remote sensing

MDPI

Technical Note

RSPS-SAM: A Remote Sensing Image Panoptic Segmentation Method Based on SAM

Zhuoran Liu [1], Zizhen Li [2], Ying Liang [2,*], Claudio Persello [3], Bo Sun [4], Guangjun He [2] and Lei Ma [5]

[1] School of Geodesy and Geomatics, Beijing University of Civil Engineering and Architecture, Beijing 102627, China; 2022060201l0@stu.bucea.edu.cn
[2] State Key Laboratory of Space-Earth Integrated Information Technology, Beijing Institute of Satellite Information Engineering, Beijing 100095, China; lizizhen@spacestar.com.cn (Z.L.); hegj@spacestar.com.cn (G.H.)
[3] Department of Earth Observation Science, Faculty ITC, University of Twente, 7500AE Enschede, The Netherlands; c.persello@utwente.nl
[4] College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China
[5] School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China; maleinju@nju.edu.cn
* Correspondence: liangying@spacestar.com.cn

Abstract: Satellite remote sensing images contain complex and diverse ground object information and the images exhibit spatial multi-scale characteristics, making the panoptic segmentation of satellite remote sensing images a highly challenging task. Due to the lack of large-scale annotated datasets for panoramic segmentation, existing methods still suffer from weak model generalization capabilities. To mitigate this issue, this paper leverages the advantages of the Segment Anything Model (SAM), which can segment any object in remote sensing images without requiring any annotations and proposes a high-resolution remote sensing image panoptic segmentation method called Remote Sensing Panoptic Segmentation SAM (RSPS-SAM). Firstly, to address the problem of global information loss caused by cropping large remote sensing images for training, a Batch Attention Pyramid was designed to extract multi-scale features from remote sensing images and capture long-range contextual information between cropped patches, thereby enhancing the semantic understanding of remote sensing images. Secondly, we constructed a Mask Decoder to address the limitation of SAM requiring manual input prompts and its inability to output category information. This decoder utilized mask-based attention for mask segmentation, enabling automatic prompt generation and category prediction of segmented objects. Finally, the effectiveness of the proposed method was validated on the high-resolution remote sensing image airport scene dataset RSAPS-ASD. The results demonstrate that the proposed method achieves segmentation and recognition of foreground instances and background regions in high-resolution remote sensing images without the need for prompt input, while providing smooth segmentation boundaries with a panoptic segmentation quality (PQ) of 57.2, outperforming current mainstream methods.

Keywords: panoptic segmentation; segment anything model; remote sensing; deep learning

check for updates

Citation: Liu, Z.; Li, Z.; Liang, Y.; Persello, C.; Sun, B.; He, G.; Ma, L. RSPS-SAM: A Remote Sensing Image Panoptic Segmentation Method Based on SAM. *Remote Sens.* **2024**, *16*, 4002. https://doi.org/10.3390/rs16214002

Academic Editor: Andrea Garzelli

Received: 28 August 2024
Revised: 15 October 2024
Accepted: 24 October 2024
Published: 28 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the improvement in spatial resolution of remote sensing images, the images carry more detailed ground object information and richer hierarchical structures, making the differences between targets and backgrounds more pronounced [1], thereby enabling the possibility of the panoptic segmentation of remote sensing images [2]. The goal of remote sensing image panoptic segmentation is to simultaneously segment countable foreground instances, known as "things" (such as airplanes, ships, and vehicles), and uncountable background regions, known as "stuff" (such as impervious surfaces, bare land, and water). Since panoptic segmentation focuses on both foreground targets and background

environments, it can achieve a more effective understanding and recognition of the overall image compared with tasks such as object detection and semantic segmentation. It holds significant value in various applications including land use change monitoring [3–5], crop planting area extraction [6], disaster early warning assessment [7], and ground object state recognition [8].

Remote sensing images differ significantly from natural images in terms of imaging perspectives and image content, making it challenging to directly apply panoptic segmentation datasets and models designed for natural images to the remote sensing domain. To address this issue, existing research on remote sensing image panoptic segmentation typically focuses on both data samples and methods. For instance, Carvalho et al. [9] extended panoptic segmentation as an extension of semantic segmentation. In this way, they designed software that generates COCO-format annotations for panoptic, instance, and semantic segmentation from GIS tools tailored to the characteristics of remote sensing images. They also introduced the BSB Aerial Dataset, which is suitable for semantic, instance, and panoptic segmentation. Garnot et al. [3] proposed an end-to-end method for panoptic segmentation of the remote sensing image time series, utilizing a temporal self-attention mechanism to extract rich and adaptive multi-scale spatiotemporal features. A dataset for panoptic segmentation of the remote sensing image time series named PASIST was constructed. Zhao et al. [10] introduced a multi-task joint interpretation dataset that integrates image captioning, semantic segmentation, instance segmentation, and panoptic segmentation. They also proposed a joint-optimized Panoptic Perception Model that simultaneously achieves panoptic segmentation and text description through a shared backbone network and joint loss design. Hua et al. [11] developed a cascaded panoptic segmentation network that incorporates hierarchical feature fusion and mask calibration modules to jointly extract contextual information and boundary features. Khoshboresh-Masouleh et al. [12] designed a panoptic building change detection method using uncertainty estimation and squeeze-and-attention CNNs, achieving more accurate change detection than traditional methods with panoptic segmentation. Fernando et al. [13] proposed a multispectral teacher network based on a cross-spectral attention fusion strategy to improve segmentation accuracy by leveraging multimodal data. Šarić et al. [14] introduced a real-time panoptic segmentation network based on pyramid fusion and boundary awareness to efficiently extract multi-scale information from remote sensing images.

Despite these advancements, the current remote sensing image panoptic segmentation methods still face challenges such as inaccurate extraction of object boundaries and weak model generalization capabilities, due to limitations in the scale of annotated samples and the blurring of object boundary information caused by convolutional operations.

In the field of computer vision, foundation models are defined as those pre-trained on extensive datasets with the aim of capturing universal features and structures within the data. These models typically employ self-supervised or supervised learning methods, undergoing large-scale training processes to achieve generalization capabilities across various downstream tasks. The core advantage of foundation models lies in their ability to provide a robust starting point for specific applications, which can be effectively adapted to new tasks through fine-tuning or adaptive training, significantly reducing the reliance on large amounts of annotated data. However, most foundation models such as CLIP [15], Florence [16], and BEiT [17], while supporting multiple tasks including detection, segmentation, and image captioning, do not achieve the precision of dedicated models in segmentation tasks.

SAM (Segment Anything Model) [18], open-sourced by Meta in 2023, is a foundation model specialized in image segmentation. It can segment any object in any image without the need for any annotations and supports zero-shot generalization, enabling the segmentation of unfamiliar objects and images. This characteristic has made SAM applicable to various domains including remote sensing image processing [19], medical image analysis [20], and autonomous driving [21].

In the remote sensing domain, the exceptional generalization performance and flexibility of SAM have led to its widespread application in tasks such as annotation sample generation and various segmentation tasks. Wang et al. [22] addressed the high cost of remote sensing image annotation by proposing a method that combines SAM with remote sensing object detection data to generate large-scale datasets for semantic and instance segmentation. Similarly, leveraging its powerful image segmentation capabilities, some studies have applied SAM in unsupervised and weakly supervised learning to improve the quality of training data annotations. Chen et al. [23] utilized SAM's robust localization and segmentation performance to develop an unsupervised shadow segmentation model called ShadowSAM. Qiao et al. [24] proposed a weakly supervised training method based on SAM that generates pseudo-fully supervised labels through weak annotations.

Due to the differences between remote sensing images and natural images, SAM faces challenges in remote sensing image segmentation, such as the diversity of ground objects and complex backgrounds. Osco et al. [25] and Ji et al. [26] conducted experiments to analyze the advantages and issues of applying SAM directly to the remote sensing field. Although SAM can achieve good segmentation results for targets with regular boundaries, such as buildings, it performs poorly in segmenting small and weakly defined targets and struggles with low-resolution images. To address these issues, some studies have optimized and improved SAM to better adapt the model to the complexity of remote sensing images and enhance its segmentation performance.

Yan et al. [27] came up with a solution to the problem of SAM not working well in multimodal remote sensing scenes, especially in SAR images. They called their model RingMo-SAM and it is based on multimodal remote sensing image segmentation. This model can segment various objects in both optical and SAR data and identify their categories. Similarly, Chen et al. [28] designed an automatic prompt generation method for remote sensing instance segmentation based on SAM, addressing the issue of SAM's inability to recognize categories and its subpar performance on remote sensing images. For panoptic segmentation of the remote sensing images, Nguyen et al. [29] used a target detection network to make bounding boxes. This allowed them to use panoptic segmentation on crop plants or leaves in fields. It was also shown that SAM is very good at both segmenting and generalizing, which can help with problems like multi-scale issues and unclear boundaries in remote sensing image panoptic segmentation tasks.

Based on the above analysis, the task of panoramic segmentation for remote sensing images still faces the following challenges: on the one hand, there are not many public datasets for the panoramic segmentation of remote sensing images, and the scale of the annotated data is small. This means that existing methods for panoramic segmentation of remote sensing images are not very good at generalization. However, the large widths of remote sensing images often necessitate cropping into smaller image patches for training and inference, resulting in a significant loss of contextual information from the original images.

To make panoramic segmentation for remote sensing images more accurate and to cut down on the need for large amounts of annotated data, this paper suggests RSPS-SAM, an end-to-end remote sensing image panoptic segmentation method based on SAM. Firstly, we designed a Batch Attention Pyramid to address the challenges of scale variations in remote sensing images and the loss of target information due to cropping in large remote sensing images during training. This module calculates batch attention for features at various scales and achieves multi-scale feature fusion through information exchange, thereby fully extracting multi-scale information from remote sensing images and long-range contextual information between cropped patches. Secondly, we designed a Mask Decoder to address the challenges of SAM's prompt input requirement and its inability to determine object categories. This module utilized multi-scale features to compute mask attention, outputting mask prompts and category information for the masks, thereby enabling automatic prompt generation and category determination for segmentation. Finally, the original SAM decoder

used image embeddings and mask prompts to output mask segmentation results for each target, achieving panoptic segmentation of remote sensing images.

The main contributions are as follows:

- An end-to-end remote sensing image panoptic segmentation method based on SAM, RSPS-SAM, was proposed. By designing a Mask Decoder, the method achieved automatic generation of mask prompts and determination of mask category information.
- A Batch Attention Pyramid was designed to extract multi-scale information and long-range contextual information from remote sensing images through multi-scale batch attention calculation and multi-scale feature fusion.
- Experimental results demonstrated that the proposed RSPS-SAM achieved a panoptic segmentation quality (PQ) of 57.2 on high-resolution remote sensing images.

2. Proposed Methods

Panoptic segmentation, formally introduced by Kirillov et al. [2] in 2019, is an image segmentation task aimed at simultaneously segmenting the foreground instances and background regions within an image. Figure 1 illustrates an example of panoptic segmentation applied to a remote sensing image.

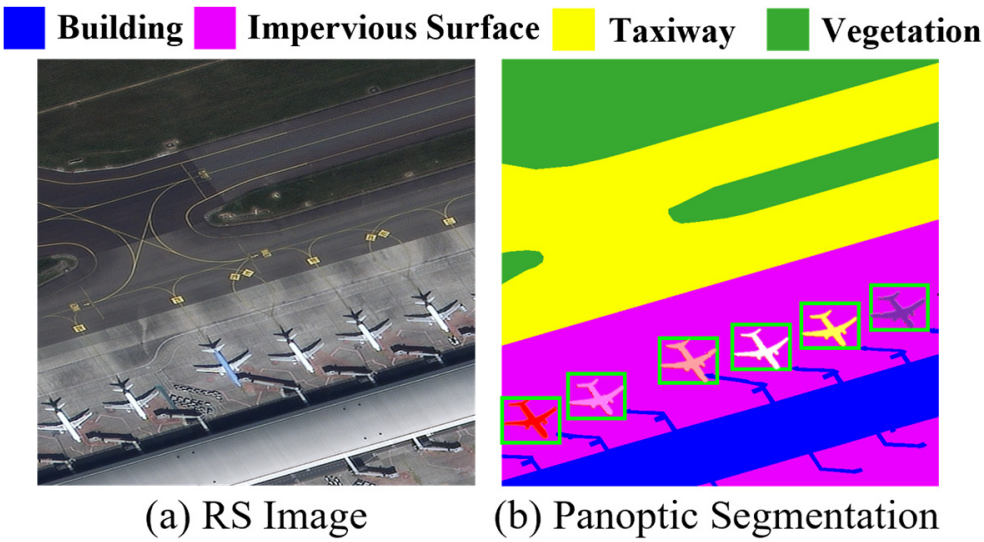


Figure 1. Example of remote sensing image panoptic segmentation: (**a**) original remote sensing image; (**b**) corresponding panoptic segmentation image. To illustrate the requirement of panoptic segmentation to delineate instances of foreground objects, the figure uses distinctly different colors to represent individual airplane instances and outlines them with bounding boxes.

The RSPS-SAM method proposed in this paper for the task of remote sensing image panoptic segmentation is illustrated in Figure 2. Based on SAM, the RSPS-SAM combines a Batch Attention Pyramid to extract multi-scale features and cropped patch features from remote sensing images. Additionally, a Mask Decoder is designed to automatically generate mask prompts and determine the categories of the segmentation results.

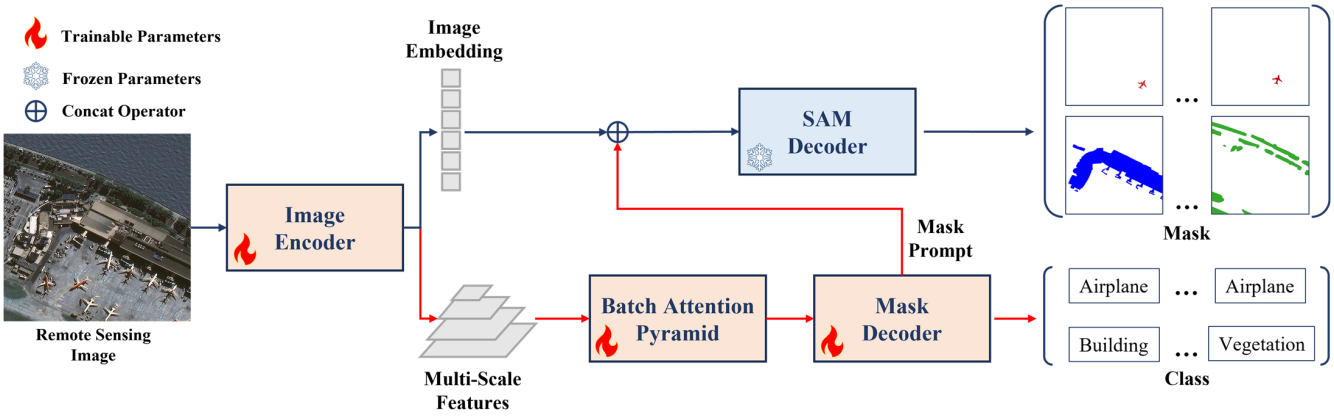


Figure 2. The structure of RSPS-SAM. A Batch Attention Pyramid and a Mask Decoder are added to SAM, with the improvements marked by red lines. These enhancements include the addition of a Batch Attention Pyramid and a Mask Decoder. During the training process, SAM's pre-training weight is adopted, and the parameters of the SAM Decoder are frozen so that they will not change. Other network parts participate in the training normally, and different parameter updating methods are marked with different colors and icons in the figure.

2.1. Network Architecture

The original SAM, as a promptable segmentation method, is composed of an image encoder, a prompt encoder, and a SAM decoder. Trained on a large dataset called SA-1B, it can perform image segmentation tasks based on four types of prompts: points, boxes, text, and masks. The image encoder is a Vision Transformer (ViT) pre-trained using the MAE method, which converts the input image into an image embedding. The prompt encoder processes input points, boxes, text, and masks using positional encoding, CLIP encoding, and convolutional embedding, respectively. The SAM decoder computes prompt self-attention and cross-attention to map the image embedding and prompt encoding to a predicted mask.

Although SAM exhibits strong segmentation performance, there are challenges when applying it directly to remote sensing image panoptic segmentation. Firstly, the category of objects is crucial for distinguishing between foreground and background in panoptic segmentation and plays a significant role in practical remote sensing applications. However, SAM's output lacks category information. Secondly, the imaging angle results in significant scale variations among ground objects in remote sensing images, and SAM's handling of multi-scale issues is suboptimal.

To address these issues, this paper modifies the original SAM by removing the prompt encoder and introducing a new branch for feature extraction, mask prompt generation, and category prediction. Firstly, a Batch Attention Pyramid was designed, as described in Section 2.2, which utilized multi-scale feature maps to fuse multi-scale features and computed batch attention between image patches at each level of the feature maps. Secondly, a Mask Decoder was designed, as detailed in Section 2.3, which primarily generated mask prompts and corresponding category predictions by computing mask attention [30]. Finally, the original SAM decoder was used to map the image embedding and mask prompts to the mask prediction results.

This modified architecture, RSPS-SAM, enhanced the model's ability to handle the complexity and diversity of remote sensing images, particularly in terms of providing category information and addressing multi-scale challenges. The integration of the Batch Attention Pyramid and the Mask Decoder ensures that the model can perform panoptic segmentation effectively without the need for manual prompts.

2.2. Batch Attention Pyramid

The Batch Attention Pyramid takes the multi-scale feature maps extracted by the image encoder as input and aims to mine multi-scale information within the image features as well as contextual information between patches. The module can calculate the batch attention of each scale level feature and realize the feature fusion between scale levels The structure of the Batch Attention Pyramid is illustrated in Figure 3, where batch attention is calculated between image patches for each scale level and then image features from lower levels are upsampled and fused with adjacent higher-level features. To ensure the effectiveness of batch attention, the experiment reads image patches in the order of their spatial positions, thereby maintaining the continuity of image patches.

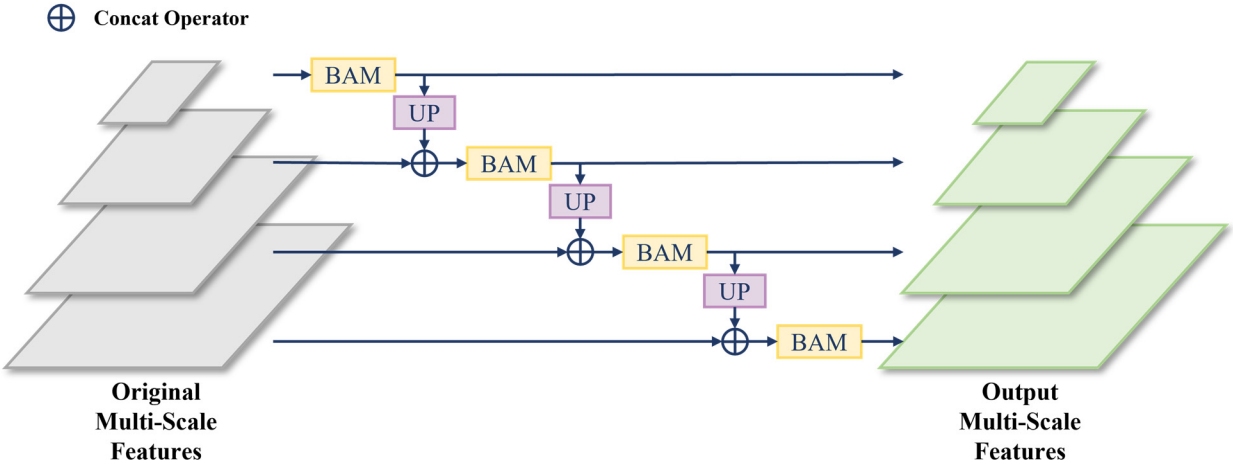


Figure 3. The structure of the Batch Attention Pyramid, where BAM is the Batch Attention Module and UP is the upsampling.

Due to limitations in memory resources and computing power, most remote sensing images with large pixel sizes cannot be directly used for training. Therefore, the original remote sensing images needed to be cropped before training, leading to a loss of contextual information. This issue is especially significant for large-scale surface objects (such as buildings and airport runways), which may lose a substantial amount of feature information because they cannot be fully contained within a single image patch. To address this problem, we utilized a Batch Attention Module (BAM) [31], which extracts long-range contextual information by calculating attention between different image patches within the same batch. This approach helps preserve critical contextual details that would otherwise be lost during cropping.

The structure of the BAM is shown in Figure 4. This module directly calculated attention between image patches from 1/4 features $X \in R^{BS \times C \times H \times W}$, which were extracted from the backbone network. First, the feature X was reshaped and transposed to obtain $X_Q \in R^{BS \times N}$, $X_K \in R^{N \times BS}$ and $X_V \in R^{BS \times N}$, where $N = C \times H \times W$. Then, X_Q was multiplied by X_K. The softmax normalization was applied to obtain the batch attention map $A \in R^{BS \times BS}$, and the calculation process of batch attention can be expressed as follows:

$$a_{ij} = \frac{\exp(X_i, X_j)}{\sum_{j=1}^{BS} \exp(X_i, X_j)} \tag{1}$$

$$e_i = \beta(a_{ij} X_j) + X_i \tag{2}$$

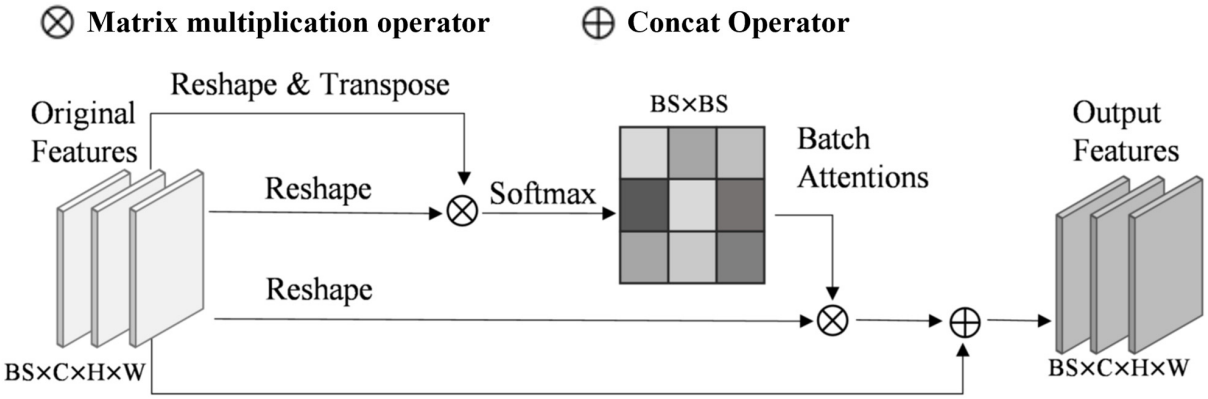


Figure 4. The structure of BAM. BS represents the training unit per GPU during training, C denotes number of channels in features, and H and W represent the height and width of the features.

Here, $a_{ij} \in A$ represents the impact of the j^{th} image patch on the i^{th} image patch. Finally, by multiplying the calculated attention map A with X_V and then reshaping again before addition to the original feature X, the final output feature $E \in R^{BS \times C \times H \times W}$ can be obtained by Formula (2), where $e_i \in E$ is the output feature of the i^{th} image patch, and β is a learnable scale parameter.

Building upon this, we define the original multi-scale features as F_l, where Batch Attention Module outputs the multi-scale features as F'_l, with $l = 1, 2, 3, 4$ representing the scale levels. Higher-level features have larger feature map resolutions. The calculation process of Batch Attention Module can be expressed as follows:

$$F'_l = BAM(F_l + F'_{l-1}) \tag{3}$$

Here, BAM denotes the computation of the Batch Attention Module, F'_{l-1} is the output result of the previous level's features, and F'_0 is set to 0. The final multi-scale feature output of the Batch Attention Module is composed of the output features F'_l from each level.

2.3. Mask Decoder

The structure of the Mask Decoder is illustrated in Figure 5, which is an improvement over the transformer decoder in Mask2Former [30]. This decoder contains two main modifications compared with the standard transformer decoder. Firstly, multi-scale feature maps are cyclically input into consecutive decoder layers according to their levels, thereby enhancing the efficiency of utilizing multi-scale features. Secondly, the original cross-attention is modified by adding an attention mask that only computes attention relationships at masked positions. Compared with the original structure in Mask2Former, the Mask Decoder in this paper, which only needs to generate low-resolution mask prompts, reduces the number of Mask Attention Layers to match the levels of the input multi-scale feature maps.

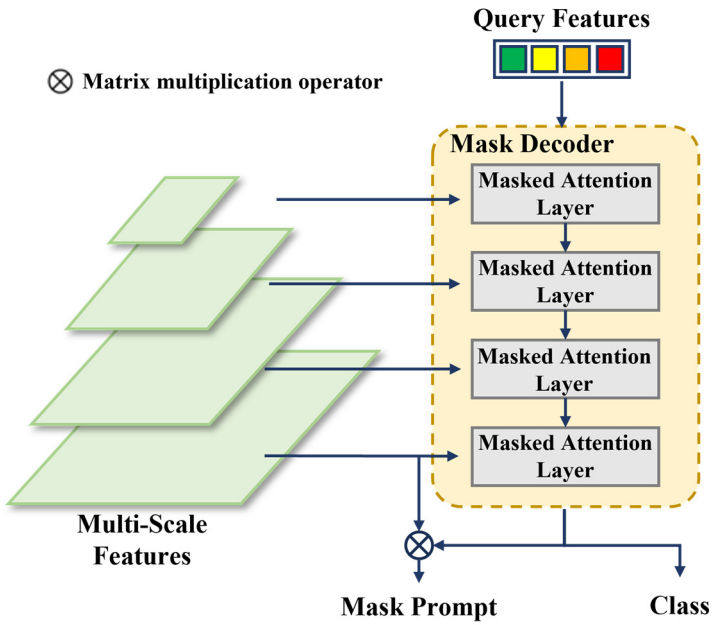


Figure 5. The structure of the Mask Decoder structure. The number of network layers in the Mask Decoder corresponds to the total levels of the input multi-scale feature maps, with each layer inputting a feature map from one scale level.

In each Mask Attention Layer, the inputs include multi-scale features, query features, and an attention mask. The layer primarily computes mask attention and self-attention, followed by a Feed-Forward Network (FFN) layer, which outputs a new attention mask for use in the next layer. The structure of the Mask Attention Layer is shown in Figure 6.

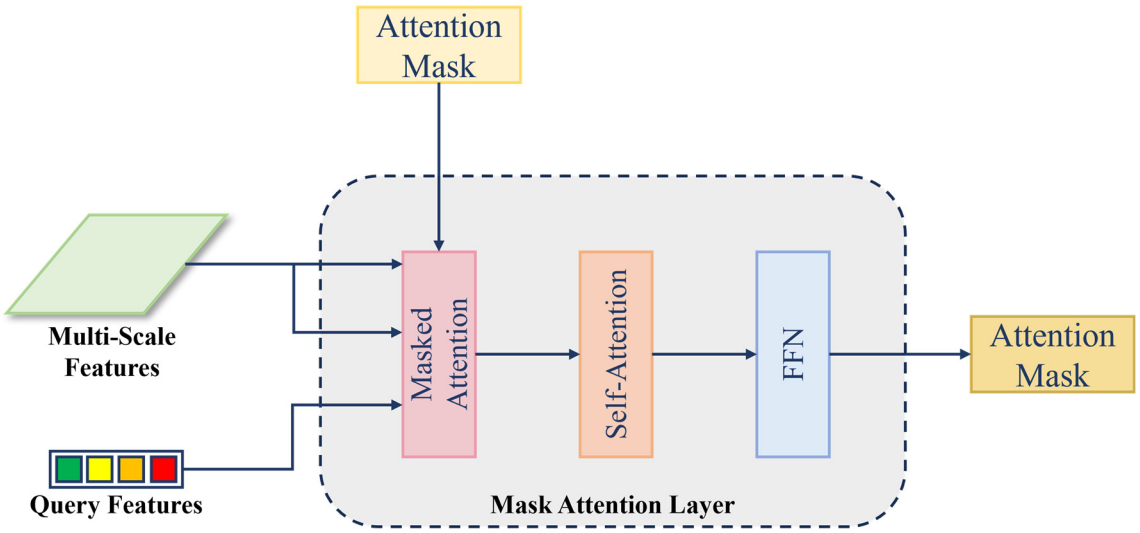


Figure 6. The structure of the Mask Attention Layer.

The most critical component, mask attention, is used to extract contextual information within the scope of the attention mask. The mask attention is a variant of cross-attention that only attends within the foreground region of the predicted mask for each query. Standard cross-attention (with residual path) computes as follows:

$$X_l = softmax\left(f_Q(X_{l-1}) f_K(F_l)\right) f_V(F_l) + X_{l-1} \quad (4)$$

Here, $l = 1, 2, 3, 4$ represents the multi-scale feature levels and the corresponding mask decoder levels, F_l represents the input multi-scale features, X_l is the query feature

output for each layer, X_0 is the query feature input to the decoder, and f_Q, f_K, f_V are linear transformations.

Let the attention mask be M_l. The calculation process of mask attention can be expressed as follows:

$$X_l = softmax\big(M_{l-1} + f_Q(X_{l-1})f_K(F_l)\big)f_V(F_l) + X_{l-1} \tag{5}$$

For the value of M_{l-1}, it is given by the following:

$$M_{l-1}(x,y) = \begin{cases} 0, \text{ if } m_{l-1}(x,y) = 1 \\ -\infty, \quad \text{otherwise} \end{cases} \tag{6}$$

where $m_{(l-1)}(x,y)$ is the pixel value of the attention mask at position (x,y).

3. Experiments and Analysis

3.1. Experimental Setup

3.1.1. Experimental Data

The experiments in this paper were conducted on the RSAPS-ASD, which comprises 102 high-resolution remote sensing images corresponding to 102 typical airports worldwide. Each image had a size of 4000 × 4000 pixels. All data were sourced from the SuperView-1 commercial satellite, which includes one panchromatic band and four multispectral bands (red, green, blue, and near-infrared) with a spatial resolution of 0.5 m. The panoptic segmentation annotations in this dataset categorize ground objects into 9 classes, including Building, Impervious Surface, Bare Land, Vegetation, Water, Taxiway, Runways, and Other, with Airplane singled out as a foreground class, as shown in Figure 7. The dataset can be accessed via the following link: https://github.com/SKLSEIIT/RSAPS-ASD accessed on 26 September 2024.

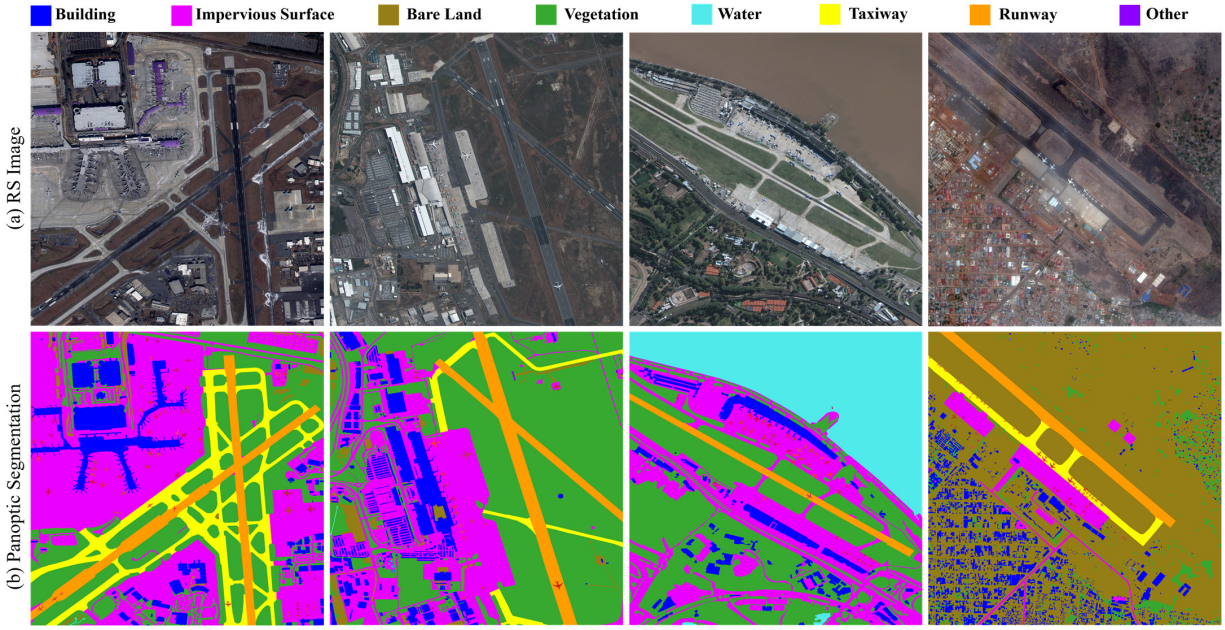


Figure 7. The examples of the RSAPS-ASD: (**a**) shows the true-color image and (**b**) displays the panoptic segmentation annotations, where background classes are represented by colors as indicated in the legend and each foreground airplane object instance is distinguished by a unique pixel value.

The experiments utilized the default dataset split of RSAPS-ASD, with all images divided into segments of 800 × 800 pixels with a stride of 600 pixels.

3.1.2. Evaluation Method

For the panoptic segmentation task, the Panoptic Quality (PQ) metric was adopted as the evaluation standard. This metric was introduced by [2] and is calculated as follows.

First, for a segmentation ground truth g and its predicted segmentation p, the Intersection over Union (IoU) is defined as the following:

$$IoU_{(p,g)} = \frac{|p \cap g|}{|p \cup g|} \tag{7}$$

where $IoU_{(p,g)}$ is IoU for ground truth g and its predicted segmentation p. A predicted segmentation p is considered to match the ground truth g if $IoU_{(p,g)} > 0.5$. For matched segmentation pairs, there are two scenarios: true positives (TP) where the prediction is correct, and false positives (FP) where the prediction is incorrect. Unmatched ground truth segments g are defined as false negatives (FN).

Based on these definitions, the Panoptic Quality (PQ) is calculated using the following formula:

$$PQ = \frac{\sum_{(p,g)\in TP} IoU_{(p,g)}}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \tag{8}$$

This evaluation metric takes into account both detection and segmentation aspects, providing a unified measure to assess the performance of the algorithm on both foreground and background object detection and segmentation.

3.1.3. Detailed Settings

The experiments were conducted on a computer equipped with four NVIDIA A100 GPUs, each with 40 GB of memory, for both training and testing validation. The following mainstream panoptic segmentation methods were selected for comparative experiments: Panoptic-FPN [32], Mask2Former [30], Panoptic SegFormer [33], and Mask DINO [34]. Panoptic-FPN uses ResNet-50 [35] as the backbone network with pre-trained weights from ImageNet. Mask2Former, Panoptic SegFormer, and Mask DINO utilize Swin-L [36] as the backbone network with pre-trained weights from ImageNet. The proposed RSPS-SAM method employs ViT-H [37] as the backbone network with pre-trained weights from SAM. More specifically, we used the AdamW optimizer and the step learning rate schedule. We used an initial learning rate of 0.0001 and a weight decay of 0.05 for all backbones. We applied a learning rate multiplier of 0.1 to the backbone and decayed the learning rate at 0.9 and 0.95 fractions of the total number of training steps by a factor of 10. A fixed batch size of 16 and a fixed training duration of 100 epochs were used for all methods. Other parameters were kept consistent with those in the original papers.

3.2. Results

3.2.1. Comparative Experiments

Following the experimental setup described above, the results of the comparative experiments are presented in Table 1. It is evident that the proposed RSPS-SAM achieves the best panoptic segmentation performance compared with other methods. Notably, RSPS-SAM shows a significant advantage in panoptic segmentation of background classes, outperforming the second-best method, Mask DINO, by 4.4 PQ in this category. Overall, RSPS-SAM improves the panoptic segmentation quality by 25.8 PQ compared with Panoptic-FPN, by 8.9 PQ compared with Mask2Former, by 7.8 PQ compared with Panoptic SegFormer, and by 3.2 PQ compared with Mask DINO.

Figure 8 displays the visual comparison results to provide a more intuitive representation of RSPS-SAM's panoptic segmentation performance. From the visual comparison, it is clear that RSPS-SAM has a significant advantage in the segmentation of various categories. By fully leveraging SAM's strong segmentation capabilities, RSPS-SAM achieves better segmentation results for ground objects with regular boundaries.

Table 1. Comparative experiment results.

Methods	PQ	PQ^{th} [1]	PQ^{st} [1]	Params.	GFLOPs	FPS
Panoptic-FPN	31.4	26.7	33.1	45.6M	170	14.5
Mask2Former	48.3	42.0	50.6	216M	868	4.1
Panoptic SegFormer	49.4	40.1	53.8	221M	886	3.9
Mask DINO	54.0	53.0	54.3	223M	1326	6.0
RSPS-SAM	57.2	52.8	58.7	740M	2973.7	1.2

[1] PQ^{th} and PQ^{st} represent the panoptic segmentation quality for the foreground (Airplane) and background (Building, Impervious Surface, Bare Land, etc.) classes, respectively.

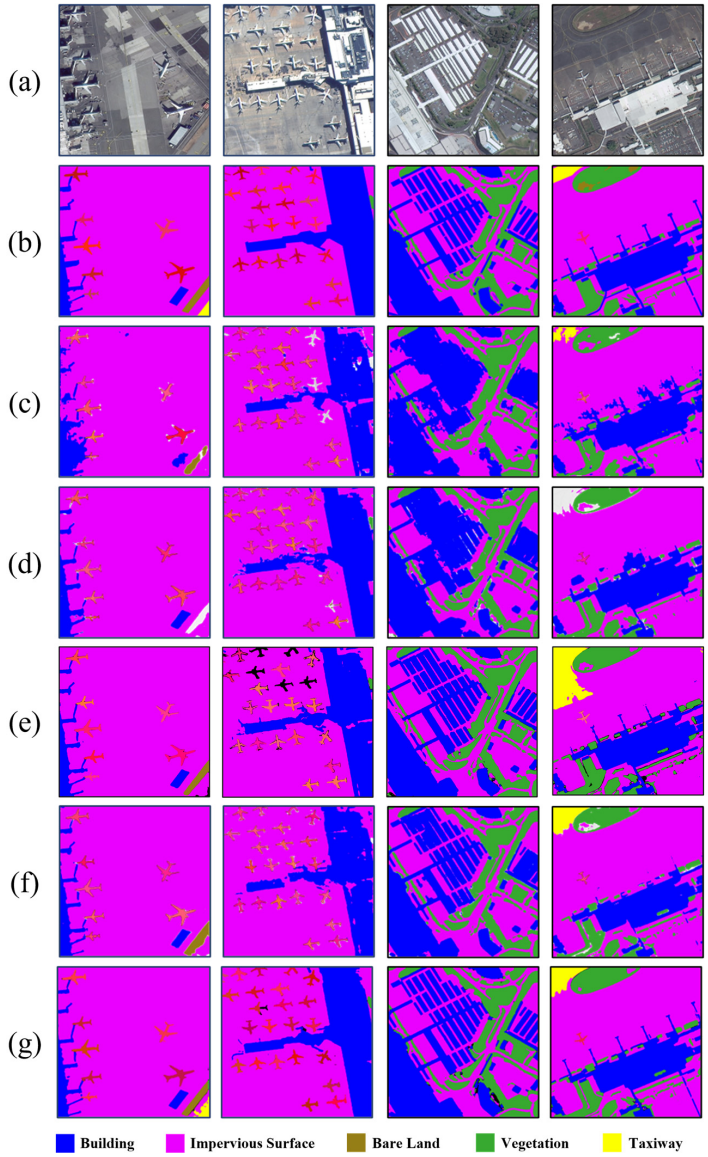


Figure 8. Visual comparison of experimental results. (**a**) True-color remote sensing image, (**b**) ground truth annotations, (**c**) results from Panoptic-FPN, (**d**) results from Mask2Former, (**e**) results from Panoptic Segformer, (**f**) results from Mask DINO, and (**g**) results from the proposed RSPS-SAM. Background classes of ground objects are represented by colors as indicated in the legend, and each foreground airplane object instance is distinguished by a unique pixel value.

Specifically, as shown in Table 1, both the $PQ^{th\,1}$ and $PQ^{st\,1}$ of Panoptic-FPN are significantly lower than those of other methods. This is primarily because Panoptic-FPN processes semantic segmentation and instance segmentation through two separate branches, resulting in insufficient feature sharing. Additionally, Panoptic-FPN uses a purely convolutional network, which is bad at understanding context and gathering global information, especially when dealing with scenes that are diverse and complicated. This makes it hard to achieve high performance.

Mask2Former, Panoptic SegFormer, and Mask DINO are all Transformer-based methods, and their panoptic segmentation accuracy is notably higher than that of Panoptic-FPN. Among them, Mask2Former has the lowest $PQ^{st\,1}$, mainly due to its weaker ability to capture detailed features of background categories, such as the blue building category in Figure 8d. Mask DINO has the highest panoptic segmentation accuracy among the three, especially in the $PQ^{th\,1}$ metric. Mask DINO's design, based on DINO (an improved version of DETR), leverages the Transformer's self-attention mechanism for object detection and segmentation, enabling it to process multiple objects simultaneously, thereby enhancing panoptic segmentation accuracy. Moreover, Mask DINO not only excels in panoptic segmentation but also inherits DINO's strengths in object detection, enabling it to better recognize and segment foreground objects in images.

The proposed RSPS-SAM leverages the advantages of SAM, reducing its dependence on large-scale labeled data and improving its generalization capability. By introducing the Batch Attention Pyramid, the model's ability to capture global information and understand context is further enhanced. As shown in the first column of Figure 8, the RSPS-SAM accurately segmented the taxiway in the lower right corner. Additionally, the method utilized mask attention, which enhanced the extraction of detailed features of background objects and improved segmentation accuracy. As illustrated in the third and fourth columns of Figure 8, the RSPS-SAM successfully segmented the detailed structures of the buildings.

3.2.2. Ablation Studies

To further validate the effectiveness of the proposed method, ablation experiments were designed on the RSAPS-ASD dataset following the experimental configuration described above. The ablation experiments primarily focus on the Batch Attention Pyramid. Firstly, the RSPS-SAM was compared with a model where the Batch Attention Pyramid was removed, and the results are shown in Table 2. It is evident that the use of the Batch Attention Pyramid effectively enhanced the panoptic segmentation performance of the model.

Table 2. Ablation experiment results for the Batch Attention Pyramid. All experiments were conducted with samples read in the order of their spatial positions.

Methods	Batch Attention Pyramid	PQ
RSPS-SAM	without	56.1
	with	57.2

Secondly, to validate the effectiveness of batch attention within the BAP, experiments were conducted using the method referenced from [31], with samples read in random order (where images within the batch do not correlate) and in sequential order (where images within the batch have spatial correlation). The results are shown in Table 3. These results indicate that batch attention can leverage long-range contextual information from sequential image slices to improve the performance of remote sensing image panoptic segmentation.

Table 3. Ablation experiment results for batch attention.

Methods	Loading Method	PQ
RSPS-SAM	Shuffled	55.8
	Ordered	57.2

By integrating and extracting multi-scale information and long-range contextual information between image patches, the Batch Attention Pyramid is able to accurately segment surface objects of various scales, even those that lack global characteristics due to image cutting. For instance, the BAP was required to correctly segment the "Taxiway" part in Figure 9a. Similarly, in Figure 9b, the BAP can utilize contextual information between image patches to accurately segment the "Impervious Surface".

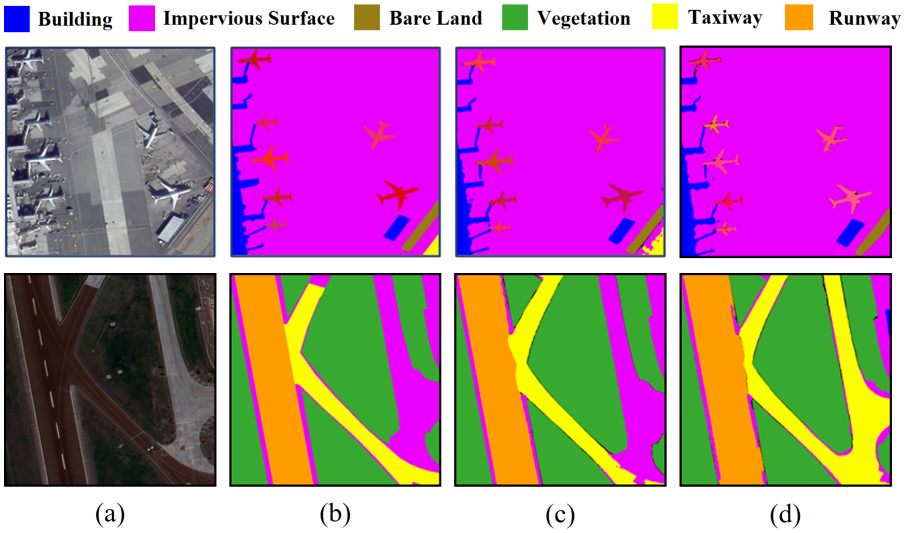


Figure 9. Visualization results of the BAP. (**a**) The RGB image patches, (**b**) ground truth, (**c**) the segmentation results with the BAP, (**d**) the segmentation results without the BAP.

3.2.3. Complexity and Computational Efficiency

As shown in Table 1, the Panoptic-FPN adopted a pure convolution structure, so it had the lowest parameter number and the highest inference speed, and its PQ value was the lowest. Among all the methods compared, the RSPS-SAM method proposed in this paper had the highest PQ. Due to the use of ViT-H with 632M parameters as the backbone, our method had a higher number of parameters and a lower inference speed. In the future, we will use technologies such as knowledge distillation to lightweight the model and improve computational efficiency.

4. Discussion

There are still some limitations that need further exploration. Firstly, although mask-based panoptic segmentation methods have proven effective in natural images [30,38], due to the presence of small and densely distributed targets in remote sensing images, the current model's segmentation performance for foreground objects is relatively poorer compared with the background. Secondly, as there are differences in geographical location and imaging time among airports in the dataset, there are also variations in ground object morphological features, texture characteristics, and coverage types between images. These differences affect the model's segmentation performance. It is necessary to explore deeper methods for ground object feature extraction to enhance the performance of remote sensing image panoptic segmentation.

5. Conclusions

This paper proposes a remote sensing image panoptic segmentation method based on SAM, called RSPS-SAM. The model achieves segmentation and classification of both foreground instances and background regions without the need for prompt input. By designing a Batch Attention Pyramid (BAP), the model integrates and extracts multi-scale information and long-range contextual information between cropped slices in remote sensing images, effectively improving panoptic segmentation performance on remote sensing images. Addi-

tionally, a Mask Decoder was constructed that utilized mask-based attention segmentation methods to automate the generation of mask prompts and category prediction. Experiments on the high-resolution remote sensing image airport scene dataset RSAPS-ASD have validated the effectiveness of the proposed method, achieving a final PQ of 57.2. In the future, we will continue to explore deeper methods for ground object feature extraction to further enhance the performance of remote sensing image panoptic segmentation.

Author Contributions: The following contributions were made to this research effort: Conceptualization, Z.L. (Zhuoran Liu) and G.H.; methodology, Z.L. (Zizhen Li), Y.L., and L.M.; software, Z.L. (Zhuoran Liu), Y.L., and B.S.; validation, G.H. and Y.L.; writing—original draft preparation, Z.L. (Zhuoran Liu) and Z.L. (Zizhen Li); writing—review and editing, Y.L., L.M., and C.P.; funding acquisition, L.M.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the funding provided by the National Natural Science Foundation of China (42171304), the Open Research Fund of the State Key Laboratory of Space–Earth Integrated Information Technology (SKL_SGIIT_20240303), and the Key Laboratory of Land Satellite Remote Sensing Application, Ministry of Natural Resources of the People's Republic of China (KLSMNR-K202301), Civil Space Advance Research Project of China (D040404).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, D.; Tong, Q.; Li, R.; Gong, J.; Zhang, L. Current Issues in High-Resolution Earth Observation Technology. *Sci. China Earth Sci.* **2012**, *55*, 1043–1051. [CrossRef]
2. Kirillov, A.; He, K.; Girshick, R.; Rother, C.; Dollar, P. Panoptic Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 9396–9405.
3. Fare Garnot, V.S.; Landrieu, L. Panoptic Segmentation of Satellite Image Time Series with Convolutional Temporal Attention Networks. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 4852–4861.
4. Zhang, Y.; Yang, J.; Li, X.; Wang, J. Autonomous Remote Sensing Investigation and Monitoring Technique of Typical Classes of Natural Resources and Its Application. *Geomat. World* **2022**, *29*, 66–73.
5. Xu, Y.; Qin, Y. Parameter selection experiment of urban block object segmentation based on Landsat 8. *J. Spatio Temporal Inf.* **2023**, *30*, 33–40. [CrossRef]
6. Weyler, J.; Läbe, T.; Behley, J.; Stachniss, C. Panoptic Segmentation with Partial Annotations for Agricultural Robots. *IEEE Robot. Autom. Lett.* **2024**, *9*, 1660–1667. [CrossRef]
7. Li, X.; Chen, D. A Survey on Deep Learning-Based Panoptic Segmentation. *Digit. Signal Process.* **2022**, *120*, 103283. [CrossRef]
8. Sakaino, H. PanopticRoad: Integrated Panoptic Road Segmentation Under Adversarial Conditions. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 20–22 June 2023; pp. 3591–3603.
9. De Carvalho, O.L.F.; De Carvalho Júnior, O.A.; Silva, C.R.E.; De Albuquerque, A.O.; Santana, N.C.; Borges, D.L.; Gomes, R.A.T.; Guimarães, R.F. Panoptic Segmentation Meets Remote Sensing. *Remote Sens.* **2022**, *14*, 965. [CrossRef]
10. Zhao, D.; Yuan, B.; Chen, Z.; Li, T.; Liu, Z.; Li, W.; Gao, Y. Panoptic Perception: A Novel Task and Fine-Grained Dataset for Universal Remote Sensing Image Interpretation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. [CrossRef]
11. Hua, X.; Wang, X.; Rui, T.; Shao, F.; Wang, D. Cascaded Panoptic Segmentation Method for High Resolution Remote Sensing Image. *Appl. Soft Comput.* **2021**, *109*, 107515. [CrossRef]
12. Khoshboresh-Masouleh, M.; Shah-Hosseini, R. Building Panoptic Change Segmentation with the Use of Uncertainty Estimation in Squeeze-and-Attention CNN and Remote Sensing Observations. *Int. J. Remote Sens.* **2021**, *42*, 7798–7820. [CrossRef]
13. Fernando, T.; Fookes, C.; Gammulle, H.; Denman, S.; Sridharan, S. Towards On-Board Panoptic Segmentation of Multispectral Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–12. [CrossRef]
14. Šarić, J.; Oršić, M.; Šegvić, S. Panoptic SwiftNet: Pyramidal Fusion for Real-Time Panoptic Segmentation. *Remote Sens.* **2023**, *15*, 1968. [CrossRef]
15. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning, Online, 18–24 July 2021; pp. 8748–8763.
16. Yuan, L.; Chen, D.; Chen, Y.-L.; Codella, N.; Dai, X.; Gao, J.; Hu, H.; Huang, X.; Li, B.; Li, C.; et al. Florence: A New Foundation Model for Computer Vision. *arXiv* **2021**, arXiv:2111.11432.
17. Bao, H.; Dong, L.; Piao, S.; Wei, F. BEiT: BERT Pre-Training of Image Transformers. *arXiv* **2021**, arXiv:2106.08254.

18. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment Anything. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 3992–4003.
19. Ren, Y.; Yang, X.; Wang, Z.; Yu, G.; Liu, Y.; Liu, X.; Meng, D.; Zhang, Q.; Yu, G. Segment Anything Model (SAM) Assisted Remote Sensing Supervision for Mariculture—Using Liaoning Province, China as an Example. *Remote Sens.* **2023**, *15*, 5781. [CrossRef]
20. Wu, J.; Ji, W.; Liu, Y.; Fu, H.; Xu, M.; Xu, Y.; Jin, Y. Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation. *arXiv* **2023**, arXiv:2304.12620.
21. Zhao, Z. Enhancing Autonomous Driving with Grounded-Segment Anything Model: Limitations and Mitigations. In Proceedings of the 2023 IEEE 3rd International Conference on Data Science and Computer Application (ICDSCA), Dalian, China, 27–29 October 2023; pp. 1258–1265.
22. Wang, D.; Zhang, J.; Du, B.; Xu, M.; Liu, L.; Tao, D.; Zhang, L. SAMRS: Scaling-up Remote Sensing Segmentation Dataset with Segment Anything Model. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Curran Associates, Inc.: New York, NY, USA, 2023; Volume 36, pp. 8815–8827.
23. Chen, X.; Wu, W.; Yang, W.; Qin, H.; Wu, X.; Mao, X. Make Segment Anything Model Perfect on Shadow Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–13. [CrossRef]
24. Qian, X.; Lin, C.; Chen, Z.; Wang, W. SAM-Induced Pseudo Fully Supervised Learning for Weakly Supervised Object Detection in Remote Sensing Images. *Remote Sens.* **2024**, *16*, 1532. [CrossRef]
25. Osco, L.P.; Wu, Q.; De Lemos, E.L.; Gonçalves, W.N.; Ramos, A.P.M.; Li, J.; Marcato, J. The Segment Anything Model (SAM) for Remote Sensing Applications: From Zero to One Shot. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *124*, 103540. [CrossRef]
26. Ji, W.; Li, J.; Bi, Q.; Liu, T.; Li, W.; Cheng, L. Segment Anything Is Not Always Perfect: An Investigation of SAM on Different Real-World Applications. *Mach. Intell. Res.* **2024**, *21*, 617–630. [CrossRef]
27. Yan, Z.; Li, J.; Li, X.; Zhou, R.; Zhang, W.; Feng, Y.; Diao, W.; Fu, K.; Sun, X. RingMo-SAM: A Foundation Model for Segment Anything in Multimodal Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–16. [CrossRef]
28. 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 Trans. Geosci. Remote Sens.* **2024**, *62*, 1–17. [CrossRef]
29. Nguyen, K.D.; Phung, T.-H.; Cao, H.-G. A SAM-Based Solution for Hierarchical Panoptic Segmentation of Crops and Weeds Competition. *arXiv* **2023**, arXiv:2309.13578.
30. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-Attention Mask Transformer for Universal Image Segmentation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–20 June 2022; pp. 1280–1289.
31. Li, Z.; He, G.; Fu, H.; Chen, Q.; Shangguan, B.; Feng, P.; Jin, S. RS DINO: A Novel Panoptic Segmentation Algorithm for High Resolution Remote Sensing Images. In Proceedings of the 2023 11th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Wuhan, China, 25–28 July 2023; pp. 1–5.
32. Kirillov, A.; Girshick, R.; He, K.; Dollar, P. Panoptic Feature Pyramid Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 6392–6401.
33. Li, Z.; Wang, W.; Xie, E.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P.; Lu, T. Panoptic SegFormer: Delving Deeper into Panoptic Segmentation with Transformers. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 1270–1279. [CrossRef]
34. Li, F.; Zhang, H.; Xu, H.; Liu, S.; Zhang, L.; Ni, L.M.; Shum, H.-Y. Mask DINO: Towards A Unified Transformer-Based Framework for Object Detection and Segmentation. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 3041–3050.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
36. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 9992–10002.
37. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
38. Cheng, B.; Schwing, A.; Kirillov, A. Per-Pixel Classification Is Not All You Need for Semantic Segmentation. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–14 December 2021; Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W., Eds.; Curran Associates, Inc.: New York, NY, USA, 2021; Volume 34, pp. 17864–17875.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.