



Article

A Lightweight Pyramid Transformer for High-Resolution SAR Image-Based Building Classification in Port Regions

Bo Zhang ^{1,2}, Qian Wu ¹, Fan Wu ^{3,4,*} , Jiajia Huang ^{1,2} and Chao Wang ^{2,4}

¹ Airborne Remote Sensing Center, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; zhangbo@radi.ac.cn (B.Z.); wuqian20@mails.ucas.ac.cn (Q.W.); huanjiajia22@mails.ucas.ac.cn (J.H.)

² College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China; wangchao@radi.ac.cn

³ Key Laboratory of Earth Observation of Hainan Province, Hainan Aerospace Information Research Institute, Sanya 572000, China

⁴ International Research Center of Big Data for Sustainable Development Goals, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China

* Correspondence: wufan@aircas.ac.cn

Abstract: Automatic classification of buildings within port areas from synthetic aperture radar (SAR) images is crucial for effective port monitoring and planning. Yet, the unique challenges of SAR imaging, such as side-looking geometry, multi-bouncing scattering, and the compact arrangement of structures, often lead to incomplete building structures and blurred boundaries in classification results. To address these issues, this paper introduces SPformer, an efficient and lightweight pyramid transformer model tailored for semantic segmentation. The SPformer utilizes a pyramid transformer encoder with spatially separable self-attention (SSSA) to refine both local and global spatial information and to process multi-scale features, enhancing the accuracy of building structure delineation. It also integrates a lightweight all multi-layer perceptron (ALL-MLP) decoder to consolidate multi-scale information across various depths and attention scopes, refining detail processing. Experimental results on the Gaofen-3 (GF-3) 1 m port building classification dataset demonstrate the effectiveness of SPformer, achieving competitive performance compared to state-of-the-art models, with mean intersection over union (mIoU) and mean F1-score (mF1) reaching 77.14% and 87.04%, respectively, while maintaining a compact model size and lower computational requirements. Experiments conducted on the entire scene of SAR images covering port area also show the good capabilities of the proposed method.

Keywords: synthetic aperture radar image; building classification; semantic segmentation; pyramid transformer; deep learning



Citation: Zhang, B.; Wu, Q.; Wu, F.; Huang, J.; Wang, C. A Lightweight Pyramid Transformer for High-Resolution SAR Image-Based Building Classification in Port Regions. *Remote Sens.* **2024**, *16*, 3218. <https://doi.org/10.3390/rs16173218>

Academic Editor: Dusan Gleich

Received: 16 July 2024

Revised: 25 August 2024

Accepted: 28 August 2024

Published: 30 August 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

Ports serve as comprehensive transportation and storage hubs, encompassing various man-made infrastructures such as industrial factories, container yards, and oil storage facilities. These structures perform crucial roles such as energy provision, cargo warehousing, and transportation services. Understanding these structures' spatial distribution and utilization within the port area is essential for planning and development to accommodate the dynamic needs of cargo logistics and storage. This necessitates research dedicated to classifying typical buildings in the port area. While traditional optical remote sensing encounters limitations in complex port environments, synthetic aperture radar (SAR) technology offers consistent, high-quality radar imagery irrespective of weather conditions. Moreover, advancements in high-resolution SAR data enable the capture of finer texture details and structural characteristics, facilitating the monitoring of diverse and smaller buildings. However, the dense arrangement of buildings in the port area poses

challenges, as intra-class and inter-class scattering characteristics become intertwined in high-resolution SAR images, thus presenting an obstacle to accurate building classification within port SAR imagery.

Deep learning methods significantly enhanced the precision and resilience of image segmentation and classification. Convolutional neural networks (CNNs) excel in capturing local information through convolutional operations, exhibiting both locality and shift invariance [1,2]. Innovations such as pyramid pooling, encoder–decoder structures, and extended convolutions expanded the receptive fields, yielding networks such as U-Net [3], PSPNet [4], and DeepLabV3/V3+ [5,6]. Nevertheless, CNNs encounter difficulties in modeling long-range semantic dependencies due to fixed receptive fields and local feature extraction. In contrast, transformer-based techniques excel in handling long-range dependencies with global receptive fields and parallel computation [7], rendering them more robust than CNNs [8], a vital consideration for high-resolution remote sensing imagery. The vision transformer (ViT) demonstrated that pure transformer architectures can achieve state-of-the-art performance in image classification tasks [9]. However, the ViT generates single-scale, low-resolution feature maps, which are less suitable for high-resolution semantic segmentation. To address this, the pyramid-based transformer was introduced. It maintains linear complexity and leverages the strengths of both CNNs and transformers, significantly enhancing computational efficiency and the ability to process multi-scale features [10]. Furthermore, its design allows it to serve as a direct substitute for the CNN backbone in dense prediction tasks, enhancing its practical applicability.

In port areas, structures such as factory buildings, containers, and oil tanks are often densely distributed. They usually show double-bounce bright lines, layovers, and shadows in high-resolution SAR images. Moreover, the features may be overlapped due to the short distances between them. Those may lead to geometric and radiative distortion in the image and make the task of pixel classification face many difficulties, and the classification results are prone to lack of structure and unclear boundaries. Traditional classification methods may struggle with the multi-scale and multi-class classification of large-scene SAR images. Therefore, there is a need for a semantic segmentation network that offers both high precision and high generalization capabilities for intensive prediction tasks in high-resolution SAR images. Additionally, in the context of remote sensing applications, the computational complexity of the model is a crucial factor to consider. It is essential to strike a balance between the model's accuracy and its efficiency to enhance its practical utility. This balance ensures that the model can handle the demands of large-scale image analysis without compromising on the quality of the segmentation results.

Considering the complexities of SAR image segmentation in port areas, we introduce SPformer, a lightweight and efficient model tailored for classifying typical port structures. The key contributions of this work are as follows:

- (1) Development of the SPformer model, which integrates pyramid transformer encoding technology based on spatially separable self-attention (SSSA). It enhances both local and global spatial information encoding, as well as multi-scale feature processing capabilities. As a result, it significantly improves the integrity of surface object structure extraction in SAR images.
- (2) A simple and effective lightweight multilayer perceptron (MLP) decoding method suitable for a transformer feature pyramid is designed, which can aggregate multi-scale semantic information of different depths and different attention ranges to improve the edge detail processing capability of ground objects.

2. Related Work

2.1. Image Semantic Segmentation

Semantic segmentation typically employs CNN-based models that capture local information through convolution operations and increase receptive fields to capture a larger range of context information through strategies such as pooling layers or skipping connections [11]. Fully convolutional networks (FCN) are first applied to the semantic seg-

mentation of optical images [12]. Strategies such as pyramid pooling, encoding–decoding structures, dilated convolution, and skip connections are then used to build new networks. However, those CNNs train weight parameters in the convolution kernel to extract and combine features from the receptive field, and they struggle with a limited receptive field [1]. To address these limitations, transformer-based networks for semantic segmentation emerged. Transformers leverage a self-attention mechanism that assesses the similarity between all pairs of patches through a dot product of the vectors. This allows for adaptive feature extraction and mixing features across the entire set of patches. The mechanism provides a broad, global receptive field, mitigating the model’s bias and effectively capturing long-distance information. Additionally, the parallel computing capability and learnable attention mechanism of the transformer can effectively tackle the challenge of information bottleneck [7]. Therefore, compared to CNNs and similar multi-layer perceptron structures, transformer is considered to have a more robust generalization capability [8].

In dense prediction tasks such as semantic segmentation, the extraction of high-resolution feature maps is particularly crucial. ViT, which has a global receptive field, typically stacks transformer encoder layers of the same size, outputting low-resolution features of a single scale. This limits the transformer model’s ability to capture the critical local details of specific tasks. Although increasing the input image resolution and reducing path size can enhance feature granularity, it leads to an increase in ViT’s computational load. The quadratic computational complexity makes it difficult for ViT to be applied to high-resolution image segmentation tasks. Moreover, as the depth of the transformer model increases, ViT’s columnar stacking structure may lead to a dispersion of attention. To overcome the performance limitations of ViT, transformer encoders based on pyramid structures were applied to semantic segmentation tasks. Compared to ViT, the pyramid-based transformer encoders not only inherit the advantages of CNNs and transformers, but also significantly reduce the computational load of feature maps and output hierarchical feature maps at various scales, especially high-resolution ones. Pyramid transformers can replace the CNN backbone network in dense prediction tasks and integrate directly with existing backends to support a variety of downstream tasks. The pyramid vision transformer (PVT) is the first model to introduce the pyramid hierarchical structure from CNN into the transformer backbone network. It can serve as a general backbone for computer vision and is more suitable for dense prediction tasks [10]. Subsequently, some improved linear complexity pyramid transformer models emerged, aiming to enhance the local continuity of features and eliminate fixed-size positional encoding. For example, the Swin transformer uses relative position encoding to replace fixed-size positional embedding and proposes multi-resolution feature extraction and shifted window-based self-attention transformer modules, limiting self-attention to local windows [13]. Twins, building on PVT and Swin, introduced a self-attention mechanism that combines local and global features, achieving a good balance between computational effort and receptive field, and obtaining stronger feature representation [14]. Twins are capable of replacing traditional CNN backbone networks in common dense prediction tasks such as object detection and semantic segmentation.

2.2. Building Classification with SAR Image

In recent years, semantic segmentation networks based on deep learning became the mainstream method for high-resolution SAR image terrain classification, aiming at classifying every pixel in the image [15]. At present, various examples of literature used convolutional neural network-based methods to extract and classify ground objects such as buildings, water bodies, roads, and grasslands in remote sensing images. Wu et al. [16] employed an improved UNet framework along with 10 m resolution GF-3 SAR image to map built-up areas across China. The extraction accuracy for 34 provinces and regions ranged from 76.05% to 93.45%. Li et al. [17] proposed a dual-attention transformer model based on Segformer for building area extraction in medium-resolution SAR images. Utilizing GF-3 10 m and Sentinel-1 20 m SAR images, they achieved an overall accuracy of 86% for the

mapping of building areas in China for the year 2020. Wu et al. [18] designed a SAR image semantic segmentation algorithm based on cross-regional context and noise regularization (CCNR). The CCNR has three heads to output the segmentation results, the representation features of each pixel, and the reconstructed image. Self-attention and contrast learning are used to explore the regional and pixel-level semantic relationships between different images and realize the aggregation of similar pixels. The experimental results on three large-scene SAR images demonstrate the effectiveness of CCNR. Kang et al. [19] proposed a building extraction method based on the DisOptNet network, which extracted useful semantic knowledge from optical images into a network trained only with SAR data through knowledge distillation. Compared with the UNet model, dice accuracy was improved by 3.49%. In addition to the research on building extraction algorithms for high-resolution SAR images, Wangiyana et al. [20] studied the impact of data enhancement methods on the performance of building extraction networks. Experiments show that geometric transformations such as scaling and rotation are more efficient than pixel transformations and that data enhancement is better for both training and testing than for only the training phase. Due to the lack of a high-resolution SAR open-source dataset and the limitations of applicability, some researchers use small-sample SAR data to carry out building extraction and ground object segmentation research. Yue et al. [21] introduced a multi-scale attention network for terrain classification in high-resolution airborne SAR images. This network enhances the model's feature learning capabilities by employing multi-scale feature extraction, channel attention mechanisms, and spatial attention mechanisms. Sun et al. [22], leveraging submeter resolution TerraSAR-X SL model SAR images of the Berlin area, proposed a conditional GIS aware network (CG-Net) that incorporates building profile information from GIS data as ancillary data. CG-Net is designed to learn multi-level visual features and use standardized building contour features, thereby accomplishing the segmentation of individual building instances within ultra-high resolution SAR images of large-scale urban areas.

The quality and size of datasets are key factors in the performance of deep learning models, yet there remains a scarcity of publicly available datasets for high-resolution SAR image semantic segmentation. Xia et al. [23] assembled a dataset for semantic segmentation by selecting 11 scenes of GF-3 spotlight (SL) mode 1 m resolution SAR images from nine cities across seven countries, using OSM vector data to construct building labels. Shi et al. [24] created the FUSAR-MAP dataset for semantic segmentation of surface features, including buildings, water bodies, roads, and vegetation, by choosing eight ultra-fine stripmap (UFS) 3 m resolution SAR images from six provinces in China. Furthermore, a 0.5 m resolution open dataset known as SpaceNet 6 was published for building extraction. This dataset is based on Capella SAR images and Worldview-2 optical images [25]. It incorporates a variety of data types, such as SAR, PAN, RGB, and RGB-NIR, and is annotated using DEM data derived from 3D LiDAR point clouds. The dataset includes 48,000 buildings, predominantly consisting of low-rise structures, so the imaging area of buildings on SAR image is less different from the actual building footprint location. Currently, high-precision DEM data, OSM vector data, and TomoSAR point clouds are among the auxiliary measurement data used to reference building annotations in high-resolution SAR images. This helps to obtain accurate footprint information for annotation purposes. However, due to the side-view imaging mechanism of SAR images in most datasets, there is often a discrepancy between the backscatter area of mid-to-high-rise buildings in SAR images and their actual footprints [26]. This limitation hinders the network's ability to learn the full range of imaging features for buildings, leading to reduced accuracy in neural network-based building extraction.

3. Materials and Methods

Buildings in high-resolution SAR images often display complex features such as double-bounce bright lines, layovers, and shadows. Transformer models excel in processing distant visual dependencies. For dense remote sensing predictions, low computational

complexity is crucial. Thus, this study proposes a streamlined network, SPformer, optimized for high-resolution SAR building classification. SPformer consists of two modules (as in Figure 1): (1) a pyramid-structured transformer encoder Twins-SVT [14] based on SSSA, providing low-level high-res spatial and high-level low-res semantic features; and (2) a lightweight ALL-MLP decoder that top-down combines multiscale feature data, forming a robust segmentation representation through local and global attention features across layers.

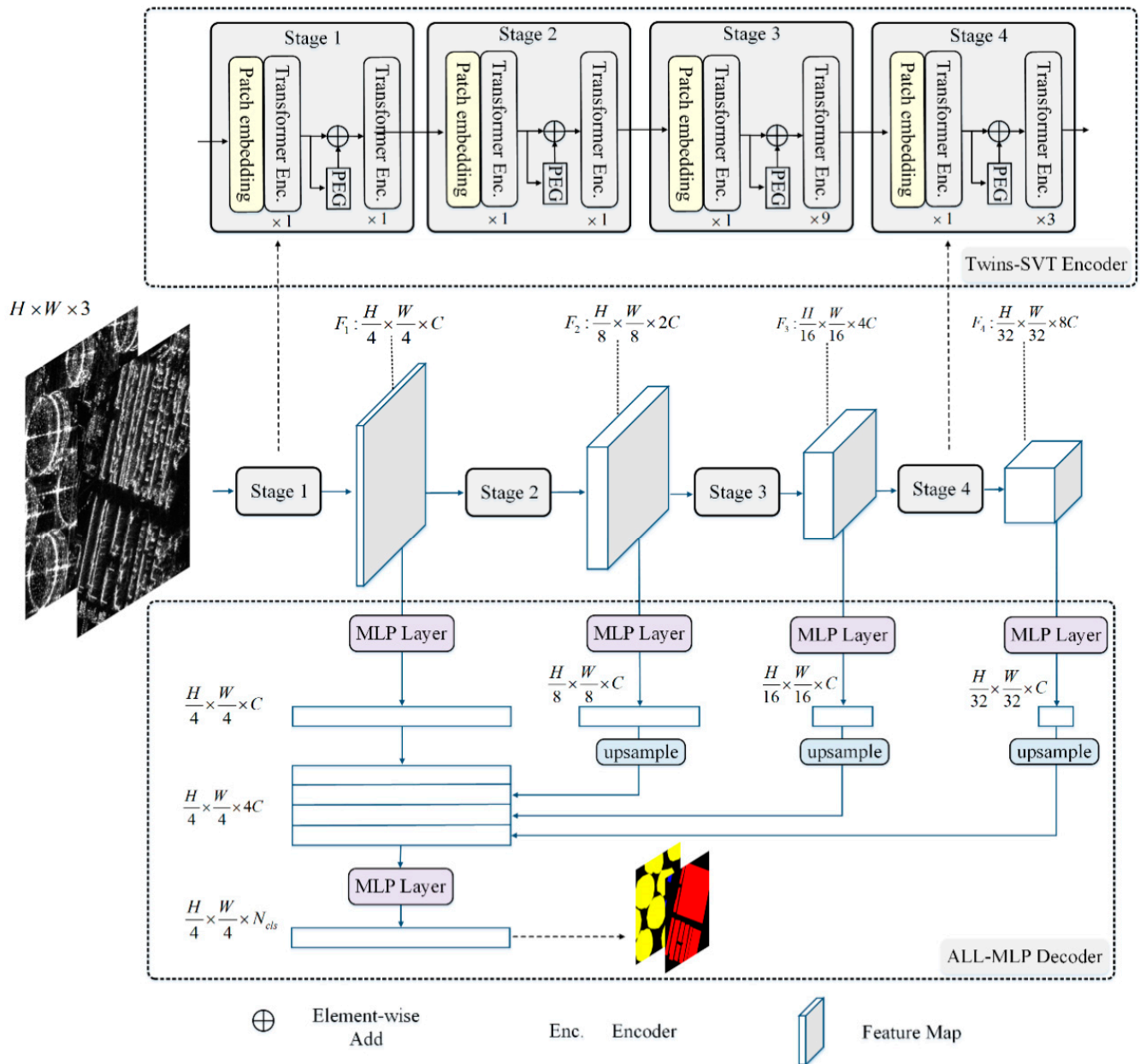


Figure 1. The overall architecture of the proposed method SPformer.

3.1. Twins-SVT Encoder

SPformer utilizes the linear complexity network, namely Twins-SVT-S, as the encoder. The Twins-SVT encoder adopts a flexible pyramid hierarchical structure, which is primarily divided into four stages, to generate feature maps of different scales and channels (as in Figure 1). Each stage comprises a patch embedding module, a position encoding generator (PEG) module [27], and multiple transformer encoder modules.

The input image sample size to the encoder is $512 \times 512 \times 3$, and the dimension C of stage one in Twins-SVT-S is 64. At each stage, the encoder outputs feature maps with varying sizes, namely F1: $128 \times 128 \times 64$, F2: $64 \times 64 \times 128$, F3: $32 \times 32 \times 256$, and F4: $16 \times 16 \times 512$, respectively.

The patch embedding module initially processes the input image. The model's initial input image, denoted as $I \in R^{H \times W \times C}$, can be considered as a sequence of $H \times W$ one-dimensional vectors, each with a dimension of C . By reducing the number of patch tokens through the convolution layer, the dimension of each patch token is increased. From stage one to stage four of the Twins SVT-S model, there are 128×128 , 64×64 , 32×32 , and 16×16 patch token sequences, with corresponding dimensions of 64, 128, 256, and 512, respectively.

The patch embedding module is an important part of the transformer encoder. As shown in Figure 2, the main steps of the module include: The input patch tokens are first converted into three-dimensional (3D) feature maps. Then the number of patch tokens is reduced and each token dimension is increased by convolution and LayerNorm. Finally, the 3D feature maps obtained by convolution are converted into one-dimensional patch token vectors for input into subsequent transformer encoders for processing.

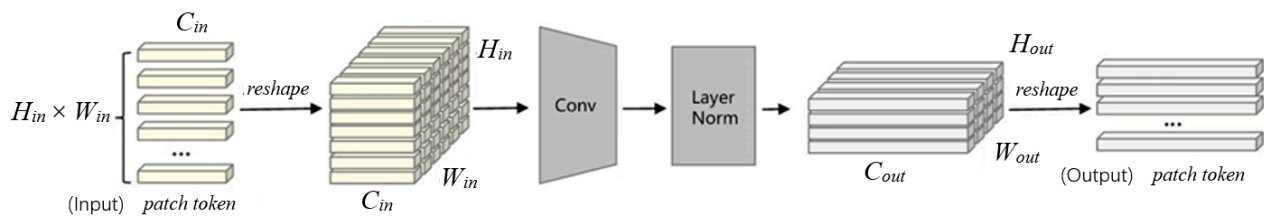


Figure 2. Schematic diagram of the patch embedding module.

3.2. Spatially Separable Self-Attention

To balance the computational efficiency and receptive field of the self-attention mechanism, Twins-SVT introduces SSSA to perceive features at both local and global levels. The SSSA module can divide image features into multiple groups in the spatial dimension, calculate self-attention within each local group, respectively, and finally use global self-attention to fuse information between groups. In the twin-SVT model, the SSSA module plays an important role, because it not only effectively captures spatial relationships between features, but also significantly reduces computational effort and improves model efficiency [14].

Specifically, SSSA consists of two modules: locally grouped self-attention (LSA) and global sub-sampled self-attention (GSA) (as shown in Figure 3). In different stages of the Twins-SVT encoder, LSA and GSA modules alternate to fuse local and global receptive fields of features effectively.

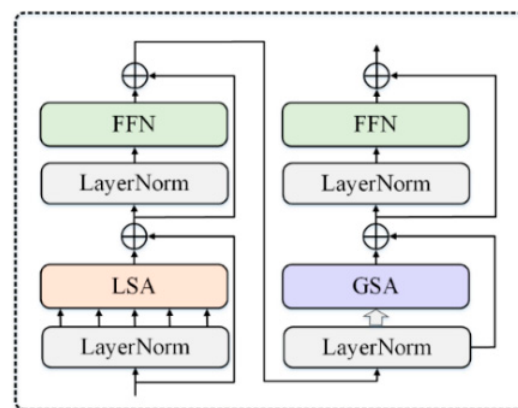


Figure 3. Schematic diagram of transformer encoder based on spatially separable self-attention.

Each transformer encoder contains a self-attention module and a feedforward network (FFN) module. The FFN module consists of two fully connected layers and one GeLU nonlinear layer. Each self-attention module and FFN module is preceded by a layer normalization (LayerNorm) module. The LayerNorm module accelerates and stabilizes the training process and prevents overfitting. Additionally, a residual connection exists between the input of LayerNorm layer and the output of the self-attention module and the FFN module. The difference between the two encoders lies in their use of different self-attention modules, namely the LSA module or the GSA module.

The input encoder's patch token sequence is presented as $F_{in} = (F_{in}^1, F_{in}^2, \dots, F_{in}^n)$, where $F_{in}^i \in R^{C_j}$, and the output patch token sequence of the encoder is denoted as $F_{out} = (F_{out}^1, F_{out}^2, \dots, F_{out}^n)$, where $F_{out}^i \in R^{C_j}$. The transformer encoder can be expressed as follows:

$$\hat{F} = SA(\text{LayerNorm}(F_{in})) + F_{in} \quad (1)$$

$$F_{out} = \text{FFN}(\text{LayerNorm}(\hat{F})) + \hat{F}. \quad (2)$$

SA is the self-attention module used in the corresponding transformer encoder, i.e., the LSA or GSA module.

As illustrated in Figure 4, LSA can partition image features into multiple groups along the spatial dimension and calculate self-attention for each local group independently. Subsequently, GSA is employed to integrate information across different groups. The SSSA module is a critical component of the Twins-SVT model, as it not only effectively captures spatial relationships among features, but also reduces computational expenses significantly, thereby enhancing model efficiency.

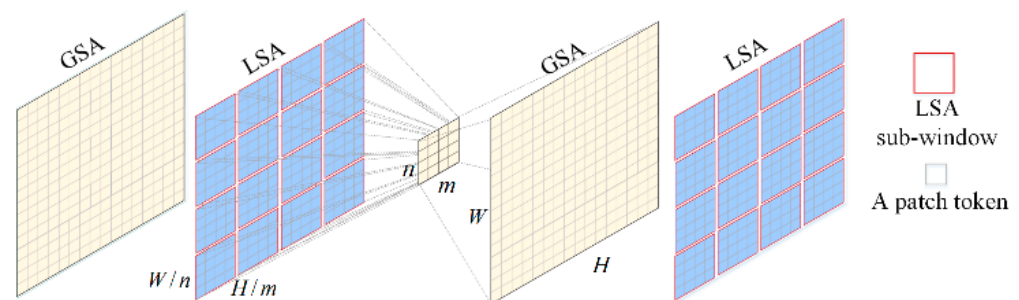


Figure 4. Schematic view of the locally grouped attention (LSA) and global sub-sampled attention (GSA).

3.3. ALL-MLP Decoder

In this study, a lightweight and effective multi-level feature aggregation ALL-MLP decoder is proposed. The key is to leverage the multi-scale features generated by the pyramid transformer encoder at different levels, where shallow attention tends to preserve local features. In contrast, deep attention focuses on global features. The ALL-MLP decoder can obtain a segmentation representation containing local details and global context through simple upsampling and parallel fusion. Compared to conventional convolutional decoders, it maintains high performance while reducing computational complexity and model parameters and has better scalability and practicality.

The ALL-MLP decoder, consisting solely of MLP layers, was originally introduced in Segformer [28]. In SPformer, as depicted in Figure 1, the ALL-MLP decoder is employed to aggregate the multi-scale feature outputs from four stages, encompassing low-dimensional high-resolution features and high-dimensional low-resolution features. These fused features are then upsampled to generate the semantic segmentation mask output. The primary process involves four steps: firstly, using MLP layers to transform the multi-scale output feature map F_i from each stage of the Twins-SVT encoder to a feature map with C channels;

secondly, upsampling the high-dimensional low-resolution feature map to match the size of the feature map in stage one and concatenating them; thirdly, fusing the concatenated features using MLP layers; and finally, mapping the fused features to the $\frac{H}{4} \times \frac{W}{4} \times N_{cls}$ dimensional semantic segmentation mask using another MLP layer, where N_{cls} denotes the number of categories. Formally, the ALL-MLP decoder can be expressed as:

$$\hat{F}_i = \text{Linear}(C_i, C)(F_i), \forall i \quad (3)$$

$$\hat{F}_i = \text{Upsample}\left(\frac{W}{4} \times \frac{W}{4}\right)(F_i), \forall i \quad (4)$$

$$F = \text{Linear}(4C, C)(\text{Concate}(\hat{F}_i)), \forall i \quad (5)$$

$$M = \text{Linear}(C, N_{cls})(F) \quad (6)$$

Here, M is the predicted mask and $\text{Linear}(C_{in}, C_{out})(\cdot)$ represents linear layers with input and output vector dimensions of C_{in} and C_{out} , respectively.

4. Experiments and Results

4.1. Experimental Dataset

This study focuses on oil tanks, containers, and factory buildings within port areas. Due to the lack of a public dataset, we therefore created a multi-category semantic segmentation dataset at the pixel level for model training and testing.

From the previous analysis, it can be seen that artificial structures mainly contain layover, multipath scattering bright lines, and shadows on high-resolution SAR images. Buildings of various types display a range of imaging features, details, and patterns within the images. However, the building footprints fail to precisely delineate the full extent of the imaging area of structures on high-resolution SAR images. Annotations relying solely on building footprints can result in the omission of significant ground scattering characteristics, particularly for mid-to-high-rise buildings. Consequently, this study, leveraging optical imagery and expert experience, adopts the actual scattering feature area of buildings on SAR images as the benchmark for annotations. Pixel-level marking areas cover the complete scattering characteristics of the buildings, including layover, scattering bright lines, and shadows. In particular, when the arrangement of objects is too dense to distinguish individual buildings, such as stacked containers, the annotation is based on dense areas and grid lines as the basic unit for the same type of object area. Figure 5 illustrates some examples of typical building annotations in the port area.

In this study, 19 scenes of GF-3 1 m resolution spotlight mode SAR images from seven different port areas in four countries were selected as the data source, and a semantic segmentation dataset of typical features in the port area at the pixel level of high-resolution SAR images was constructed, as shown in Table 1. The incidence angles of the images range from 22° to 48°, with 9 ascending orbit images and 10 descending orbit images included.

Table 1. Basic information of data set of port area object classification.

Area	Image Quantity	Imaging Mode	Resolution	Band
Port of Yokohama	19 scenes, including 9 ascending and 10 descending scenes	Spotlight mode	1 m	C
Port of Shanghai				
Port of Honolulu				
Port of Gosport				
Port of Qingdao				
Zhoushan Port				
Port of Xiamen Gaoqi				

The dataset includes 1630 effective sample patches, each with a size of 512×512 , with a training-to-testing data ratio of 8:2. Compared with containers and oil tanks, factory

buildings exhibit larger sizes, with the number of pixels in a single factory considerably surpassing that of containers and oil tanks. Therefore, the pixel ratio of the factory building, container, and oil tank samples in the dataset is approximately 6.5:2:1.5. The test data include randomly selected slices from images of six different port areas, with the pixel proportions of factories, containers, and oil tanks consistent with the overall dataset.

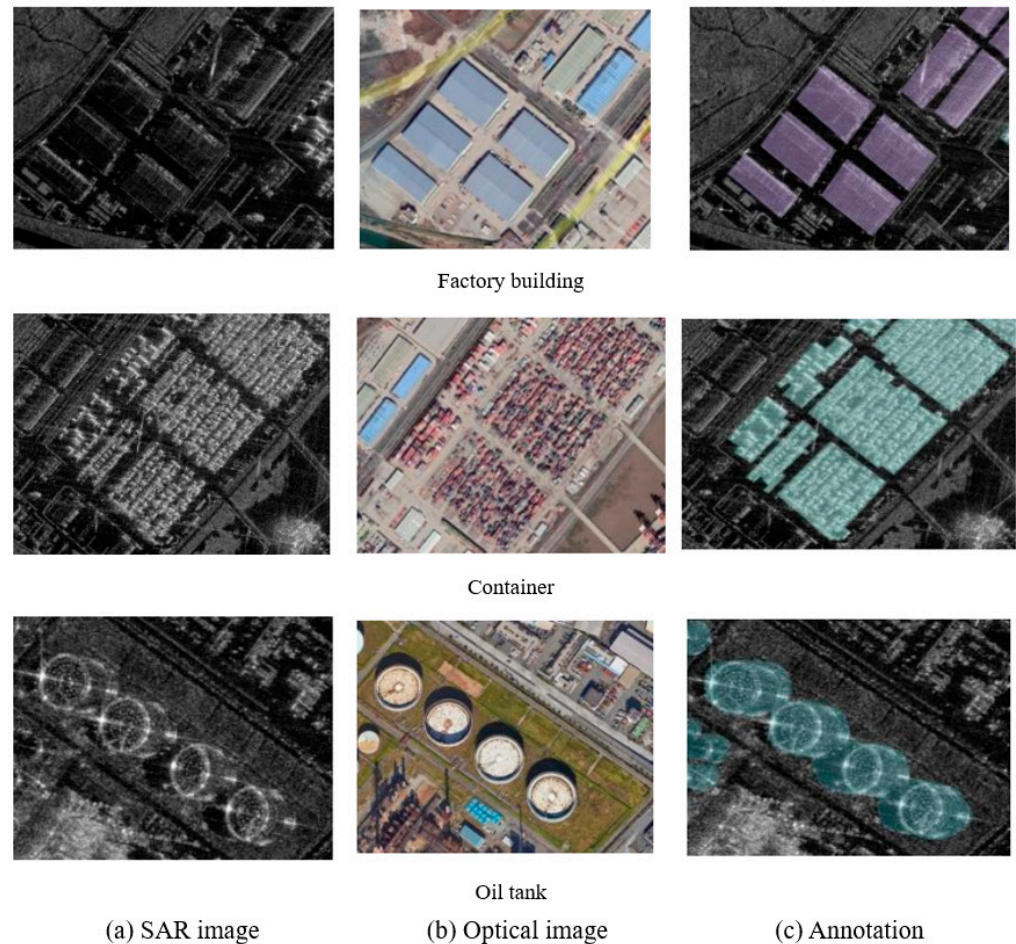


Figure 5. Examples of typical building annotations in the port area.

4.2. Experimental Setting

To enhance the model's robustness and generalization while minimizing the risk of overfitting, the training samples were preprocessed using techniques such as scaling, cropping, flipping, and image normalization. Multi-scale scaling and flipping were also applied to the test data for augmentation. Specifically, training samples were randomly scaled within a range of $[0.5, 2]$, with a maximum threshold of 0.75 for the proportion of pixels per class in the cropped patches, and a random flipping probability of 0.5 was applied.

The experiments were conducted using the Pytorch deep learning framework and the MMSegmentation library (<https://github.com/open-mmlab/mmdetection>, accessed on 13 July 2024). The training iteration number was set to 80 k, with AdamW as the optimizer. The cross-entropy loss function was employed. The initial learning rate was set to 0.0001, with a weight decay of 0.0001, and a multi-stage learning rate decay strategy was employed to dynamically adjust the learning rate, ensuring stable model convergence. Additionally, a linear warm-up strategy was applied during the initial training phase to further enhance model stability and convergence. The encoders of the model were initialized with pre-trained weights, and the batch size was set to 16, 12, or 8 per iteration, depending on memory constraints, to balance training efficiency and model performance. GFLOPs were calculated based on an input size of 512×512 to assess the computational complexity

of the model. All experiments were run on a Lenovo Thinkstation, which is made of China, equipped with two NVIDIA GeForce RTX 3090Ti GPUs running Ubuntu 18.04 operating systems, utilizing single-machine dual-card distributed training to improve training speed and effectiveness. For inference, a sliding window of 512×512 resolution with a 50% overlap was used to handle different SAR image sizes, matching the training data dimensions. Through meticulous parameter settings, the experiment achieved exceptional segmentation performance while maintaining computational efficiency.

Table 2 presents the main parameters of the backbone network in SPformer. The pyramid backbone network adheres to two principles inspired by ResNet [29]: firstly, as network depth increases, the hidden dimension gradually expands while the output resolution gradually decreases; secondly, computational resources are mainly allocated to the third stage of the backbone network.

Table 2. Configuration details of the backbone of SPformer.

Stage	Output Size	Layer	Parameter
Stage 1	128×128	Patch embedding	$P_1 = 4; C_1 = 64$
		Transformer encoder with PEG	$\begin{bmatrix} LSA \\ GSA \end{bmatrix} \times 1$
Stage 2	64×64	Patch embedding	$P_2 = 2; C_2 = 128$
		Transformer encoder with PEG	$\begin{bmatrix} LSA \\ GSA \end{bmatrix} \times 1$
Stage 3	32×32	Patch embedding	$P_3 = 2; C_3 = 256$
		Transformer encoder with PEG	$\begin{bmatrix} LSA \\ GSA \end{bmatrix} \times 5$
Stage 4	16×16	Patch embedding	$P_4 = 2; C_4 = 512$
		Transformer encoder with PEG	$[GSA] \times 4$

Where P_i is patch size, which refers to the size of the image divided into patch tokens, and then embedded into a one-dimensional vector by projection; C_i is the dimension of the output feature of each layer. The input feature size of stage four is 32×32 , with relatively lower computational requirements. A transformer encoder module based on global attention, namely GSA, is adopted for feature extraction.

4.3. Evaluation Metric

To quantitatively evaluate the results, five indices are selected as evaluation metrics: the mean intersection over union (mIoU), mean F1 score (mF1), billion floating point operations per second (GFLOPs), model parameters (Params), and frame per second (FPS). IoU and F1 scores are the two most commonly used accuracy evaluation metrics in semantic segmentation. IoU measures the overlap between predicted and true pixel labels, it is formally defined as:

$$IoU = \frac{T_p}{T_p + F_n + F_p} \quad (7)$$

where T_p represents the number of correctly predicted positive samples, F_p represents the number of negative samples predicted as positive, and F_n represents the number of positive samples predicted as negative. The F1 score calculates the similarity between the predicted and true pixel categories. It is defined as:

$$F1 = \frac{2 * T_p}{(T_p + F_n) + (T_p + F_p)} \quad (8)$$

FPS represents the speed of processing, formally defined as:

$$FPS = \frac{N}{T} \quad (9)$$

where N is the number of samples in the test set, and T is the time required for processing on the test set.

4.4. Semantic Segmentation Results for Quantitative Comparison

4.4.1. Comparison of Different Models

In order to evaluate the effectiveness of the proposed method for classifying typical buildings in high-resolution SAR images of port areas, comparison experiments were conducted using mainstream models based on a CNN and transformer. The selected models and the comparison results are displayed in Table 3. To ensure a fair comparison of the efficiency of SPformer, pyramid-structured transformer networks Swin, Twins-PCPVT, and Poolformer employ the commonly used FPN decoder [30]. Additionally, the appropriate network size is selected to minimize parameter discrepancies between different models.

Table 3. Comparison of segmentation results of different models on our constructed dataset.

Model	Backbone	Architecture	mIoU	mF1	Per-Class IoU			GFLOPs	Params (M)
					Fact.	Cont.	oilT.		
Segmenter	ViT-B	Trans. Enc.	71.51	83.30	67.01	69.45	78.06	69.6	102.39
FCN	Resnet50	CNN Enc.	74.48	85.26	67.08	76.04	80.33	197.69	49.49
HRNet	HRNet-w32	CNN Enc.	74.67	83.57	69.21	74.95	79.84	90.96	29.54
PoolFormer	Pool-S36	Trans. Enc.	75.11	85.71	70.21	73.86	81.25	47.3	34.59
Twins-PCPVT	PCPVT-S	Trans. Enc.	75.46	85.96	71.16	74.54	80.70	36.99	27.87
DeeplabV3+	Resnet50	CNN Enc.	75.71	86.05	68.60	76.41	82.11	176.23	43.58
PSPNet	Resnet50	CNN Enc.	75.96	86.22	68.63	76.80	82.45	178.45	48.96
Swin	Swin-T	Trans. Enc.	76.44	86.59	71.47	76.73	81.11	46.08	31.88
SPformer	SVT-S	Trans. Enc.	77.14	87.04	72.59	76.84	81.99	31.84	25.1

As shown in Table 3, SPformer achieves the highest classification performance for building classification in port areas, with mIoU and mF1 of 77.14% and 87.04%, respectively. Among the three types of buildings in high-resolution SAR images, factory classification poses the most significant challenge due to the larger size of factories, greater intra-class and contextual variations, and ambiguous boundaries at both the single convolution kernel and image patch levels. These factors contribute to poor pixel classification accuracy. Compared to CNN models, models employing a pyramid-structured transformer as the encoder, such as PoolFormer [31], Twins-PCPVT, Swin, and SPformer, demonstrate superior pixel classification accuracy in the factory category (as in Table 3). The mIoU for the factory category exceeds 70% for these models. This improvement can be attributed to the transformer's ability to capture the long-range intra-class and inter-class dependencies, as well as global semantic information of objects. Consequently, they excel in capturing the high-level semantic information regarding how the basic elements of factories, such as double bounce lines, layovers, and shadows, are related to form complete factory features.

Figure 6 visually depicts the classification performance of four models, namely SPformer, Twins-PCPVT, HRNet [32], and PSPNet. Notably, SPformer exhibits clearer and more accurate pixel classification results, particularly in factory and container scenes. It excels in extracting more complete ground structures and capturing precise edge details.

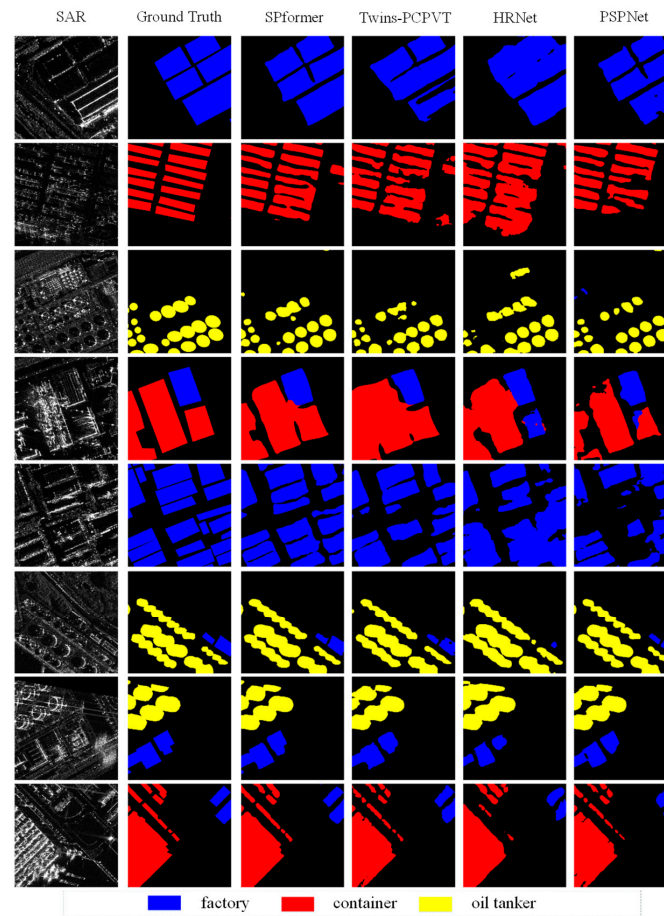


Figure 6. Examples of classification results with different models.

4.4.2. Comparison of Different Decoders

SPformer employs a lightweight ALL-MLP decoder, whose effectiveness and low computational complexity were validated through comparison experiments with other mainstream networks. To further assess the effectiveness of the lightweight ALL-MLP decoder in the pyramid transformer encoder, this study integrates the decoder of the common UPerNet [33] from the Swin transformer into the Swin, Twins-PCPVT, and SPformer networks. Table 4 shows the results of the six comparative experiments, demonstrating that under the same decoder configuration, the main network Twin-SVT adopted by SPformer exhibits superior performance advantages, despite a slight reduction in inference speed. In terms of performance, the use of a lightweight MLP decoder increases the mIoU by 2.16% compared to using the UPerNet decoder, while doubling the inference speed. This underscores the high performance and low computational complexity of the SPformer model, which is particularly critical for processing high-resolution images of remote sensing or urban scenes.

Table 4. Comparison of segmentation results of pyramid transformer networks with encoders of different size.

Model	Decoder	mIoU	GFLOPs	Params	FPS
Swin	Upernet	74.01	234.86	59.83	24.28
	All-MLP	75.65	27.69	27.95	55.01
Twins-PCPVT	Upernet	74.29	225.26	53.97	22.25
	All-MLP	75.14	34.11	25.17	40.97
SPformer	Upernet	74.98	224.22	53.86	21.05
	All-MLP	77.14	31.84	25.1	44.31

4.5. Semantic Segmentation Results of Whole-Scene SAR Images

To further investigate the classification effectiveness of the proposed SPformer model in the application of high-resolution SAR image classification of typical buildings in port areas, whole-scene SAR images of Qingdao Port and Yokohama Port in Japan were selected for end-to-end testing.

4.5.1. Classification Results of SAR Image of Qingdao Port

Figure 7a shows the classification results of typical buildings in SAR images of the Qingdao port area. This area is characterized by a higher concentration of factory buildings and fewer oil tanks. To compare the classification results of large-scene SAR images, the PSPNet model, representing the good performance among the CNN model, and the Twins-PCPVT model, representing the transformer models, were selected.

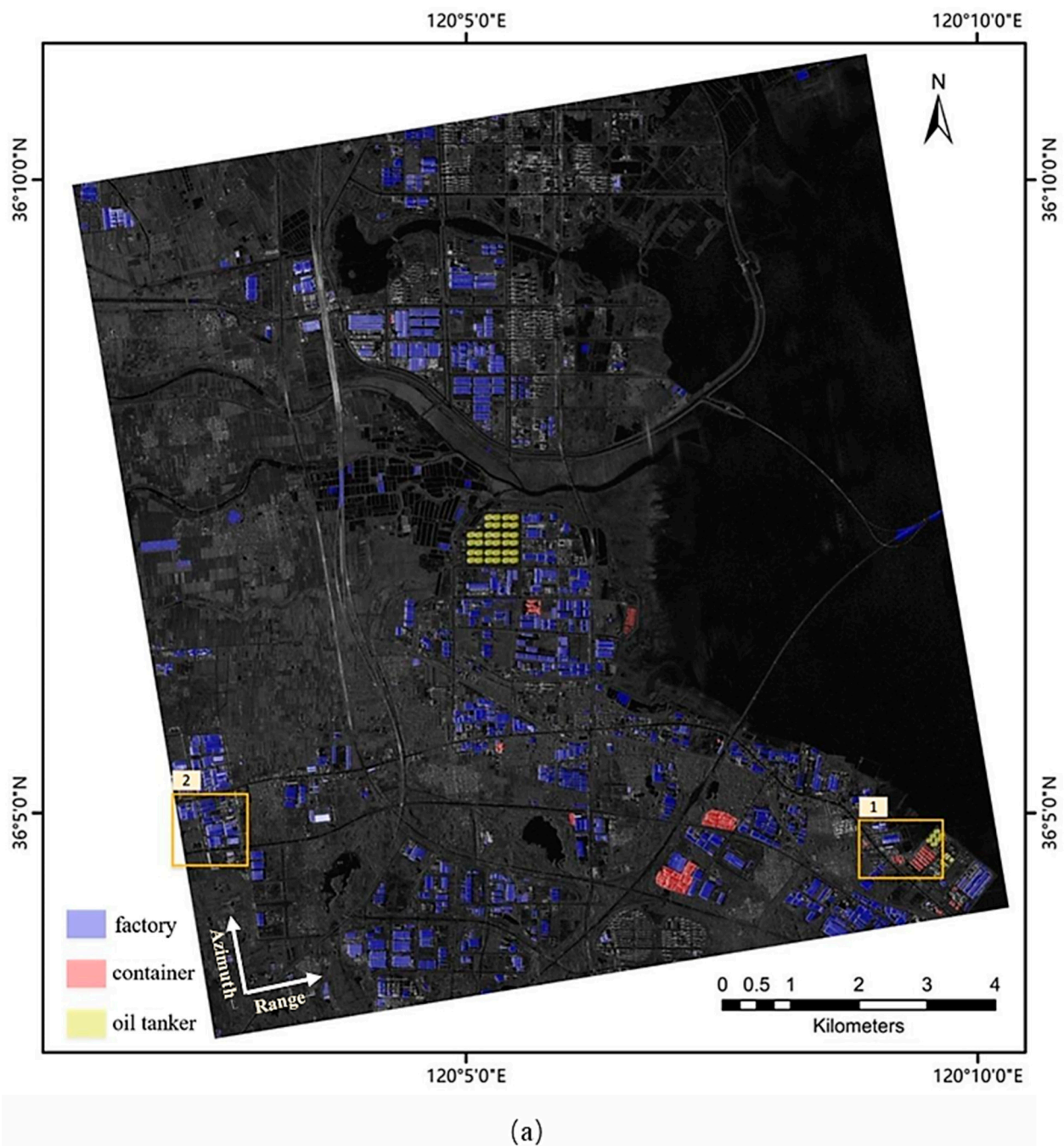


Figure 7. Cont.

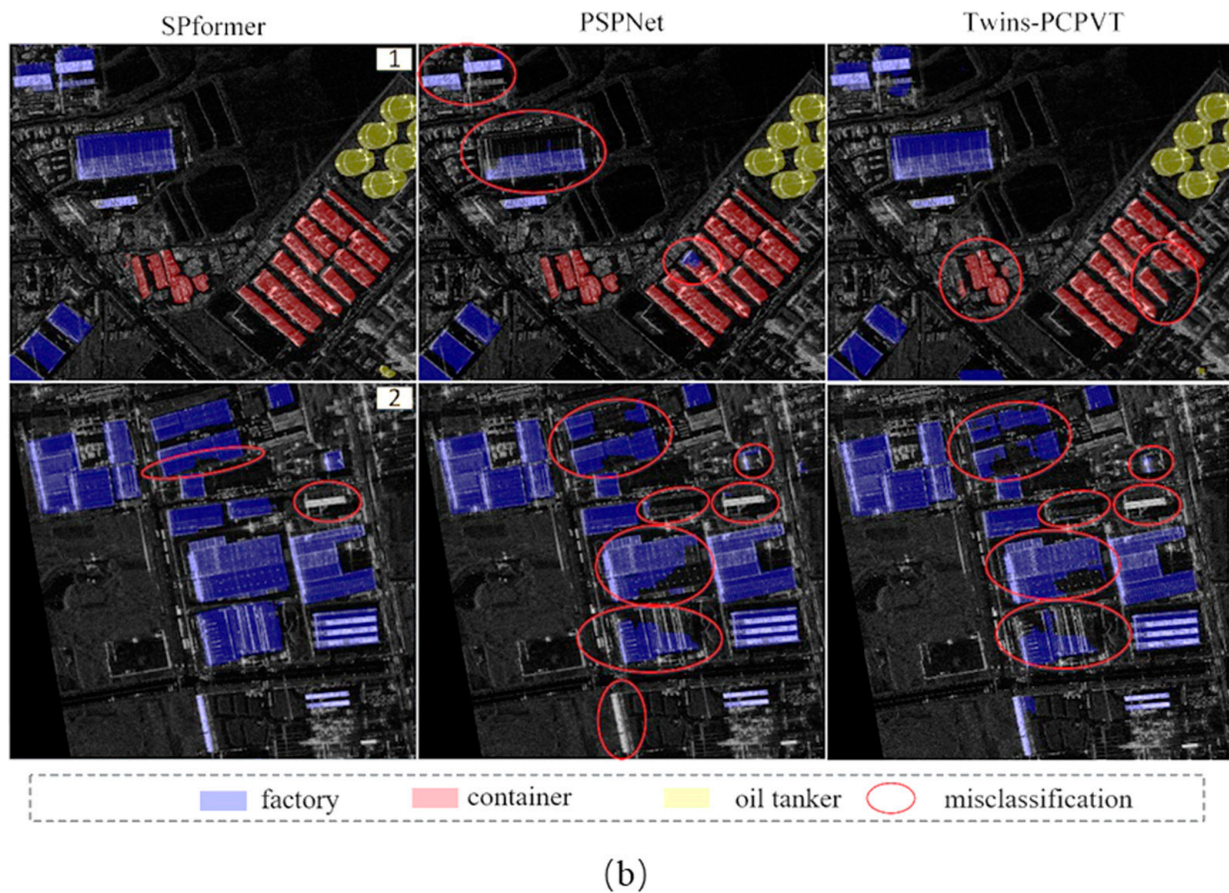


Figure 7. The classification results of the SAR image of Qingdao Port. (a) Classification results with the whole SAR image scene by SPformer network. (b) Zoom-in classification results of three networks in the numbered 1 and 2 experimental areas.

Figure 7b zooms in on the building classification results of the three models within rectangular boxes numbered 1 and 2 in Figure 7a. The red circle highlights the area where pixel errors in classification occur. Both PSPNet and Twins-PCPVT exhibit a significant issue of missing structure in the classification of factory buildings, where pixels corresponding to factory buildings are misclassified as background or other land cover types. In contrast, the proposed SPformer method demonstrates superior capability in building classification, providing more accurate and detailed results.

4.5.2. Classification Results of SAR Image of Yokohama Port

Figure 8a depicts the classification results of buildings in whole-scene SAR images of the Yokohama Port area in Japan. This area exhibits a higher density of oil tanks and containers, with factory buildings mainly distributed along the coastline. Figure 8b zooms in on the building classification results of the SPformer, PSPNet, and Twins-PCPVT models within the rectangular box areas numbered 1 and 2 in Figure 8a, respectively. SPformer demonstrates greater accuracy in factory classification, extracting more complete building structures and edges. These results underscore the practical application potential and value of the proposed method for building classification in large-scene high-resolution SAR images of port areas.

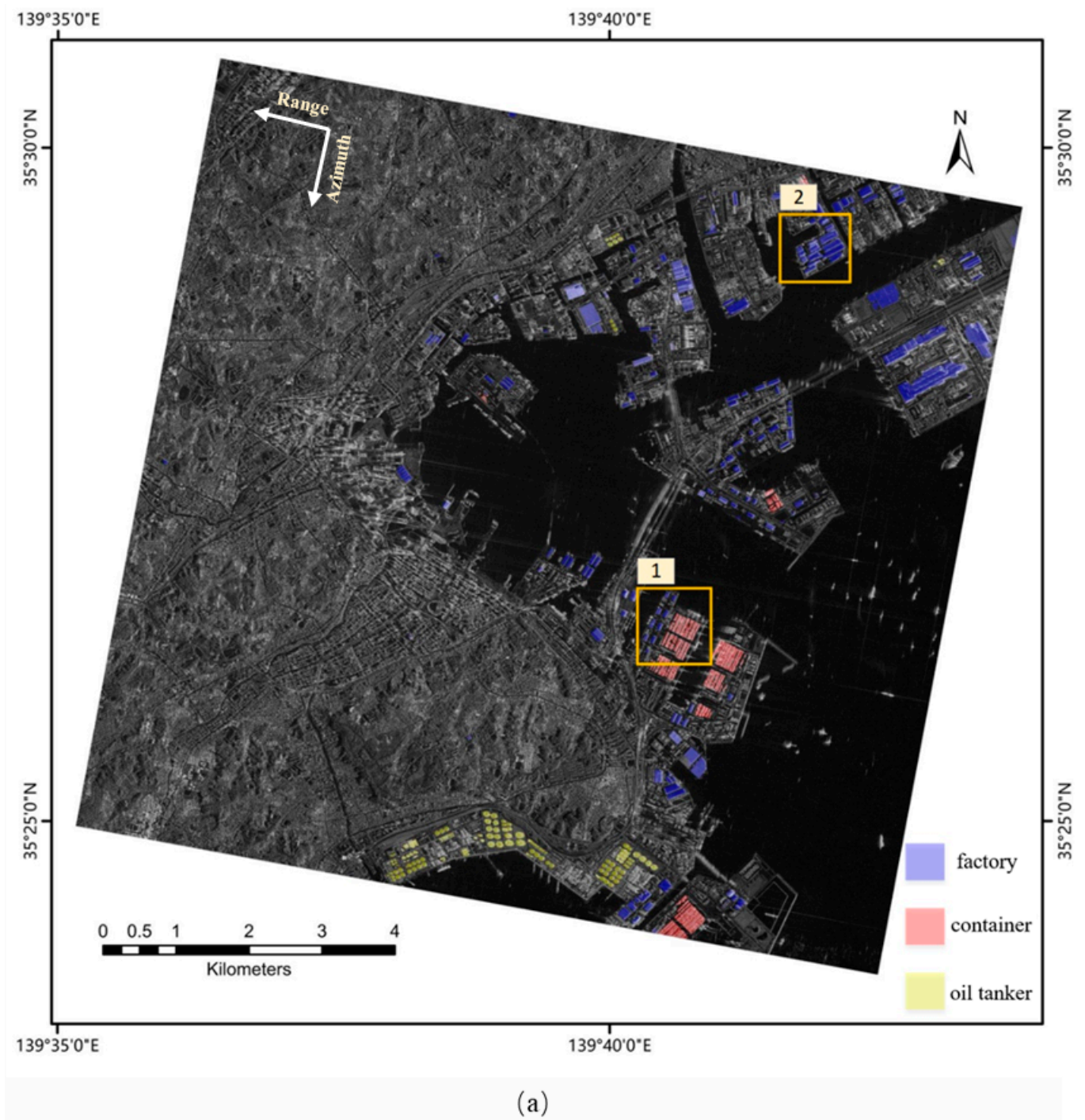


Figure 8. Cont.

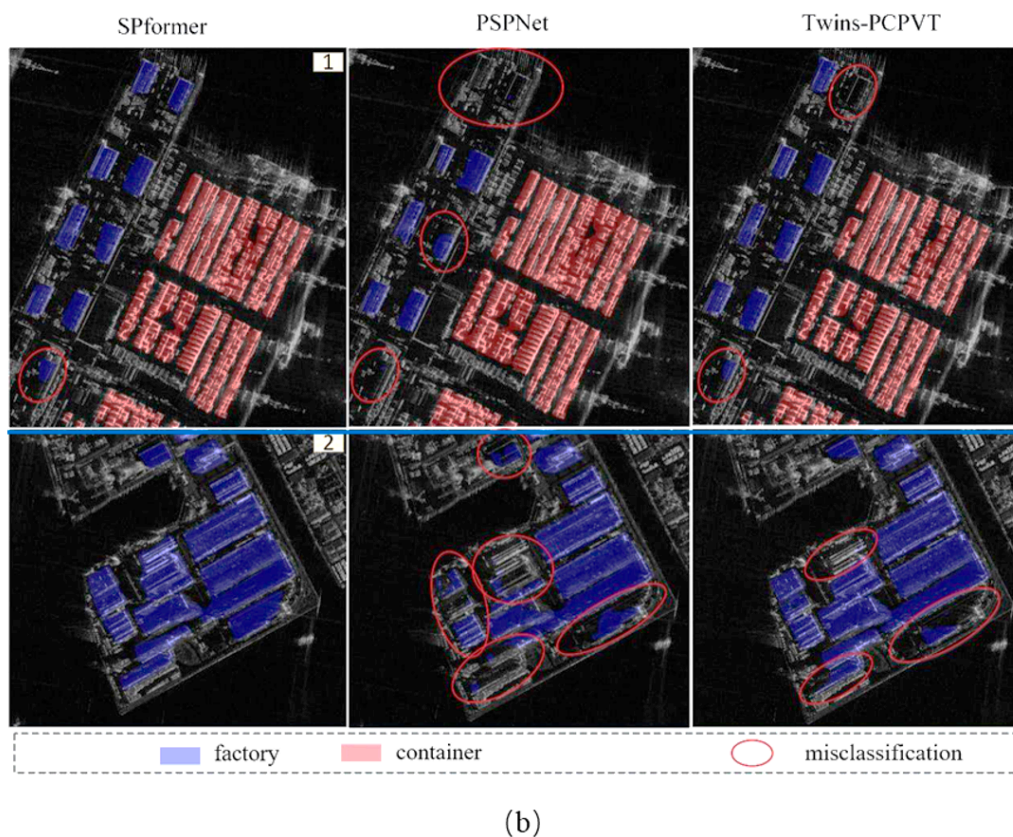


Figure 8. The classification results of SAR images of Yokohama Port. (a) Classification result of typical buildings in SAR images of the Yokohama Port area by the SPformer network. (b) Zoom-in classification results of three networks in the numbered 1 and 2 experimental areas.

5. Discussion

In this paper, we constructed a multi-class semantic segmentation dataset based on 1 m-resolution GF-3 SL mode SAR images. The dataset includes three typical buildings commonly found in port areas: factory buildings, containers, and oil tanks. We then conducted a classification task of those buildings in port areas in high-resolution SAR images. The SPformer model, proposed in this study, combines pyramid transformer encoding technology with a lightweight multi-level feature aggregation MLP decoding method. This integration allows for the extraction of complete building structures and edge details, resulting in higher accuracy in building classification. Specifically, the pyramid transformer encoding technology, which employs a spatially separable attention mechanism, enhances the transformer's local feature extraction capability by grouping attention locally. Additionally, it improves information interaction between local windows through globally subsampled attention, thus enhancing the extraction of global semantic information, and improving the integrity of building structure extraction.

The lightweight MLP decoding method can efficiently aggregate multi-level features, enhancing the interaction between low-level spatial information and high-level semantic information. This improvement enhances the model's ability to process details and extract the edge of a building. From the perspective of actual deployment capability and detection performance, SPformer demonstrates superior classification accuracy while maintaining fewer model parameters and lower calculation costs. Comparison experiments of different decoders, as shown in Table 4, validate that compared to the UPerNet decoder with a more complex configuration, SPformer's lightweight MLP decoder with a simpler configuration effectively improves model performance and efficiency. The mIoU value is increased by 2.16%, and inference speed is doubled. In addition, results of building classification on large-scale SAR images show that SPformer can better deal with the problems of structure

loss and edge ambiguity in building classification within port areas, showcasing significant application potential.

While this paper made significant results in the classification of port areas using GF-3 1 m resolution SAR images, there remain challenges due to the complex imaging mechanism of SAR sensors and the diverse scattering characteristics of buildings. One issue lies in the use of footprint vector data as auxiliary data for labeling buildings in high-resolution SAR images. This approach suffers from a low coincidence between the label and the complete scattering feature area of buildings, resulting in limited model effectiveness. Additionally, the visual interpretation method employed in this paper relies on expert experience, and the labeling process covers the complete scattering feature area of buildings. However, due to factors such as subjectivity, negligence, and inconsistency in manual labeling, errors may occur, and the labeling process can be costly. Exploring an automatic and efficient method of generating a building sample dataset is crucial for the application of high-resolution SAR image building classification.

Secondly, the mIoU value on the building classification by SPformer is 77.14%. Specifically, the mIoU values for factory buildings, containers and oil tanks are 72.59%, 76.84%, and 81.99%, respectively. As illustrated in Figure 9, buildings with small size and dense distribution areas lacking obvious scattering features, or exhibiting significant scattering aliases with other land cover types, are prone to being missed in classification. For instance, when the imaging features of nearshore factory buildings are obscured by the scattering features of surface objects such as cranes and ships, erroneous classification is likely to occur.

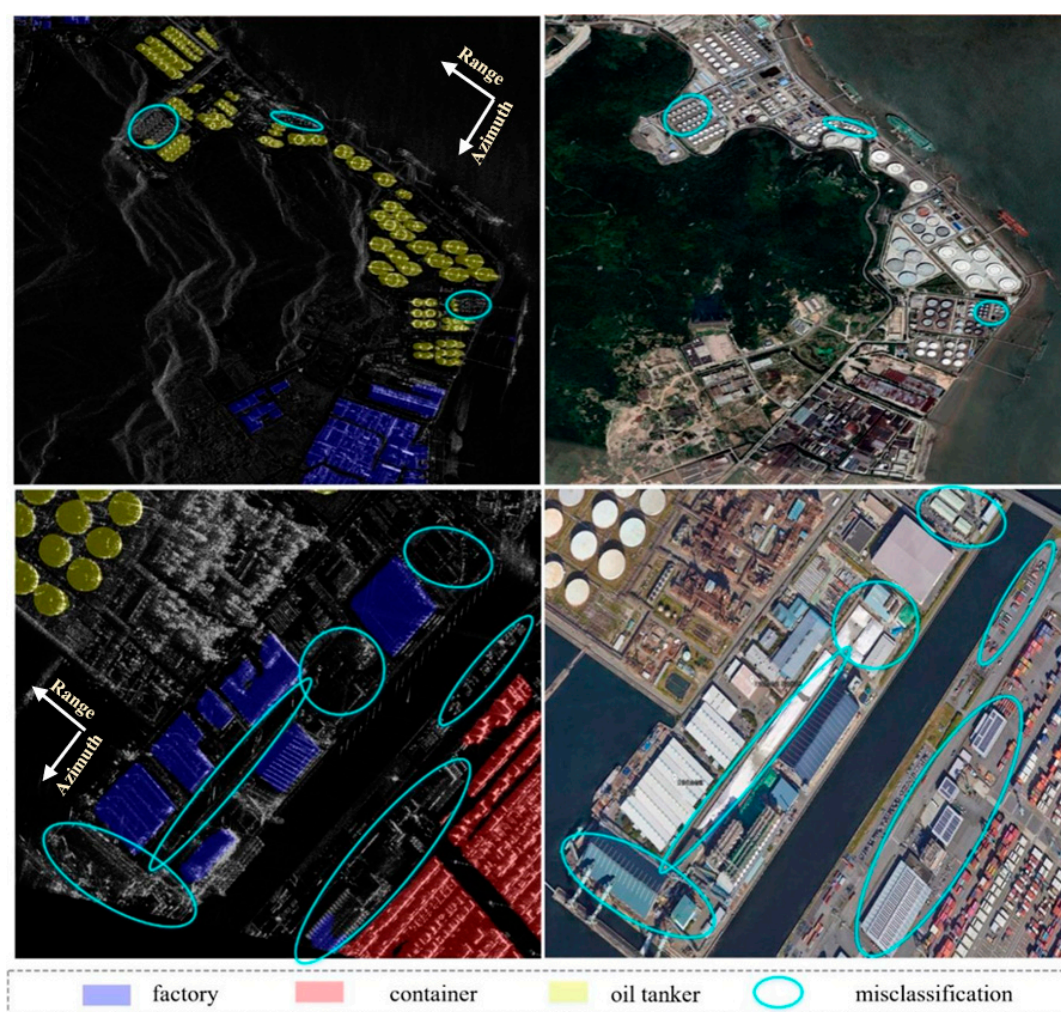


Figure 9. Examples of feature-prone missed detection area.

6. Conclusions

To address the issues of structural loss and boundary ambiguity that often occur in building classification, this letter proposes a lightweight and efficient SAR image semantic segmentation model called SPformer. The network introduces a pyramid transformer encoder based on spatially separable self-attention to enhance local and global spatial information encoding, as well as multi-scale feature extraction capabilities.

This enhancement improves the integrity of building structure extraction. Additionally, a lightweight MLP decoding method tailored for the pyramid transformer encoder is proposed. It combines output features from the encoder at different levels and areas of interest to obtain a segmented representation containing both local details and global context. This approach not only improves the model's ability to process details and the accurately extract building edges, but also significantly reduces model complexity and the demand for computational resources. Experimental results demonstrate that SPformer can extract more complete building structures and edge details while maintaining a compact model size and lower computational complexity in typical building classification tasks involving high-resolution SAR images in port areas, with mIoU and mF1 indicators of 77.14% and 87.04%, respectively.

Author Contributions: Conceptualization, B.Z., C.W. and F.W.; methodology, Q.W., B.Z. and J.H.; software, Q.W. and J.H.; validation, J.H. and Q.W.; formal analysis, B.Z.; investigation, Q.W. and B.Z.; resources, B.Z., C.W. and F.W.; data curation, B.Z. and F.W.; writing—original draft preparation, J.H. and Q.W.; writing—review and editing, B.Z., C.W., F.W. and J.H.; visualization, Q.W. and J.H.; supervision, B.Z. and F.W.; project administration, B.Z., C.W. and F.W.; funding acquisition, B.Z., C.W. and F.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 42201503), the Hainan Provincial Natural Science Foundation of China (No. 622MS104), the Key Program of National Natural Science of China (No. 41930110), and the Science and Disruptive Technology Program, AIRCAS (No. 2024-AIRCAS-SDTP-11).

Data Availability Statement: The data presented in this study are available on request from the first author.

Acknowledgments: The authors would like to thank the China Centre for Resource Satellite Data and Application for providing the GF-3 SAR data. Zhen Li offered valuable suggestions regarding the experiments and provided the necessary funding. Our sincere thanks also extend to the anonymous reviewers and editors for their insightful comments and constructive suggestions.

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

References

1. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for Semantic Segmentation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 7242–7252.
2. Li, Z.; Zhang, W.; Pan, J.; Sun, R.; Sha, L. A Super-Resolution Algorithm Based on Hybrid Network for Multi-channel Remote Sensing Images. *Remote Sens.* **2023**, *15*, 3693. [\[CrossRef\]](#)
3. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
4. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
5. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
6. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
7. Khan, S.; Naseer, M.M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in Vision: A survey. *ACM Comput. Surv.* **2022**, *15*, 1–41. [\[CrossRef\]](#)
8. Naseer, M.M.; Ranasinghe, K.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M. Intriguing Properties of Vision Transformers. *arXiv* **2015**, arXiv:2015.10497.

9. 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* **2021**, arXiv:2010.11929.
10. Wang, W.; Xie, E.; Li, X.; Fan, D.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 548–558.
11. Wang, R.; Lei, T.; Cui, R.; Zhang, B.; Meng, H.; Nandi, A.K. Medical Image Segmentation Using Deep Learning: A Survey. *IET Image Process.* **2022**, *16*, 1243–1267. [[CrossRef](#)]
12. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
13. 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.
14. Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; Shen, C. Twins: Revisiting the Design of Spatial Attention in Vision Transformers. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2021**, *34*, 9355–9366.
15. Minaee, S.; Boykov, Y.Y.; Porikli, F.; Plaza, A.J.; Kehtarnavaz, N.; Terzopoulos, D. Image Segmentation Using Deep Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3523–3542. [[CrossRef](#)]
16. Wu, F.; Wang, C.; Zhang, H.; Li, J.; Li, L.; Chen, W.; Zhang, B. Built-up Area Mapping in China from GF-3 SAR Imagery Based on the Framework of Deep Learning. *Remote Sens. Environ.* **2021**, *262*, 112515. [[CrossRef](#)]
17. Li, T.; Wang, C.; Wu, F.; Zhang, H.; Tian, S.; Fu, Q.; Xu, L. Built-Up Area Extraction from GF-3 SAR Data Based on a Dual-Attention Transformer Model. *Remote Sens.* **2022**, *14*, 4182. [[CrossRef](#)]
18. Wu, Z.; Hou, B.; Guo, X.; Ren, B.; Li, Z.; Wang, S.; Jiao, L. CCNR: Cross-Regional Context and Noise Regularization for SAR Image Segmentation. *Int. J. Appl. Earth Obs.* **2023**, *121*, 103363. [[CrossRef](#)]
19. Kang, J.; Wang, Z.; Zhu, R.; Xia, J.; Sun, X.; Fernandez-Beltran, R.; Plaza, A. DisOptNet: Distilling Semantic Knowledge from Optical Images for Weather-independent Building Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [[CrossRef](#)]
20. Wangiyana, S.; Samczyrski, P.; Gromek, A. Data Augmentation for Building Footprint Segmentation in SAR Images: An Empirical Study. *Remote Sens.* **2022**, *14*, 2012. [[CrossRef](#)]
21. Yue, Z.; Gao, F.; Xiong, Q.; Wang, J.; Hussain, A.; Zhou, H. A Novel Attention Fully Convolutional Network Method for Synthetic Aperture Radar Image Segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4585–4598. [[CrossRef](#)]
22. Sun, Y.; Hua, Y.; Mou, L.; Zhu, X.X. CG-Net: Conditional GIS-Aware Network for Individual Building Segmentation in VHR SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [[CrossRef](#)]
23. Xia, J.; Yokoya, N.; Adriano, B.; Zhang, L.; Li, G.; Wang, Z. A Benchmark High-Resolution GaoFen-3 SAR Dataset for Building Semantic Segmentation. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* **2021**, *14*, 5950–5963. [[CrossRef](#)]
24. Shi, X.; Fu, S.; Chen, J.; Wang, F.; Xu, F. Object-Level Semantic Segmentation on the High-Resolution Gaofen-3 FUSAR-Map Dataset. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* **2021**, *14*, 3107–3119. [[CrossRef](#)]
25. Shermeyer, J.; Hogan, D.; Brown, J.; Van Etten, A.; Weir, N.; Pacifici, F.; Hansch, R.; Bastidas, A.; Soenen, S.; Bacastow, T.; et al. SpaceNet 6: Multi-Sensor All Weather Mapping Dataset. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 768–777.
26. Wu, Q.; Zhang, B.; Xu, C.; Zhang, H.; Wang, C. Dense Oil Tank Detection and Classification via YOLOX-TR Network in Large-Scale SAR Images. *Remote Sens.* **2022**, *14*, 3246. [[CrossRef](#)]
27. Chu, X.; Tian, Z.; Zhang, B.; Wang, X.; Shen, C. Conditional Positional Encodings for Vision Transformers. *arXiv* **2023**, arXiv:2102.10882.
28. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. Segformer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2021**, *34*, 12077–12090.
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
30. Kirillov, A.; Girshick, R.; He, K.; Dollár, P. Panoptic Feature Pyramid Networks. In Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
31. Yu, W.; Luo, M.; Zhou, P.; Si, C.; Zhou, Y.; Wang, X.; Feng, J.; Yan, S. Metaformer Is Actually What You Need for Vision. *arXiv* **2022**, arXiv:2111.11418.
32. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
33. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified Perceptual Parsing for Scene Understanding. *arXiv* **2018**, arXiv:1807.10221.

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.