducing local details to the image encoder so that we can obtain a more fine-grained segmentation result map.
- Extensive experiments demonstrate that our method achieves superior performance on multiple remote sensing datasets compared to common segmentation models.

2 Relate Work

Osco et al. [18] evaluate the ability of the SAM model in RSS and explore the adaptability of the model to this specific application field. R.I. Sultan et al. [24] propose GeoSAM, which adopts a fine-tuning strategy of dense visual cues from zero-shot learning and sparse visual cues from pre-trained CNN segmentation models, and makes significant progress in segmenting roads and pedestrian infrastructure in geographic images using the base model. Chen et al. [3] develop anchor-based cues and query-based cues for instance segmentation based on satellite images. Extensive experimental results on the WHU Building, NWPU VHR-10, and SSDD datasets validate the effectiveness of the proposed method. Zhang et al. [31] propose UV-SAM, which first generates hybrid cues for urban villages using a small semantic segmentation model, including masks, bounding boxes, and image representations, which are then fed into SAM for fine-grained boundary recognition. Extensive experimental results on two Chinese datasets show that UV-SAM outperforms existing baselines. Ma et al. [17] propose a simple and general framework that fully utilizes the raw output of SAM as well as a general remote sensing image semantic segmentation model, and develops an auxiliary optimization strategy using two loss functions, object consistency loss and boundary preservation loss, revealing the potential of large models such as SAM in remote sensing. Zhang et al. [30] eliminate the need for manual intervention to provide cues. In the multi-head attention block of the SAM encoder part, a set of complementary scaling modules are proposed. In addition, Adapter-Feature is inserted between ViT blocks. These Modules are designed to combine high-frequency image information and image embedding features to generate image-informed cues. Qi et al. [19] propose an automatic prompt learning method that leverages guided masks to generate coarse pixel-level prompts for SAM. Extensive experiments on the DLRSD dataset highlight the superiority of the method over other available few-shot methods.

In previous research on SAM in the context of RSS, scholars typically overlook certain performance limitations resulting from SAM's inherent structure. Specifically, the encoder structure of SAM lacks multi-scale information and possesses a simplistic overall design, rendering it hard to capture detailed information in RSS. We present a modified version of SAM, which incorporates aggregated multi-scale information and enhances fine segmentation capabilities.

3 Proposed Method

3.1 RS-SAM Overview

As shown in Figure 1, our method is constructed utilizing the SAM architecture, which incorporates a LoRA-adapted encoder for transferring from the domain of natural images to the domain of remote sensing images, a Multi-scale Neck for capturing objects of various scales, and a Multi-scale Progressive Refinement Module (MPRM) for generating a more fine-grained segmentation map. We have provided pseudocode to clearly describe the architecture in algorithm 1.

Algorithm 1 RS-SAM Forward Pass

Encoder
$XF, F_{early} \leftarrow$ SAM.Image_Encoder(Image)
$F_{1/4}, F_{1/8}, F_{1/16}, F_{1/32} \leftarrow$ Multi-scale_Neck(XF)
Decoder
$M_{1/32} \leftarrow$ Decoder_32($F_{1/32}$, mask_tokens)
$M_{1/16} \leftarrow$ Decoder_16($F_{1/16}, M_{1/32}$)
$M_{1/8} \leftarrow$ Decoder_8($F_{1/8}, M_{1/16}$)
$M_{1/4} \leftarrow$ Decoder_4($F_{1/4}, M_{1/8}$)
MPRM
$D_{\text{fusion}} \leftarrow \text{MFF}(M_{1/32}, M_{1/16}, M_{1/8}, M_{1/4})$
$F_{\text{fusion}} \leftarrow \text{MHFF}(F_{1/32}, F_{1/16}, F_{1/8}, F_{1/4}, F_{early})$
Predict
Pred $\leftarrow F_{\text{fusion}} \times D_{\text{fusion}}$

3.2 LoRA-adapted Image Encoder

The LoRA (Low-Rank Adaptation) fine-tuning technique is a parameter-efficient approach designed to enhance the effectiveness of large pre-trained models on specific tasks. Following the method proposed by Zhang et al. [8], we use the original encoder of SAM and freeze all the layers to retain the pre-learned knowledge. The projection matrices of queries and values in the multi-head attention layers are added by learnable low rank matrices for adapting to the domain of remote sensing, as shown in Figure 2. Given input queries or values $X_{\text{in}}^{\{Q,V\}}$, the output is $X_{\text{out}} = X_{\text{in}}(\text{SG}(W) + W_d W_u)$, where operator $\text{SG}(\cdot)$ represents stopping the gradient.

3.3 Multi-scale Neck

Multi-scale features are crucial for RSS, as the images usually cover a wide geographical area containing objects at different scales. Using multi-scale features, the model can effectively capture spatial details from micro to macro, thereby distinguishing different land cover types and land object boundaries more accurately. Therefore, we utilize the last layer of the ViT backbone and generate multi-scale feature maps inspired by the feature pyramid structure [15] for object detection.

Figure 3 illustrates the multi-scale feature neck which uses lightweight up sampling layers,transposed convolution, and max pooling layers for generating features with different resolution. This streamlined yet powerful approach ensures both effective feature enhancement and dimensionality reduction, delivering robust performance with simplicity.

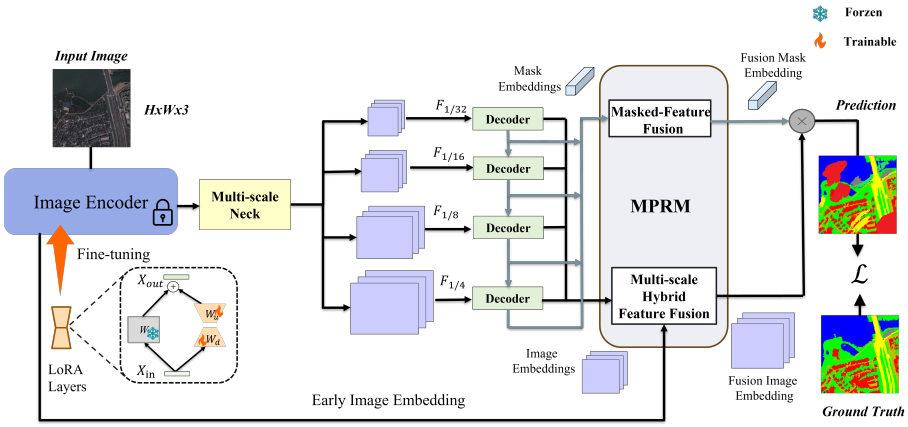

Fig. 1: The overall pipeline of our proposed RS-SAM. The input image starts from the LoRA-adapted image encoder and then passes through the Multi-scale Neck to generate multi-scale embeddings. Each scale embedding generates an enhanced image and mask embedding through the corresponding decoder, where each decoder also receives the mask embedding from the previous scale. Further refinement is handled by MPRM, which integrates multi-scale embeddings and early image embedding to generate a refined segmentation map. The black arrows represent the data flow of image embeddings, and the blue arrows represent the data flow of mask embeddings.

Given the image feature from the last layer $X_F = \text{Encoder}(I)$, the formulas of the Multi-scale Neck are computed as follows:

$$\begin{cases} F_{1/4} &= \text{Neck} \circ U_2 \circ U_1(X_F) \\ F_{1/8} &= \text{Neck} \circ U_3(X_F) \\ F_{1/16} &= \text{Neck}(X_F) \\ F_{1/32} &= \text{Neck} \circ D(X_F) \end{cases}, \tag{1}$$

where $U(\cdot)$ is the upsampling operation, $D(\cdot)$ is the downsampling operation, and $\text{Neck}(\cdot)$ represents a sequence of convolutional layers and normalization layers to perform certain processing to improve the performance and accuracy of segmentation.

3.4 Decoder

As the Multi-scale Neck generates features of different resolutions, a series of SAM decoder are utilized to generate segmentation results at the corresponding resolution. Each decoder consists of two Transformer modules, including Token-to-Image Attention, MLP, and Image-to-Token Attention. The formulas are as

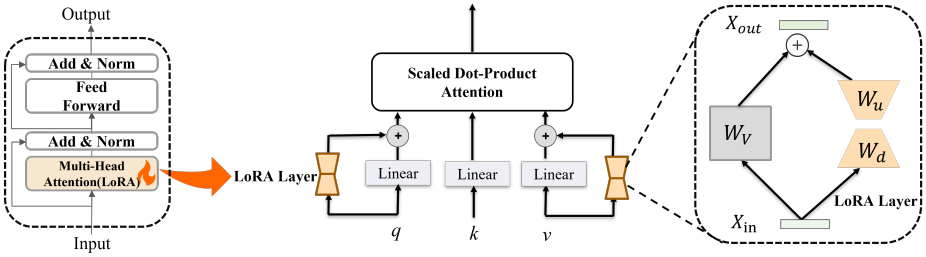

Fig. 2: LoRA-adapted image encoder.

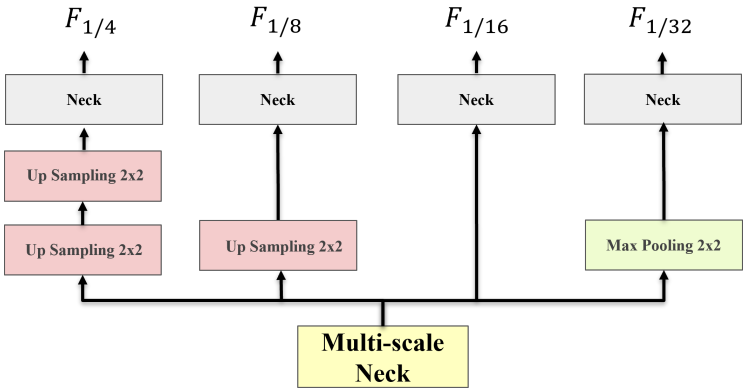

Fig. 3: Multi-scale Feature Neck Module. Using the last layer of features in ViT, we generate feature maps at four scales: 1/32, 1/16, 1/8, and 1/4 to enhance the spatial details in remote sensing features.

follows:

$$\widehat{F}_i, M_i = \text{Decoder}(F_i, M_{i-1}), \tag{2}$$

where F_i denotes the feature map from the correspond neck, M_{i-1} denotes the mask embedding from the previous scale. $\widehat{F}_i$ and M_i denote the image embedding and mask embedding updated through the self-attention and cross-attention.

3.5 Multi-scale Progressive Refinement Module

In order to mine the rich multi-scale semantic features obtained by the decoders at a more fine-grained level and solve the problem that the simple structure of SAM makes it difficult to obtain fine segmentation results based on semantic information in RSS. We propose a Multi-scale Progressive Refinement Module (MPRM), which contains Multi-scale Hybrid Feature Fusion (MHFF) and a Mask Feature Fusion (MFF) to solve the above problems. The prediction result

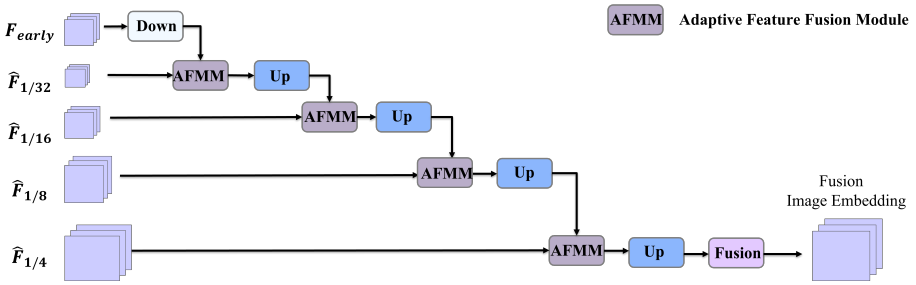

Fig. 4: This module aims to fuse multi-scale feature information. Through this module, more fine-grained fusion features containing rich semantic information can be obtained. The Up, Down and Fusion Modules are composed of convolution, normalization and activation functions.

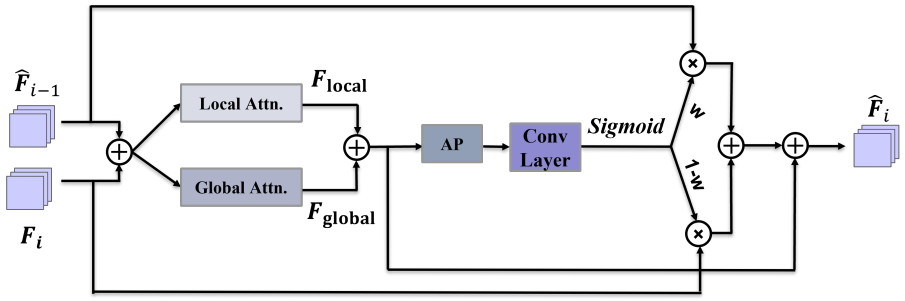

Fig. 5: The structure of the AFFM.

is expressed as:

$$\begin{cases} \text{Prediction} = F_{\text{fusion}} \times D_{\text{fusion}}, \\ F_{\text{fusion}} \quad = \text{MHFF}(F_{\text{early}}, \widehat{F}_{1/32}, \widehat{F}_{1/16}, \widehat{F}_{1/8}, \widehat{F}_{1/4}) \ , \\ D_{\text{fusion}} \quad = \text{MFF}(M_{1/32}, M_{1/16}, M_{1/8}, M_{1/4}) \end{cases} \tag{3}$$

where F_{early} represents early image embedding. F_{fusion} and D_{fusion} represents fusion image embedding and fusion mask embedding.

Multi-scale Hybrid Feature Fusion As shown in Figure 4, this module is used to adaptively integrate multi-scale image embeddings for ViT block, progressively optimize segmentation results. The module consists of the Adaptive Feature Fusion Module (AFFM), Up Module, Down Module, and Fusion Module. As shown in Figure 5, AFFM aggregates image embeddings at two scales. First, the up module scales up lower-resolution features to match higher-resolution features, then refined by local and global attention modules, which can extract and optimize features with broader contextual information. Finally, the enhanced

features are passed through the spatial attention module to generate the attention weight matrix (w_c) of two scale features, multiply with the corresponding features and finally fuse to get the rich features from the fusion of two scales.

This model can retain detail information across various scales, thereby improving the richness and robustness of feature expression.

The formulas are as follows:

$$\begin{cases} F_{\text{local}} &= \text{Attn}_{\text{local}}(\widehat{F}'_{i-1} + \widehat{F}_i) \\ F_{\text{global}} &= \text{Attn}_{\text{global}}(\widehat{F}'_{i-1} + \widehat{F}_i) \\ w_c &= \text{Sigmoid} \circ \text{ConvLayer} \circ \text{AP}(F_{\text{local}} + F_{\text{global}}) \\ \widehat{F}'_i &= \widehat{F}'_{i-1} \cdot w_c + \widehat{F}_i \cdot (1 - w_c) + F_{\text{local}} + F_{\text{global}} \end{cases}. \tag{4}$$

Notably, the two features are activated by the combination of w_c and $1 - w_c$, so that the model can effectively balance the importance of features at different scales, avoid the excessive dominance of a single feature, and ensure the diversity and comprehensiveness of the fused features.

Mask Feature Fusion This module is designed to effectively fuse mask features. It achieves efficient and lightweight information integration through a simple MLP layer which consists of multiple fully connected layers and ReLU activation functions. It enhances feature expression and discrimination by integrating multi-scale information and nonlinear transformations. This module also light and effective enough. The formulas are as follows:

$$f(x) = \text{ReLU} \circ \text{MLP}(x), \tag{5}$$

$$D_{\text{fusion}} = f^{(n)} \left(\sum_i M_i \right). \tag{6}$$

3.6 Training Loss

The training loss combines the pixel-level classification loss and binary mask loss for each segment prediction:

$$\mathcal{L} = \lambda_{\text{ce}} \mathcal{L}_{\text{ce}} + \lambda_{\text{dice}} \mathcal{L}_{\text{dice}}, \tag{7}$$

The pixel-level classification losses $\mathcal{L}_{\text{ce}}$ (binary cross entropy loss) and $\mathcal{L}_{\text{dice}}$ (dice loss) handle pixel classification and mask overlap, respectively. $\mathcal{L}_{\text{ce}}$ ensures detail accuracy, while $\mathcal{L}_{\text{dice}}$ addresses data imbalance and overall segmentation integrity. Weight coefficients λ_{ce} and λ_{dice} adjust the contributions of each loss to the total loss, balancing pixel-level classification and global mask accuracy to optimize model performance.

Table 1: Performance comparison on the DeepGlobe Dataset.

Model	Dice↑	Jaccard↑	Precision↑	Recall↑	Acc↑
FCN [23]	0.649	0.480	0.649	0.650	0.971
UNet [20]	0.705	0.544	0.679	0.733	0.978
TransUNet [2]	0.681	0.516	0.724	0.643	0.975
MANet [12]	0.698	0.536	0.761	0.645	0.977
HRNet [25]	0.713	0.555	0.760	0.672	0.978
UNetFormer [29]	0.736	0.582	**0.778**	0.697	0.979
SegNet [1]	0.707	0.546	0.710	0.704	0.976
SAM Adapter [20]	0.734	0.580	0.732	0.736	0.978
RSAM-Seg [30]	0.732	0.578	0.733	0.732	0.978
Baseline [8]	0.703	0.542	0.767	0.649	0.977
RS-SAM	**0.765**	**0.620**	0.765	**0.766**	**0.981**

4 Experiments

We conducted experiments on three representative public remote sensing datasets. Common segmentation evaluation metrics, including Dice, Jaccard, Precision, Recall, and Accuracy, were used for the assessment.

4.1 Dataset

Deepglobe The dataset [4] contains 6226 RGB images with a resolution of 1024×1024 and segmentation labels, covering images captured in Thailand, Indonesia, and India. The satellite images mainly cover areas containing roads. Due to hardware capacity limitations, we resize the 1024×1024 images to 512×512, and use 70% of the images as training set, 15% of the images as test set, and another 15% of the images as validation set.

WHDLD The WHDLD dataset [21,22] is an important remote sensing image dataset for high-resolution land cover classification and analysis. It is a high-density labeled dataset that can be used for multi-label tasks. The pixels of each image are annotated with the following 6 category labels, namely buildings, roads, sidewalks, vegetation, bare soil and water. The resolution is 256×256, with a total of 4940 images. 70% of the images are used as training set, 15% of the images are used as test set, and the remaining 15% of the images are used as validation set.

Aerial imagery dataset This dataset [7] published by Wuhan University, includes aerial imagery of Christchurch with an original ground resolution of 0.075 meters. It contains manually edited data for approximately 22,000 individual buildings. The dataset is divided into three parts: a training set (130,500 buildings), a validation set (14,500 buildings), and a test set (42,000 buildings). The

Table 2: Performance comparison on the Aerial Imagery Dataset.

Model	Dice↑	Jaccard↑	Precision↑	Recall↑	Acc↑
FCN [23]	0.781	0.641	0.713	0.864	0.946
UNet [20]	0.738	0.585	0.618	0.916	0.928
TransUNet [2]	0.714	0.555	0.597	0.886	0.921
MANet [12]	0.765	0.620	0.655	0.921	0.937
HRNet [25]	0.786	0.647	0.687	0.917	0.944
UNetFormer [29]	0.707	0.547	0.575	0.920	0.915
SegNet [1]	0.747	0.596	0.650	0.877	0.934
SAM Adapter [20]	0.929	0.868	0.919	**0.940**	0.984
RSAM-Seg [30]	0.913	0.841	0.905	0.922	0.981
Baseline [8]	0.908	0.831	0.915	0.900	0.980
RS-SAM	**0.937**	**0.881**	**0.938**	0.936	**0.986**

Table 3: Performance comparison on the WHDLD Dataset.

Model	Dice↑	Jaccard↑	Precision↑	Recall↑	Acc↑
FCN [23]	0.580	0.420	0.633	0.559	0.891
UNet [20]	0.590	0.433	0.594	0.596	0.897
TransUNet [2]	0.589	0.436	0.611	0.576	0.901
MANet [12]	0.586	0.426	0.617	0.574	0.893
HRNet [25]	0.602	0.447	0.623	0.589	0.901
UNetFormer [29]	0.617	0.458	0.628	0.622	0.901
SegNet [1]	0.531	0.381	0.551	0.522	0.886
SAM Adapter [20]	0.704	0.569	**0.737**	0.699	0.936
RSAM-Seg [30]	0.733	0.602	0.732	0.737	0.943
Baseline [8]	0.728	0.597	0.729	0.730	0.942
RS-SAM	**0.736**	**0.607**	0.730	**0.744**	**0.945**

aerial images have been down-sampled to a 0.3-meter ground resolution and cropped into 8189 tiles of 512×512 pixels for ease of use.

4.2 Implementation details

All our implementations are in PyTorch, and we train all our settings on 4 NVIDIA RTX A6000 GPUs. During training, we only use normalization as a means of data augmentation. The training loss is a combination of cross entropy loss and dice loss, with λ_{ce} of 0.2 and λ_{dice} of 0.8. The maximum training epoch is set to 200. We use AdamW optimizer for parameter updates and use the warm-up strategy. To ensure fairness, the model with the lowest loss on the validation set is finally selected as the best model. SAM-based experiments use lightweight Backbone ViT-B. The Baseline method chosen is SAM LoRA [8].

Table 4: Performance comparison on the WHDLD dataset, showing the Dice coefficient for each category.

Model	Building↑	Road↑	Pavement↑	Vegetation↑	Bare Soil↑	Water↑
FCN [23]	0.561	0.637	0.448	0.753	0.392	0.690
UNet [20]	0.588	0.605	0.472	0.780	0.368	0.729
TransUNet [2]	0.594	0.564	0.465	0.781	0.344	0.786
MANet [12]	0.591	0.633	0.479	0.764	0.365	0.684
HRNet [25]	0.606	0.623	0.460	0.779	0.368	0.776
UNetFormer [29]	0.633	0.635	0.488	0.790	0.428	0.725
SegNet [1]	0.553	0.516	0.392	0.742	0.252	0.731
SAM Adapter [20]	0.661	0.706	0.581	0.865	0.464	0.947
RSAM-Seg [30]	0.705	**0.723**	**0.596**	0.880	0.530	0.964
Baseline [8]	0.713	0.716	0.583	0.875	0.520	0.961
RS-SAM	**0.732**	**0.723**	0.583	**0.884**	**0.531**	**0.965**

Table 5: Performance comparison of different Modules on the DeepGlobe Dataset.

Variants			Metric	
Multi-scale Neck	MFF	MHFF	Dice	Jaccard
×	×	×	0.703	0.542
✓	×	×	0.756	0.607
✓	✓	×	0.759	0.611
✓	×	✓	0.757	0.610
✓	✓	✓	0.765	0.620

Quantitative Comparisons In Table 1, Table 2, Table 3 and Table 4, RS-SAM exhibits outstanding performance across three remote sensing datasets. Taking Dice and Jaccard indices as examples, benefit from multi-scale information and advanced optimization methods. RS-SAM significantly outperforms other SAM variants such as SAM Adapter [20], Baseline [8], and RSAM-Seg [30]. On the DeepGlobe dataset, RS-SAM's Dice index is 3.1% to 6.2% higher, and its Jaccard index is 4.0% to 7.8% higher. For the Aerial Imagery dataset, RS-SAM's Dice index is 0.9% to 2.8% higher, and its Jaccard index is 1.3% to 5.0% higher. On the WHDLD dataset, RS-SAM's Dice index is 0.3% to 3.2% higher, and its Jaccard index is 0.5% to 3.8% higher. These results clearly demonstrate the superiority of our model. Additionally, We present fair comparisons with advanced remote sensing domain segmentation methods such as UNetFormer [29], MANet [12], TransUnet [2], UNet [20], HRNet [25], SegNet [1], and FCN [23], showing that RS-SAM achieves significant performance improvements in RSS, greatly enhancing segmentation capabilities and outperforming other methods.

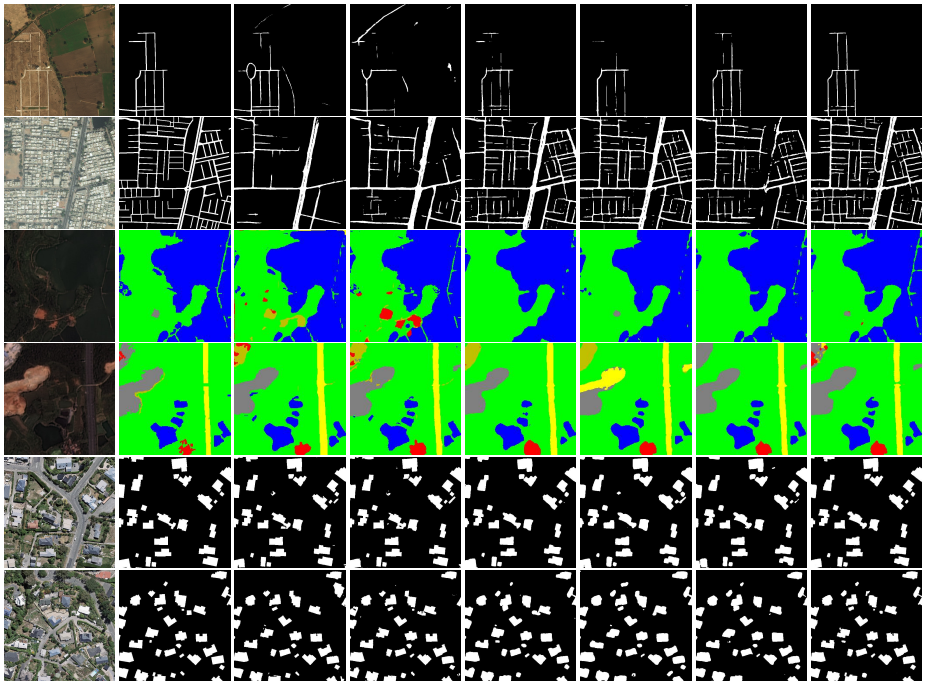

Fig. 6: Performance comparison on three datasets. Rows 1-2: DeepGlobe dataset; Rows 3-4: WHDLD dataset; Rows 5-6: Aerial Imagery dataset.

Qualitative Comparisons To demonstrate the superiority of our method more intuitively, we show a visual comparison of different methods in Figure 6. The visualization provides a clearer advantage of our method compared to previous methods. Our method generates better segmentation masks.

4.3 Analysis and Ablation Studies

In order to evaluate the contribution of different Modules in our proposed method, we conduct experiments on the Deepglobe dataset to verify the effectiveness of each module in our paper.

Table 6: Comparing Multi-Scale Feature Generation Methods.

Generating Multi-scale Features	**Dice↑**	**Jaccard↑**
layers 3, 6, 9, and 12	0.758	0.610
last layer	0.765	0.620

Table 7: Performance comparison of Image Encoders of different sizes on the DeepGlobe Dataset.

Model	Params(M)	Dice↑	Jaccard↑	Precision↑	Recall↑	Acc↑
Baseline(ViT-B) [8]	90.36	0.703	0.542	0.767	0.649	0.977
RS-SAM(ViT-B)	107.4	0.765	0.620	0.765	0.766	0.981
Baseline(ViT-L) [8]	308.12	0.761	0.615	0.754	0.769	0.980
RS-SAM(ViT-L)	324.13	0.781	0.640	0.798	0.763	0.982

Effects of Adapters in Multi-scale Neck In Table 5, we show the impact of the Multi-scale Neck on SAM. Specifically, the Dice coefficient has increased from 0.703 to 0.765, and the Jaccard coefficient has increased from 0.542. to 0.620. These results show that generat multi-scale information in SAM is very necessary and powerful in RSS. In the Table 6, we also proved that the multi-scale features generated by using only the feature map of the last layer of ViT Block in SAM have better performance than the multi-scale features of $1/4$, $1/8$, $1/16$, and $1/32$ generated by 3, 6, 9, and 12 layers respectively. This shows that using only the feature map with rich information in the last layer to generate multi-scale features is superior in SAM. This is a very effective module to solve the problem that SAM does not have multi-scale information.

Effects of Adapters in Multi-scale Progressive Refinement Module In Table 5, we show the impact of the MPRM on SAM. Using only the Multi-scale Neck, the Dice coefficient reaches 0.756 and the Jaccard coefficient is 0.607. Introducing the MFF increases the Dice to 0.759 and the Jaccard to 0.611. With only the MHFF, the Dice increases to 0.757 and the Jaccard to 0.610. Each Module independently enhances detail processing. When combined, the model significantly improve performance, achieving the Dice of 0.765 and the Jaccard of 0.620. This demonstrates their complementarity in capturing detailed and global information, optimize the segmentation result and solve the structural limitations of SAM, substantially enhancing the model's segmentation performance in remote sensing tasks.

Qualitative Results As show in Figure 7, we visualize the results of the proposed module. The fourth column shows the segmentation results using the Multi-scale Neck. It can be seen that compared with Baseline, the segmentation effect has been significantly improved, objects of different sizes can be segmented well (marked in yellow circles) , which fully proves the effectiveness of our Module. When the MPRM is further used, the areas where errors occurred when only the Multi-scale Neck was used (marked in red circles) have been greatly improved. This shows that the MPRM can further improve segmentation accuracy and solve some local problems in the original method.

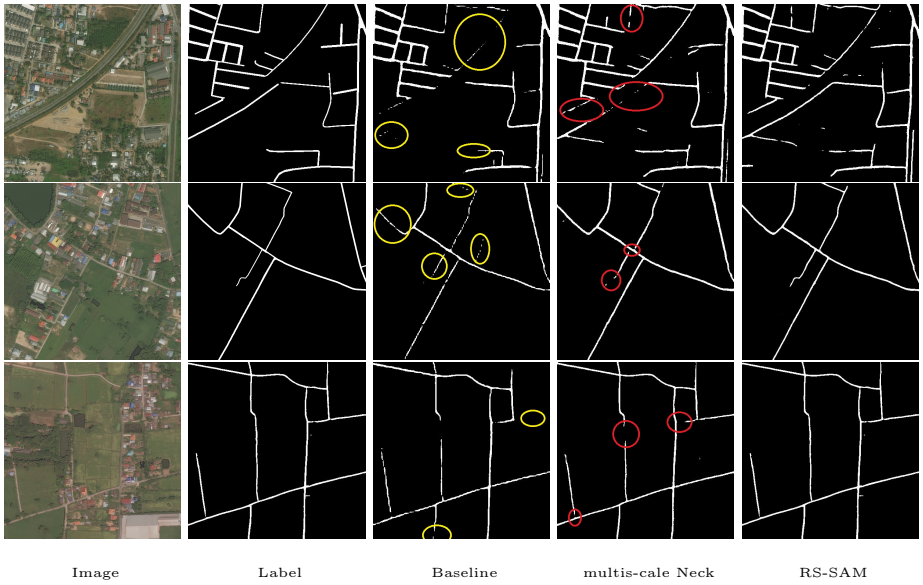

Fig. 7: Visualization of ablation results. The yellow circles indicate that SAM without multi-scale information predicts overly large and small objects and performs poorly. The red circle indicates insufficient segmentation granularity, necessitating optimization by the MPRM.

Effects of Different Image Encoder Sizes Table 7 shows the performance of different-sized Image Encoders on the DeepGlobe dataset. Notably, RS-SAM (ViT-B) outperforms Baseline (ViT-L), although the latter has a much larger number of parameters than the former, achieving a Dice of 0.765 versus 0.761 and Jaccard of 0.620 versus 0.615. This demonstrates the effectiveness of the RS-SAM modifications, as it delivers superior performance with a smaller encoder.

5 Conclusion

This paper proposes RS-SAM, a new framework for remote sensing image segmentation based on the Segment Anything Model (SAM). We start by crafting an adapter utilizing LoRA for the SAM encoder to transferring SAM to the domain of remote sensing. Next, we address the encoder's limitations by integrating a Multi-scale Neck for capturing objects in different sizes. Finally, to solve the problem of SAM's difficult in predicting fine segmentation, we propose a Multi-scale Progressive Refinement Module to aggregate multi-scale and low-level features. Extensive experiments demonstrate that our method achieves superior performance on three remote sensing datasets compared to common segmentation models.

Acknowledgments This work is supported by Zhejiang Province "Pioneering Soldier" and "Leading Goose" R&D Project (2023C01027).

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation (12 2017). https://doi.org/10.17863/CAM.17966
2. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, Alan, L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
3. Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z., Shi, Z.: Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–17 (2024). https://doi.org/10.1109/TGRS.2024.3356074
4. Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raskar, R.: 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)
5. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
6. Huang, L., Jiang, B., Lv, S., Liu, Y., Fu, Y.: Deep-learning-based semantic segmentation of remote sensing images: A survey. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **PP**, 1–28 (01 2023). https://doi.org/10.1109/JSTARS.2023.3335891
7. Ji, S., Wei, S., Lu, M.: Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Transactions on Geoscience and Remote Sensing **57**(1), 574–586 (2019). https://doi.org/10.1109/TGRS.2018.2858817
8. Kaidong, Z., Dong, L.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023)
9. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (October 2023)
10. Kolesnikov, A., Dosovitskiy, A., Weissenborn, D., Heigold, G., Uszkoreit, J., Beyer, L., Minderer, M., Dehghani, M., Houlsby, N., Gelly, S.: An image is worth 16x16 words: Transformers for image recognition at scale (2021)
11. Li, R., Wang, L., Zhang, C., Duan, C., Zheng, S.: A2-fpn for semantic segmentation of fine-resolution remotely sensed images. International Journal of Remote Sensing **43**(3), 1131–1155 (2022). https://doi.org/10.1080/01431161.2022.2030071, https://doi.org/10.1080/01431161.2022.2030071
12. Li, R., Zheng, S., Zhang, C., Duan, C., Su, J., Wang, L., Atkinson, P.M.: Multiattention network for semantic segmentation of fine-resolution remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–13 (2022). https://doi.org/10.1109/TGRS.2021.3093977
13. Li, R., Zheng, S., Zhang, C., Duan, C., Wang, L., Atkinson, P.M.: Abcnet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. ISPRS Journal of Photogrammetry and Remote Sensing **181**, 84–98 (2021). https://doi.org/https://doi.org/10.1016/j.isprsjprs.2021.09.005, https://www.sciencedirect.com/science/article/pii/S0924271621002379

14. Li, S.: A review of remote sensing image classification techniques: the role of spatio-contextual information. European Journal of Remote Sensing **47**, 389–411 (06 2014). https://doi.org/10.5721/EuJRS20144723
15. Li, Y., Mao, H., Girshick, R., He, K.: Exploring Plain Vision Transformer Backbones for Object Detection, pp. 280–296 (11 2022). https://doi.org/10.1007/978-3-031-20077-9_17
16. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 936–944 (2017). https://doi.org/10.1109/CVPR.2017.106
17. Ma, X., Wu, Q., Zhao, X., Zhang, X., Pun, M.O., Huang, B.: Sam-assisted remote sensing imagery semantic segmentation with object and boundary constraints (2023)
18. Osco, L., Wu, Q., Lemos, E., Gonçalves, W., Ramos, A.P., Li, J., Junior, J.: The segment anything model (sam) for remote sensing applications: From zero to one shot (06 2023)
19. Qi, X., Wu, Y., Mao, Y., Zhang, W., Zhang, Y.: Self-guided few-shot semantic segmentation for remote sensing imagery based on large vision models (2023)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. vol. 9351, pp. 234–241 (10 2015). https://doi.org/10.1007/978-3-319-24574-4_28
21. Shao, Z., Yang, K., Zhou, W.: Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset. Remote Sensing **10**(6) (2018). https://doi.org/10.3390/rs10060964, https://www.mdpi.com/2072-4292/10/6/964
22. Shao, Z., Zhou, W., Deng, X., Zhang, M., Cheng, Q.: Multilabel remote sensing image retrieval based on fully convolutional network. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **13**, 318–328 (2020), https://api.semanticscholar.org/CorpusID:211208779
23. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(4), 640–651 (2017). https://doi.org/10.1109/TPAMI.2016.2572683
24. Sultan, R.I., Li, C., Zhu, H., Khanduri, P., Brocanelli, M., Zhu, D.: Geosam: Fine-tuning sam with sparse and dense visual prompting for automated segmentation of mobility infrastructure (2024)
25. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR (2019)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
27. Wang, B., Zhao, Y., Chen, C.L.P.: Moving cast shadows segmentation using illumination invariant feature. IEEE Transactions on Multimedia **22**(9), 2221–2233 (2020). https://doi.org/10.1109/TMM.2019.2954752
28. Wang, L., Li, R., Duan, C., Zhang, C., Meng, X., Fang, S.: A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2022). https://doi.org/10.1109/LGRS.2022.3143368

29. Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P.M.: Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS Journal of Photogrammetry and Remote Sensing **190**, 196–214 (2022). https://doi.org/https://doi.org/10.1016/j.isprsjprs.2022.06.008, https://www.sciencedirect.com/science/article/pii/S0924271622001654
30. Zhang, J., Yang, X., Jiang, R., Shao, W., Zhang, L.: Rsam-seg: A sam-based approach with prior knowledge integration for remote sensing image semantic segmentation. ArXiv **abs/2402.19004** (2024), https://api.semanticscholar.org/CorpusID:268063200
31. Zhang, X., Liu, Y., Lin, Y., Liao, Q., Li, Y.: Uv-sam: Adapting segment anything model for urban village identification. Proceedings of the AAAI Conference on Artificial Intelligence **38**(20), 22520–22528 (Mar 2024). https://doi.org/10.1609/aaai.v38i20.30260, https://ojs.aaai.org/index.php/AAAI/article/view/30260