



Article

BSDSNet: Dual-Stream Feature Extraction Network Based on Segment Anything Model for Synthetic Aperture Radar Land Cover Classification

Yangyang Wang ^{1,2}, Wengang Zhang ², Weidong Chen ^{1,*} and Chang Chen ¹

¹ School of Information Science and Technology, University of Science and Technology of China, Hefei 230037, China; wangyy@nudt.edu.cn (Y.W.); chench@ustc.edu.cn (C.C.)

² Electronic Countermeasure Institute, National University of Defense Technology, Hefei 230037, China; zhangwg@nudt.edu.cn

* Correspondence: wdchen@ustc.edu.cn

Abstract: Land cover classification using high-resolution Polarimetric Synthetic Aperture Radar (PolSAR) images obtained from satellites is a challenging task. While deep learning algorithms have been extensively studied for PolSAR image land cover classification, the performance is severely constrained due to the scarcity of labeled PolSAR samples and the limited domain acceptance of models. Recently, the emergence of the Segment Anything Model (SAM) based on the vision transformer (ViT) model has brought about a revolution in the study of specific downstream tasks in computer vision. Benefiting from its millions of parameters and extensive training datasets, SAM demonstrates powerful capabilities in extracting semantic information and generalization. To this end, we propose a dual-stream feature extraction network based on SAM, i.e., BDSNet. We change the image encoder part of SAM to a dual stream, where the ConvNext image encoder is utilized to extract local information and the ViT image encoder is used to extract global information. BDSNet achieves an in-depth exploration of semantic and spatial information in PolSAR images. Additionally, to facilitate a fine-grained amalgamation of information, the SA-Gate module is employed to integrate local–global information. Compared to previous deep learning models, BDSNet’s impressive ability to represent features is akin to a versatile receptive field, making it well suited for classifying PolSAR images across various resolutions. Comprehensive evaluations indicate that BDSNet achieves excellent results in qualitative and quantitative evaluation when performing classification tasks on the AIR-PolSAR-Seg dataset and the WHU-OPT-SAR dataset. Compared to the suboptimal results, our method improves the Kappa metric by 3.68% and 0.44% on the AIR-PolSAR-Seg dataset and the WHU-OPT-SAR dataset, respectively.

Keywords: Polarimetric Synthetic Aperture Radar; land cover classification; Segment Anything Model



Citation: Wang, Y.; Zhang, W.; Chen, W.; Chen, C. BDSNet: Dual-Stream Feature Extraction Network Based on Segment Anything Model for Synthetic Aperture Radar Land Cover Classification. *Remote Sens.* **2024**, *16*, 1150. <https://doi.org/10.3390/rs16071150>

Academic Editor: Kevin Tansey

Received: 7 February 2024

Revised: 18 March 2024

Accepted: 22 March 2024

Published: 26 March 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

Precise land cover classification data play a vital role in regional development, such as natural resource management [1] and agricultural planning [2]. Numerous land classification tasks have relied on conventional optical imagery [3]. Nevertheless, the acquisition of high-quality optical images proves challenging due to weather-related limitations, such as clouds and fog. In light of this, PolSAR imagery has emerged as a pivotal and innovative data source for land classification. Capitalizing on its superior penetration capabilities and the ability to measure backscattering irrespective of weather conditions, PolSAR has emerged as a pivotal instrument for surmounting the limitations imposed by adverse weather conditions. Nonetheless, the energy reflected from ground objects is impacted by both the surface texture and nearby terrain features, its intrinsic morphology, and the incidence angle of the PolSAR sensor [4]. These uncertainties constrain the classification accuracy of PolSAR imagery in land cover classification tasks, particularly in complex terrain

areas. This complexity poses challenges in the identification and regional quantification of land cover features. Furthermore, the availability of datasets in the PolSAR domain is limited, further impeding the development of object classification in the field.

Existing PolSAR land cover classification methods can be categorized into traditional approaches [5,6] and deep learning methods [7–9]. Traditional classification methods are predominantly designed based on statistical features of PolSAR data, such as the Wishart distribution [10], etc. Xie et al. [5] introduced an enhanced Markov Random Field (MRF) model that combines polarimetric features with spatial information from SAR images using the Wishart distance and class confidence for interpreting polarimetric SAR data. Mishra et al. [11] proposed a machine learning decision tree classifier for land cover classification. Developed based on validated evidence and expert knowledge through experimental validation, this classifier has specific classification rules for each class due to their distinct scattering behaviors. However, these methods heavily rely on the accuracy of statistical models. Furthermore, the complexity of parameter estimation in statistical models, coupled with the complexity of land cover scattering, often leads to unsatisfactory classification results across different PolSAR datasets. For the last few years, land cover classification has seen a rise in the dominance of deep learning methods. This is attributable to their remarkable ability to extract features. Considering the pervasive triumph of convolutional neural networks (CNNs) in computer vision assignments, the majority of the deep learning approaches for PolSAR land cover classification are founded on CNNs. Zhou et al. [12] pioneered the use of a CNN with two convolutional layers and two fully connected layers for PolSAR land cover classification. Mei et al. [13] developed a neural network that contains several convolution layers (C-CNN) to assess its feature learning capabilities. With the increasing demand for handling graph-structured data, the GNN has gradually become a research hotspot, giving rise to numerous land classification studies based on the GNN. Kavran et al. [14] proposed a spatio-temporal land cover classification method based on graph neural networks. The updated approach now employs a graph neural network as a multispectral image feature extraction network and utilizes a convolutional neural network as the node classification model within its modular node classification pipeline. Zhao et al. [15] proposed a U-shaped Object Graph Neural Network, primarily composed of Self-Adaptive Graph Construction (SAGC), a hierarchical graph encoder, and a decoder. By inputting depth features extracted from convolution and multi-layer attention operations, it generates context-aware graph structures to predict land cover types. Additionally, Fang et al. [16] proposed a fully convolutional network (FCN), which exhibited an excellent classification performance. Subsequently, the emergence of Unet [17] revolutionized land cover classification tasks, attaining a superior performance across nearly all categories of land cover. The encoder–decoder framework remains widely favored in current applications.

Owing to the limited scope of convolutional operations, it encountered challenges in efficiently capturing overarching and distant semantic information, giving rise to the application of transformers based on attention mechanisms [18]. Transformers have the capability to extract spatially relevant information globally, thus providing flexible feature representation capabilities. Dong et al. [19] investigated the utilization of fewer ViT layers for the classification of land use by PolSAR. Encouraging outcomes from the SViT underscored the effectiveness and viability of employing transformer architectures in the realm of PolSAR images. Wang et al. [20] employed a deeper vision transformer (ViT) while addressing the challenge of insufficient data by pretraining the ViT backbone using a masked autoencoder (MAE) [21]. Nevertheless, the utilization of the ViT's global attention mechanism substantially escalates the computational complexity and exhibits limited sensitivity to the extraction of local features. The introduction of ConvNext adeptly reconciles the strengths of convolution and the ViT. Achieving a performance comparable to the ViT, ConvNext [22] markedly reduces the model's parameter count.

The interpretation of SAR imagery poses unique challenges compared to optical imagery, primarily due to the complex scattering mechanisms that introduce coherent

speckle noise, causing difficulties in characterizing data features. These challenges result in inadequate or challenging feature extraction, thereby hindering the achievement of satisfactory classification accuracy. Recently, the introduction of the Segment Anything Model (SAM) [23] based on the ViT has elevated the accuracy of semantic segmentation tasks to a new level. Recent research indicates that it has shown impressive performances across various downstream tasks, attributable to its robust generalization capabilities and intricate architecture. This demonstrates its feasibility in PolSAR image land cover classification tasks. To fully leverage the raw information from PolSAR images and enhance the efficiency of feature extraction and utilization, we propose a dual-stream local–global feature fusion network. Specifically, we adopt the SAM architecture as the backbone framework, with a modification to its image encoder section to incorporate a dual-stream encoder. One branch of the network employs a pure convolutional network, i.e., ConvNext, to extract local features, preserving the local detailed texture characteristics of PolSAR images. The other branch utilizes the ViT-based network inherent in SAM to extract global features from PolSAR images, modeling global dependencies between regions. Additionally, the low-rank adaptation (LoRA) [24] module is added to the transformer architecture to enhance the lightweight nature of the model. Finally, a Separation-and-Aggregation Gate (SA-Gate) [25] is employed to integrate features from both local and global representations, facilitating multi-level data interaction.

Our main contributions are as follows:

- We propose a dual-stream local–global feature encoding network, BSDSNet. It utilizes SAM as the backbone framework and consists of a ConvNext image encoder dedicated to extracting local information and another ViT image encoder to capture global features.
- We introduce a novel low-rank fine-tuning strategy, LoRA, specifically tailored to approximate low-rank updates for the parameters within the ViT image encoder. This innovative approach is designed to significantly reduce the parameter count, effectively minimizing the computational overhead.
- We introduce the SA-Gate feature fusion module to achieve a precise integration of both local and global information. This module comprises the Feature Separation (FS) module, which calibrates single-stream features, and the Feature Aggregation (FA) module, enabling dual-stream feature fusion.

The subsequent sections of this paper have the following structure: Section 2 presents the specifics of the BSDSNet land cover classification algorithm. The test discourse is provided in Section 3. The concluding remarks are encapsulated in Section 4.

2. Materials and Methods

Figure 1 illustrates the proposed BSDSNet. This network adopts a conventional encoder–decoder architecture to accomplish the land cover classification task of PolSAR through the extraction and fusion of local–global features. In this section, we first introduce the encoder, composed of a dual-stream network, with one stream utilizing the ConvNext edge feature extraction module for local information extraction and another employing a SAM variant spatial feature extraction module for global information capture. Subsequently, the SA-Gate module, introduced for the fusion of these two streams, is presented. Finally, the structure of the decoder is detailed.

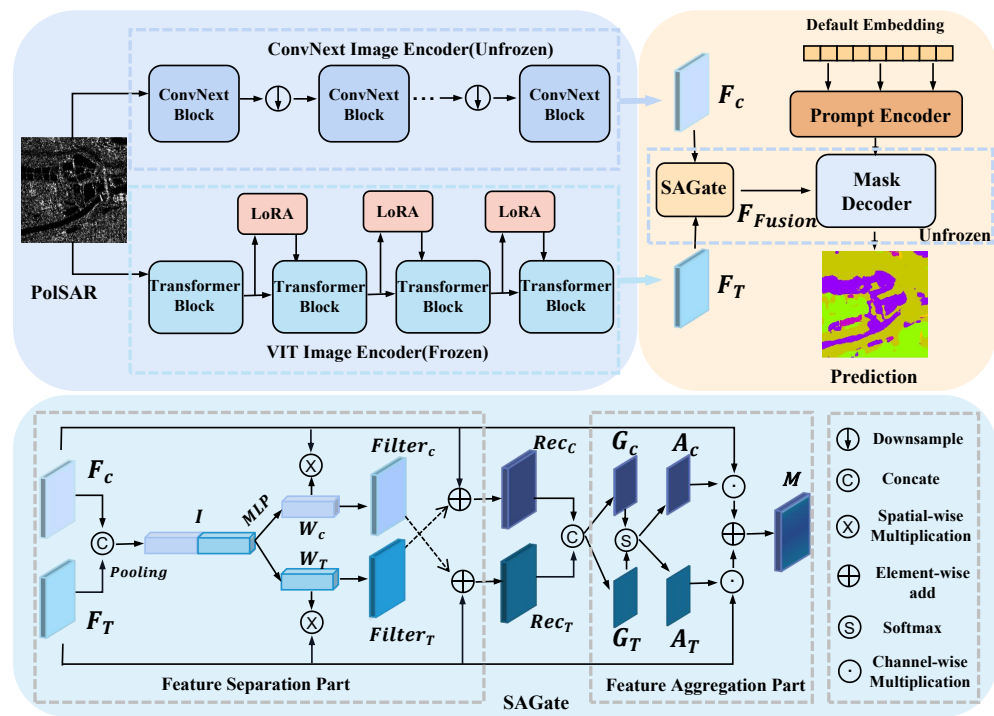


Figure 1. Overall architecture of our proposed method, BSDSNet. The network consists of a dual-stream encoder and single-stream decoder, with the ConvNext image encoder and VIT image encoder. The feature fusion module of the network also includes the SA-Gate and mask decoder. The weight parameters from the VIT image encoder are in a frozen state, while the ConvNext image encoder, SA-Gate, and mask decoder parts are trainable.

2.1. Local Feature Extraction

2.1.1. Overview

Due to the influence of PolSAR imaging mechanisms, the generated images may exhibit uneven brightness caused by multiplicative noise known as speckle noise. Speckle noise is challenging to eliminate as it is multiplicative in nature. This noise can be more pronounced in local regions, resulting in some areas being overly bright or dark [26]. To address this issue, we choose to employ a ConvNext network that focuses on extracting local features. This approach helps suppress or alleviate the impact of coherent speckle noise, enabling the network to better handle local brightness variations.

2.1.2. ConvNext

The framework diagram of the ConvNext image encoder is illustrated in Figure 1. It is designed based on the principles of the VIT and ResNet [27], constructing a network entirely composed of standard ConvNet modules. This design reduces the number of parameters and training data requirements compared to transformers while balancing accuracy and scalability. The entire ConvNext image decoder processes the input SAR image $I \in R^{H \times W \times 4}$ to generate a feature map containing rich local information. Specifically, the first convolution layer transforms the input image into the feature map $F \in R^{H \times W \times C}$. Then, the feature mapping F is processed through an encoder architecture, with each architecture comprising four stages. Each stage contains a ConvNext block and a downsampling operation. At each stage of the encoder, the image resolution is downsampled to one-half by the convolutional groups layer and the number of channels is doubled.

A detailed diagram of a ConvNext block is shown in Figure 2. Each block comprises three cascaded convolutions. The ConvNext block initially employs depthwise convolution [28]; the quantity of groups matches the quantity of channels. Depthwise convolution mixes only along the spatial dimension and therefore has fewer parameters, thus signifi-

cantly reducing the complexity and computational effort of the model. Subsequently, a layer normalization (LN) [29] layer is incorporated to facilitate a rapid network convergence and mitigate overfitting. Meanwhile, the second convolutional layer maps the feature vector to 4 times the spatial size; this operation is referred to as the reverse bottleneck structure. Then, the third convolutional layer reduces the information to the original spatial size. This strategic operation is designed to enhance the network's expressive capacity while maintaining its lightweight characteristics [30]. Following the second convolutional layer, there is a type of activation function known as GELU. Ref. [31] is introduced. In comparison to ReLU [32], GELU is a smoother activation function, contributing to the stable propagation of gradients. After the third convolutional layer, a layer scale is incorporated—a fundamentally trainable parameter used for scaling channel data. Finally, the drop path method is employed to reduce the coupling between features extracted by neurons and branches, effectively preventing overfitting. The inclusion of a shortcut connection further facilitates the training convergence.

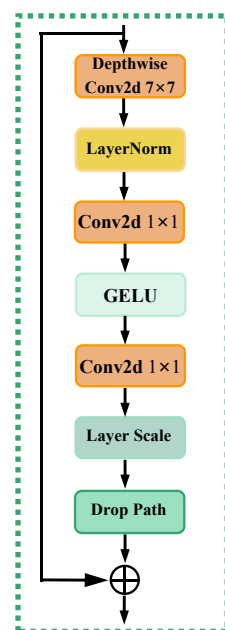


Figure 2. A flowchart of ConvNext block.

2.2. Global Feature Extraction

2.2.1. Overview

The regional information in PolSAR images exhibits variability, including factors such as the topography and land cover. Employing a network that emphasizes global feature extraction facilitates an enhanced adaptability to overarching background variations. This approach improves the model's versatility in accommodating diverse scenarios, thereby bolstering the robustness of image processing. Recently, SAM's outstanding semantic segmentation performance and strong robustness across multiple scenes have made it an indispensable method for processing images of diverse natural scenes. Having undergone training on a comprehensive dataset of 11 million images and 1.1 billion masks, SAM showcases a remarkable zero-shot performance in a spectrum of segmentation tasks. This indicates its capability to learn from new datasets and effectively leverage "prompting" techniques for novel tasks, even when having limited or no prior exposure to these tasks. SAM exhibits the ability to generalize widely across images, making it particularly appealing for remote sensing applications. SAMed [33] inherits the notable image segmentation performance of SAM and further refines segmentation boundaries. Additionally, by categorizing each segmented region into distinct meaningful organizations, SAMed achieves a thorough understanding of the semantic classes associated with each segmented area. This

improvement in SAMed further enhances its compatibility with remote sensing image land cover classification. Therefore, we opt to employ the SAMed model based on SAM, for the extraction of global information from PolSAR images and land cover classification.

SAMed mainly consists of the following two components:

- A VIT image encoder. The VIT excels in capturing global features compared to traditional CNNs. The VIT leverages the self-attention mechanism, enabling features at each position to interact globally with other positions. This enables our method to effectively grasp the overall contextual information within the image, rather than solely focusing on local intricacies.
- LoRA. This is a low-rank adaptation method proposed by Microsoft, which compresses feature maps to lower dimensions and then reflects them to the original dimension, thereby reducing the computational overhead.

2.2.2. VIT Image Encoder

The process of extracting global features using the VIT is outlined as follows. Initially, PolSAR images are segmented into multiple image blocks and embedded into feature vectors through linear projection. This positional embedding approach enhances the VIT's perception of information at different locations in the image. Subsequently, the embedded feature vector passes through a feature encoder that includes a multi-head self-attention (MSA) mechanism and multi-layer perceptron (MLP) alternately. This facilitates the further extraction of long-term relevant features in the image.

The VIT image encoder is comprised of transformer blocks, as depicted in Figure 3. This constitutes a cascading module consisting of MSA blocks and MLP blocks. In each transformer block, the input features undergo normalization through LayerNorm prior to entering an MSA block. The MSA block functions similarly to the convolutional layers in CNNs, capturing spatial features of the image by extracting remote interactions between different image patches. The weighting information is obtained through the self-attention (SA) mechanism within the MSA blocks. The feature embeddings between the patches are then weighted and summed using these weights to obtain the interaction information.

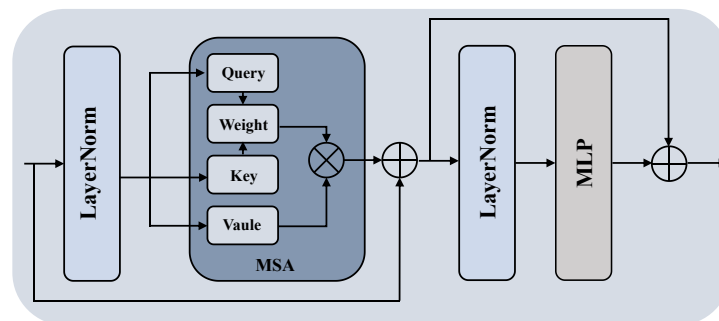


Figure 3. An overview of transformer block. MSA represents the multi-head self attention.

The SA block utilizes learnable linear matrices W_Q , W_K , and W_V to map the embedded features X of each image patch to vectors like the Q (query), K (key), and V (value). Following this, the similarity between the vector Q (query) and the vector K (key) is computed by scaling the dot product function. In this way, the weight of the interaction between different image patches is generated, focusing on local input features. The output is obtained by the weighted sum of the V vector. The SA module can be represented in matrix form as follows:

$$Attention(X, Q, K, V) = \text{soft max} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (1)$$

where $Q = XW_Q$, $K = XW_K$, and $V = XW_V$ are linear transformations using learnable matrixes W_Q , W_K , and W_V . d_k is the dimension of the key vector. QK^T represents the dot

product between the vector Q and K matrices. The softmax function is applied along the rows to obtain attention weights.

The MSA block is indeed composed of multiple SA heads running in parallel. Assuming MSA has n attention heads and the input of the l -th MSA in the model is $z^{(l-1)}$, the MSA will cut $z^{(l-1)}$ into n slices $z_i^{(l-1)}$, $i = 1, \dots, n$. Each slice $z_i^{(l-1)}$ is processed with a separate SA block. Subsequently, the outputs from the n attention heads are concatenated and fused using linear projection W_O . MSA can be represented as follows:

$$MSA(z^{(l-1)}) = [SA_1(z_1^{(l-1)})SA_2(z_2^{(l-1)})SA_3(z_3^{(l-1)})SA_4(z_4^{(l-1)})]W_O \quad (2)$$

MSA attains a unified representation of self-attention across diverse perspectives, effectively harnessing the correlations within the embedded data. In the transformer block, after applying MSA to the embedded vectors, layer normalization and transformation are further performed through an MLP block. The MLP consists of two linear transformations and an activation function, which performs a nonlinear transformation on the feature vector. The vector dimension of the intermediate layer is $L \cdot r_{MLP}$, where r_{MLP} is a predefined scaling factor. Therefore, for the input of each layer, it can be simplified to the following formula after processing by a transformer:

$$z^{(l)} = MLP(LN(MSA(LN(z^{(l-1)})))) \quad (3)$$

2.2.3. LoRA

We utilize the pretrained parameters in SAM, so the parameters associated with the VIT image encoder will be frozen and cannot be updated. This approach helps to avoid difficulties in the retraining and convergence of the entire SAM due to limited data. Additionally, it alleviates the computational resource pressure.

LoRA [24] assumes that when dealing with encoded token sequences $E \in R^{B \times N \times C_{in}}$ and the output $\hat{E} \in R^{B \times N \times C_{out}}$ from a projection layer W , updates to the projection layer should be gradual and stable. To achieve this, it recommends using a low-rank approximation to describe the gradual updating process. The low-rank approximation is a mathematical technique that reduces matrix complexity while preserving crucial information. It helps smoothly adjust the projection layer's parameters for incremental and stable updates. The utilization of LoRA is illustrated in Figure 4. It involves freezing the transformer layer to maintain the stability of W . Simultaneously, bypasses are introduced in the projection layers Q and V to accomplish low-rank approximation. As the MSA determines the regions to focus on based on cosine similarity, LoRA influences attention by being utilized on the projection layers.

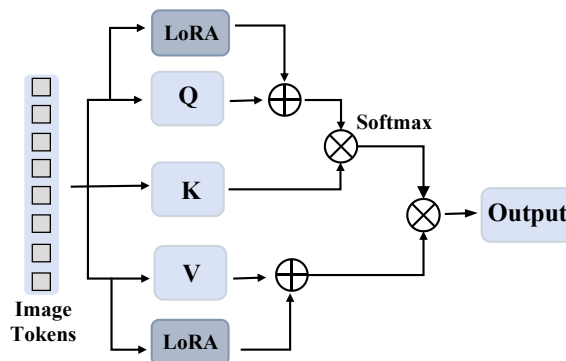


Figure 4. A flowchart of LoRA. Adding LoRA to the Q layer and V layer of the transformer.

These bypasses encompass two linear layers, denoted as L_A and L_B . The treatment of the updated layer W can be described as follows:

$$\begin{aligned}\hat{E} &= \hat{W}E, \\ \hat{W} &= W + \Delta W = W + L_A L_B.\end{aligned}\quad (4)$$

The processing strategy for multi-head self-attention will transform into the following procedure:

$$Attention(X, Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V \quad (5)$$

where

$$\begin{aligned}Q &= \hat{W}_q E = W_q E + L_{Aq} L_{Bq} E, \\ K &= W_k E, \\ V &= \hat{W}_v E = W_v E + L_{Av} L_{Bv} E.\end{aligned}\quad (6)$$

and where W_q , W_k , and W_v are derived from the frozen projection layers of SAM and L_{Aq} , L_{Bq} , L_{Av} , and L_{Bv} are trainable LoRA parameters.

2.3. SA-Gate

To ensure a better fusion of the features obtained from the dual-stream image encoder, we do not directly concatenate multimodal features using a channel-wise concatenation approach for feature fusion. Instead, we employ the SA-Gate module for this purpose. The SA-Gate module, illustrated in Figure 1, consists of two components: Feature Separation (FS) and Feature Aggregation (FA) parts.

To mitigate the propagation of error information during the aggregation of local-global features, we opt to employ FS to filter out noise signals in local regions. Initially, F_C and F_T are integrated through channel concatenation, followed by global average pooling.

$$I = \text{pooling}(F_C \parallel F_T) \quad (7)$$

where $\parallel$ represents the concatenation of feature mappings, *pooling* denotes global average pooling, and $I = (I_1, \dots, I_k, \dots, I_{2C})$ is a cross-modal global descriptor utilized for collecting expressive statistical information across the entire input.

Then, attention vectors W_C and W_T are obtained using MLP network.

$$\begin{aligned}W_C &= \sigma(F_{MLP}(I)) \\ W_T &= \sigma(F_{MLP}(I))\end{aligned}\quad (8)$$

where F_{MLP} denotes the MLP network and σ denotes the sigmoid function scaling the weight value into (0, 1).

By performing channel-wise multiplication of the input features F_C and F_T and attention vectors W_C and W_T , we obtain filtered feature maps with reduced noise, denoted as $Filter_C$ and $Filter_T$.

$$\begin{aligned}Filter_C &= F_C \otimes W_C \\ Filter_T &= F_T \otimes W_T\end{aligned}\quad (9)$$

The filtered features are pixel-wise summed with the original features to yield the calibrated feature vector.

$$\begin{aligned}Rec_C &= Filter_C + F_C \\ Rec_T &= Filter_T + F_T\end{aligned}\quad (10)$$

To fully exploit the complementary nature of local-global features, it is necessary to perform a complementary aggregation of cross-modal features at specific spatial locations based on their feature capabilities. In the initial stage, the features of these two mappings are merged through concatenation at specific spatial positions. Following this, we establish two mapping functions to link high-dimensional features to two separate spatial gates, utilizing 1×1 convolutions to execute these mapping functions.

$$\begin{aligned}F_C &: F_{concat} \rightarrow G_C \\ F_T &: F_{concat} \rightarrow G_T\end{aligned}\quad (11)$$

where G_C and G_T are the spatial-wise gate for F_C and F_T feature maps, respectively. A softmax function is applied to both gates:

$$\begin{aligned} A_C^{(i,j)} &= \frac{e^{G_C^{(i,j)}}}{e^{G_C^{(i,j)}} + e^{G_T^{(i,j)}}}, \\ A_T^{(i,j)} &= \frac{e^{G_T^{(i,j)}}}{e^{G_C^{(i,j)}} + e^{G_T^{(i,j)}}}, \\ A_C^{(i,j)} + A_T^{(i,j)} &= 1. \end{aligned} \quad (12)$$

$G_C^{(i,j)}$ and $G_T^{(i,j)}$ represent the weight factors corresponding to the position of the feature maps F_C and the F_T . The final feature fusion result M is obtained by the weighted sum of F_C and F_T .

$$M_{i,j} = F_C^{(i,j)} \cdot A_C^{(i,j)} + F_T^{(i,j)} \cdot A_T^{(i,j)} \quad (13)$$

where M represents the final input features for our mask decoder.

2.4. Decoder

The decoder section includes the mask decoder and the prompt decoder. As our input lacks bounding boxes or textual prompts, the prompt decoder is set to default embedding. In the previous section, Section 2.2.3, it is mentioned that the parameters of the transformer image encoder are initialized using the pretrained parameters of SAM, and they are not fine-tuned. Therefore, only the mask decoder undergoes training.

The mask decoder deterministically predicts each semantic class of X . Assuming there are n classes for segmentation, including one background class and $n - 1$ classes corresponding to each meaningful entity, the mask decoder simultaneously predicts n semantic masks $\hat{S}_l \in R^{h \times w \times k}$, corresponding to each semantic label. Finally, the generated predicted segmentation map is obtained as follows:

$$\hat{S} = \arg \max(\text{soft max}(\hat{S}_l, d = -1), d = -1) \quad (14)$$

where $d = -1$ indicates the softmax and argmax operations performed across the last dimension.

2.5. Loss Function

For training the model, our preference is the classical cross-entropy loss function. Widely employed in classification tasks, it defines the separation between the predicted and true probability distributions. This definition is encapsulated in the following formula:

$$Loss = - \sum_i^K y_i \cdot \log(\hat{y}_i) \quad (15)$$

where y_i represents the true probability distribution for class i and $\hat{y}_i$ represents the predicted probability distribution for class i .

2.6. Quantitative Analysis

To quantitatively assess our method's performance, three widely used metrics are utilized: the OA, Kappa, and IoU. The OA gauges the percentage of accurately classified pixels across the complete dataset. Computed as the ratio of correctly classified pixels (true positives and true negatives) to the total pixel count, it provides an indication of classification precision.

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \quad (16)$$

where TP , TN , FP , and FN represent true positive, true negative, false positive, and false negative, respectively.

The Kappa is a statistical measure, used to ensure the correct classification of small categories in the dataset and ensure the balance of the method. It provides a normalized accuracy score.

$$Kappa = \frac{(P_0 - P_e)}{1 - P_e} \quad (17)$$

where P_0 represents the proportion of correct classifications for each category among all samples. Assuming the true sample counts for each class are $a_1, a_2, \dots, a_n$ and the predicted sample counts for each class are $b_1, b_2, \dots, b_n$, with a total sample count of n , then P_e is calculated as follows: $p_e = (a_1 \times b_1 + a_2 \times b_2 + \dots + a_n \times b_n) / (n \times n)$.

The IoU measures the overlap between predicted and true pixel masks. It is calculated as the intersection of the predicted and true positive areas divided by the union of their areas.

$$IoU = \frac{Pre \cap GT}{Pre \cup GT} \quad (18)$$

where Pre represents predicted value and GT represents the ground truth.

The mIoU is defined as the average of the IoU values between each Pre and GT across all pixels in an image.

$$mIoU = \frac{1}{C} \sum_1^C IoU_C \quad (19)$$

where C is the total number of classes.

3. Results

3.1. Training Settings and Data Description

Utilizing the PyTorch framework, our proposed method operates with Python 3.8.8, the version of Torch is 1.13.0, and the version of CUDA is 11.7. The training process involves 60 rounds with an initial learning rate of 1.0×10^{-4} . To enhance the computational efficiency, all experiments are executed on a GPU. Network parameter optimization is achieved through the Adam optimizer [34].

To comprehensively demonstrate the performance of the network, we choose two widely used land cover classification datasets for evaluation.

(1) AIR-PolSAR-Seg [35]: The dataset was produced by the Chinese Academy of Sciences and is based on images taken by the Gaofen-3 satellite in the full-polarization strip mode. The spatial resolution of these image data is 8 m, covering a selected area of 9082×9805 pixel images and $2000 \times 512 \times 512$ pixel images. We used the latter part. The image includes four polarization modes: HH, HV, VV, and VH. The ground truth values are illustrated in Figure 5. For experimental convenience, four patches were cropped from the original image data, each measuring 256×256 . Each final image is composed of four polarization images. Therefore, the total number of image patches is 8000.

There are six categories in total: industrial zones, natural features, land utilization areas, water bodies, miscellaneous (other) areas, and residential (housing) zones, which can be used for various types of tasks, such as large scene terrain classification, water body segmentation, and building extraction. The labels of truth values and color coding are displayed in Figure 5. Since land use areas and other areas account for a small proportion, no statistics were conducted in this experiment. To ensure as much consistency as possible between the data distributions of the training and testing sets, a random selection process was employed where 80% of the image patches were chosen for the training set, leaving the remaining 20% to comprise the test set.

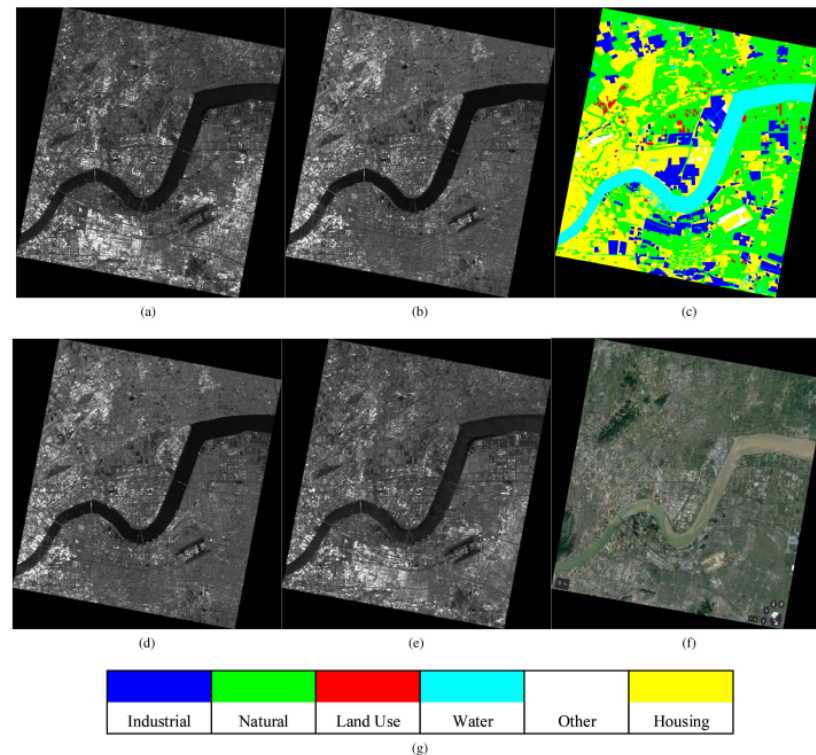


Figure 5. Examples of various images from the AIR-PolSAR-Seg dataset: (a) image of HH polarization method; (b) image of HV polarization method; (c) ground truth; (d) image of VH polarization method; (e) image of VV polarization method; (f) the corresponding optical image; (g) color representation of different categories.

(2) WHU-OPT-SAR [36]: The WHU-OPT-SAR dataset was released by Wuhan University and contains 38 SAR and optical image pairs, which were acquired through the GF-3 fine strip II. Each image measures 5556 pixels in width and 3704 pixels in height. Each pair of images has a corresponding feature category annotation. In order to facilitate the experiment, we partitioned all images into multiple smaller images of 256 pixels in length and width, allocating them to training and test sets in an 8:2 ratio. An included image is shown in Figure 6. The region exhibits a subtropical monsoon climate, ranging from a minimum elevation of 50 m to a maximum elevation of 3000 m. This dataset encompasses a wide range of remote sensing images capturing diverse topographies and vegetation types. It includes seven main categories, namely farmland, city, village, water, forests, road, and others. Figure 7 provides examples of these primary categories.

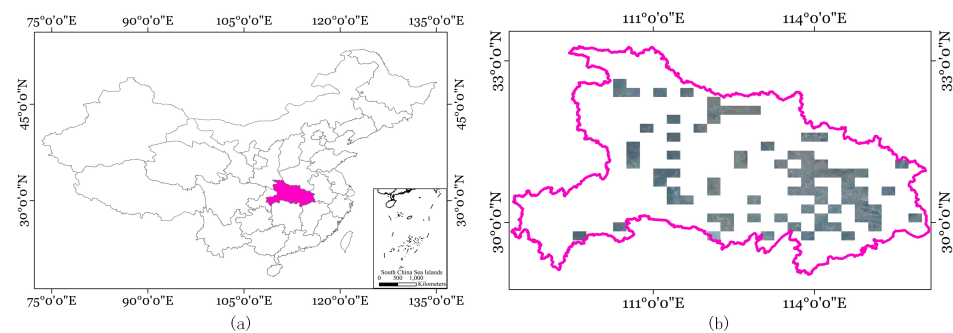


Figure 6. The details of the WHU-OPT-SAR dataset: (a) location of image collection area; (b) the specific distribution of images in the region.

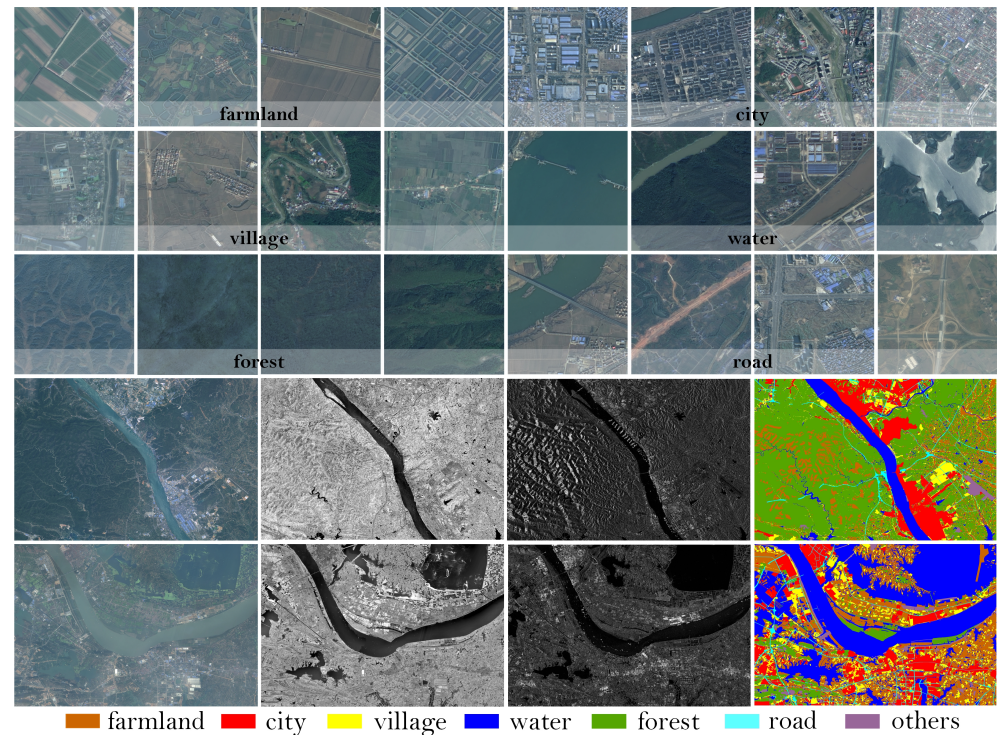


Figure 7. Examples of every category in the WHU dataset. The top part displays common scenes of the dataset, while the bottom part includes optical images, near-infrared images, SAR images, and annotations from left to right.

3.2. Land Cover Classification Experiments

3.2.1. Comparative Baselines

Within this paper, our proposed method is juxtaposed against several standard semantic segmentation algorithms for comparative evaluation:

- SwinUnet [37]: This method utilizes a Unet-inspired pure transformer for medical image segmentation. It incorporates image patches into a U-shaped encoder–decoder architecture based on transformers, enabling the acquisition of local–global semantic features.
- SegNet [38]: A comprehensive fully convolutional neural network structure, consisting of an encoder network paired with a decoder network; the architecture culminates in a layer to classify. SegNet introduces a noteworthy innovation by employing upsampling in its decoder to handle lower-resolution input feature maps.
- TransUnet [39]: A method that merits both transformers and U-Net, called TransUnet. The transformer processes input sequences derived from CNN-encoded feature maps for global context extraction. Simultaneously, the decoder upscales the encoded features and integrates them with high-resolution CNN feature maps to attain accurate segmentation.
- DANet [40]: This approach employs a dual-stream network, extracting features independently from SAR and optical images. It models semantic interdependencies on both spatial and channel dimensions by attaching two types of attention modules to the expanded fully convolutional network (FCN).
- EncNet [41]: An approach that captures the semantic context of the scene by introducing a context encoding module and selectively highlighting feature maps dependent on the class dependencies.
- Deeplabv3 [42]: A DeepLabv3 network with an encoder–decoder structure is proposed. DeepLabv3 is utilized for encoding rich contextual information, and a simple, effective decoder is employed to recover object boundaries.

- UnetFormer [43]: The method proposed a transformer-based decoder featuring an efficient global–local attention mechanism known as the global–local transformer block (GLTB), tailored for real-time urban scene segmentation.
- LSKNet [44]: The method proposed a rotating frame target detection algorithm for remote sensing images. This algorithm dynamically adjusts its large spatial sensing field to better model scenarios for ranging various objects in remote sensing scenes.

3.2.2. Comparison Experiments

To assess the capabilities of different land cover classification algorithms, we select five scenarios to observe the visualization and quantitative test results, as shown in Figure 8. It can be clearly observed that the continuity of ground object edges in other algorithms is inferior, leading to severe pixel confusion. Moreover, these algorithms generate a significant number of errors in the classification of natural, industrial, and water images. In contrast, our proposed method exhibits smaller misclassification areas and achieves more accurate detection in edge regions. Table 1 shows that the OA, Kappa, and F1 increase by 1.28%, 0.16%, and 0.58%. For nature and water, our method considerably boosts the performance with the Iou increasing by 3.76% and 0.8%, respectively. Additionally, our approach demonstrates strong robustness in accuracy across all categories. While it may not achieve an optimal performance in the industrial and housing categories, its overall accuracy remains the highest. This is attributed to its avoidance of an exceptionally poor performance in any specific category and maintaining an appropriate balance in accuracy across all categories. In contrast, SwinUnet, TransUnet, and DANet exhibit significant discrepancies in accuracy among various categories, particularly in the nature class, highlighting poor robustness.

Table 1. Results of metric values for each comparison experiment on the AIR-PolSAR-Seg dataset. The **bold** and *italic* entries indicate the optimal and suboptimal results.

Method	OA	Kappa	mIou	F1	Iou			
					Industrial	Natural	Water	Housing
SwinUnet	43.49	26.37	32.40	43.32	13.21	5.8	73.44	37.14
LSKNet	45.97	25.91	35.88	44.77	33.36	1.6	86.13	22.45
SegNet	52.25	33.30	41.57	51.85	32.57	0.64	83.28	49.78
UnetFormer	61.02	49.43	49.35	59.95	36.49	6.61	81.89	72.42
TransUnet	68.67	51.26	47.32	58.38	58.22	2.5	77.73	50.81
DANet	70.84	58.21	53.23	62.35	55.61	3.21	95.67	58.43
EncNet	73.92	63.32	59.20	67.90	66.67	<i>17.18</i>	<i>99.05</i>	53.92
Deeplabv3	70.57	59.80	56.45	69.32	<i>65.17</i>	14.33	85.66	60.62
Ours	75.20	63.48	58.79	69.90	64.38	20.94	99.85	50.01

To further assess the effectiveness of our proposed algorithm, we selected areas with complex land cover. Six diverse terrain scenes were chosen, as depicted in Figure 9. The land masks generated by our proposed method exhibit a superior edge continuity, notably outperforming in depicting challenging categories, such as farmland and water. This superiority may stem from our method’s ability to fully exploit critical semantic information distinguishing farmland and water, independent of surface visual features. The method integrates these insights at higher levels, leading to more accurate classification results. Quantitative evaluation results are summarized in Table 2, demonstrating notable advancements in metrics for our proposed method. The OA, Kappa, mIoU, and F1 have shown improvements of 0.2%, 0.34%, 1.11%, and 1.27% respectively. Notably, our method exhibits substantial gains in the IoU for specific classes, including farmland (1.85%), village (2.56%), water (2.6%), and other categories (3.6%). It is noteworthy that our algorithm maintains a high level of accuracy across all classes without exhibiting extreme deficiencies.

Table 2. Results of metric values for each comparison experiment on the WHU-OPT-SAR dataset. The **bold** and *italic* entries indicate the optimal and suboptimal results.

Method	OA	Kappa	mIou	F1	Iou						
					Farmland	City	Village	Water	Forest	Road	Others
SwinU	70.53	57.80	34.04	45.84	56.22	30.84	32.87	32.64	77.27	6.04	2.42
LSKNet	69.82	55.95	30.57	41.61	56.17	26.63	24.92	18.01	78.08	5.53	4.64
SegNet	73.21	60.51	36.00	48.85	59.35	29.55	36.03	27.76	77.88	15.48	5.93
UnetF	70.54	58.65	35.65	48.21	59.84	27.41	36.26	35.19	75.33	10.08	5.46
TransU	73.40	61.87	38.09	47.28	61.22	32.49	36.15	33.01	77.82	15.63	9.88
DANet	69.18	56.82	34.31	45.04	61.09	26.41	34.62	23.71	74.61	12.82	6.92
EncNet	73.67	61.73	35.34	51.50	63.72	25.33	32.88	32.46	78.70	10.67	3.64
Deepl3	73.59	62.33	38.44	51.98	62.46	29.01	38.76	32.67	77.96	15.43	12.82
Ours	73.87	62.67	39.55	53.25	59.87	34.34	41.32	35.61	78.36	10.94	16.42

The latest baseline models, UnetFormer and LSKNet, exhibit suboptimal performances on the aforementioned datasets. UnetFormer relies on a transformer architecture, emphasizing global attention mechanisms. However, due to the intricate nature of SAR images, a local attentional network like a CNN may yield superior results. LSKNet is designed for directed target detection, focusing on entire targets rather than individual pixel points, leading to less satisfactory outcomes in segmentation tasks.

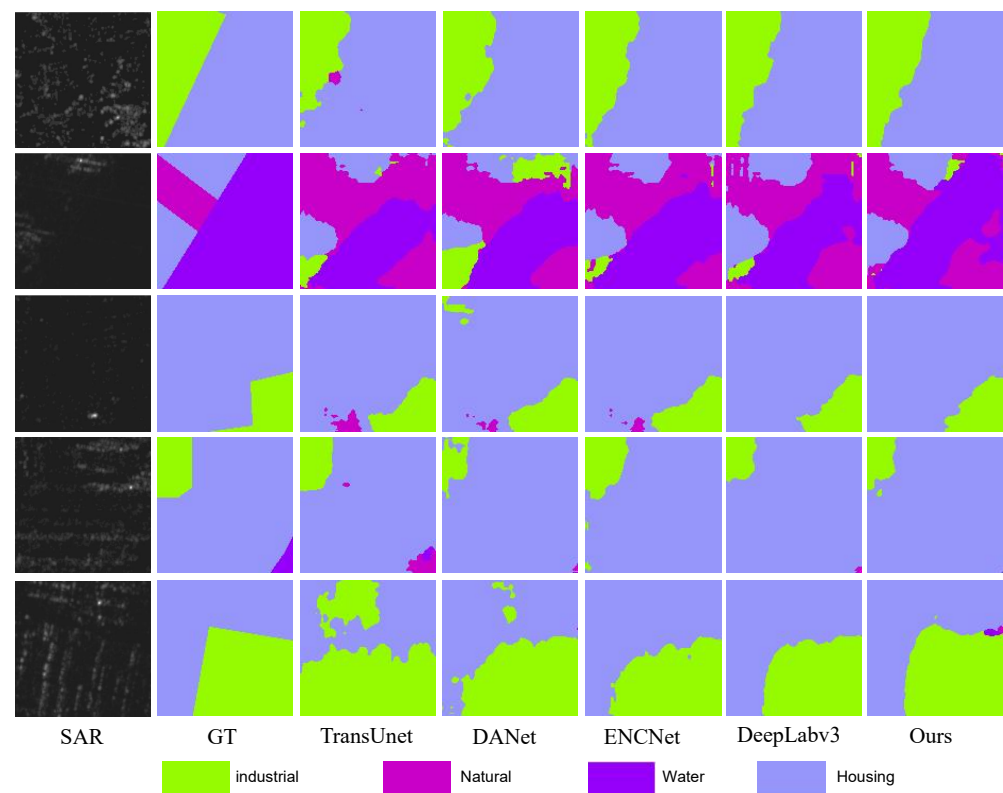


Figure 8. Land cover classification results for different comparison experiments on the AIR-PolSAR-Seg dataset. From left to right, each column represents PolSAR, ground truth, and comparison results of TransUnet, DANet, ENCNet, DeepLabV3, and our model.

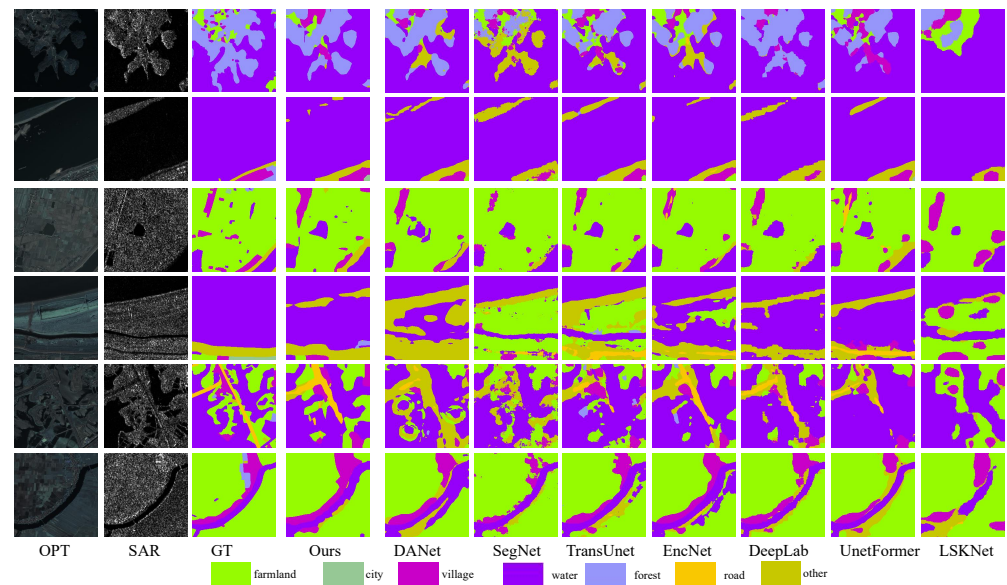


Figure 9. Land cover classification results for different comparison experiments on the WHU-OPT-SAR dataset. From left to right, each column represents PolSAR, ground truth, and comparison results of our model, DANet, SegNet, TransUnet, ENetNet, DeepLabV3, UnetFormer, and LSKNet.

3.2.3. Ablation Analysis

To assess the efficacy of the principal components within the BSDSNet proposal, we performed ablation studies on the WHU-OPT-SAR dataset. Specifically, we conducted experiments on every single component (i.e., the ConvNext image encoder, VIT image encoder, the SA-Gate module) and their combinations. Detailed results of the ablation studies are presented in Table 3, indicating that each component within the BSDSNet framework makes a positive contribution to the overall outcomes. Figure 10 presents the corresponding visual results, providing additional validation of the effectiveness of each key module.

The network is a pure SAM when only the VIT image encoder is employed. Evaluation metrics indicate that SAM encounters ill-posed challenges in land cover classification tasks due to the distinctive imaging mechanism inherent in SAR. Moreover, the integration of the ConvNext branch substantially elevates the metrics, substantiating the indispensability of local–global feature extraction. At this stage, the absence of the SA-Gate module results in the sole utilization of the concatenate operation for merging local–global features. Furthermore, the introduction of the SA-Gate module optimally aligns the metrics to their peak values, affirming the efficacy of accurately amalgamating multi-scale features.

Table 3. Ablation experiment of BSDSNet network with and without the use of VIT image encoder, ConvNext image encoder, and SA-Gate. The **bold** indicates the optimal results.

VIT	ConvNext	SA-Gate	OA	Kappa	mIou	F1
✓	×	×	68.50	55.00	32.60	44.78
✓	✓	×	71.10	58.82	35.38	49.25
✓	✓	✓	73.87	62.67	39.55	53.25

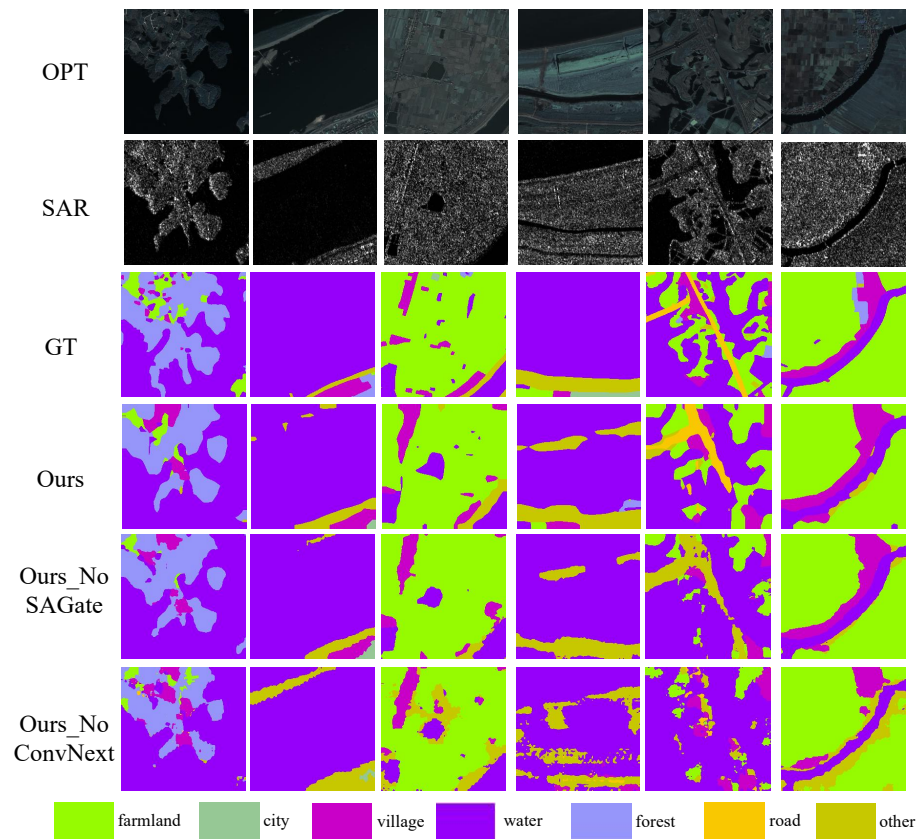


Figure 10. Land cover classification results for ablation experiment on the WHU-OPT- SAR dataset.

4. Discussion

Recently, deep learning, benefiting from its powerful data feature extraction capability, has found widespread applications in land cover classification tasks in the remote sensing domain. It achieves this by constructing multi-layered networks to extract high-dimensional features and uncover abstract spatial and semantic features from the data, thereby enabling the processing of complex datasets. We evaluate its performance by using two different datasets and compare the results with various other methods. In recent years, the emergence of SAM has propelled the development of semantic segmentation forward significantly. SAM demonstrates powerful capabilities in extracting semantic information and generalization. Therefore, we choose to directly apply it to our network architecture. We believe that, in the near future, architectures based on SAM will play a crucial role in advancing remote sensing. Additionally, many existing deep learning methods directly concatenate SAR and optical image channels to achieve data fusion. However, this approach fails to effectively fuse the features of both modalities due to their differing imaging mechanisms. Therefore, based on the characteristics of multimodal data, we design a dual-stream encoder network to better extract specific multimodal features. We utilize SA-Gate to fuse multimodal features from multiple dimensions. Multimodal feature fusion poses a significant challenge as existing research cannot directly evaluate the fusion status of features from both modalities.

5. Conclusions

Within this paper, we propose a dual-stream network for extracting features based on SAM for land cover classification. We extend SAM's image encoder to change it to a dual stream as a way to extract local–global features of PolSAR. Our method has a more powerful feature extraction capability to extract more fine-grained texture features and semantic information of PolSAR to improve the accuracy of the land classification task. Subsequently, an SA-Gate module is proposed to realize the fully effective complementary interaction of multi-scale information. Experiments involving both comparisons and abla-

tions were executed on the AIR-PolSAR-Seg and WHU-OPT-SAR datasets. Findings from the comparative experiments underscore the method's superiority over alternative deep learning PolSAR land cover classification approaches. Ablation experiments systematically explored the impact of each primary component on the classification performance, affirming their substantive contributions. In the future, fusion methods for multi-scale features can be further explored and the model can be applied to more complex classification tasks.

Author Contributions: Data collection, methodology, Y.W.; writing—original draft, W.Z. and Y.W.; writing—review and editing, Y.W. and W.C.; project administration, C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The AIR-PolSAR-Seg dataset is openly available at <https://github.com/AICyberTeam/AIR-PolSAR-Seg> (accessed on 15 January 2024). The WHU-OPT-SAR dataset is openly available at <https://github.com/AmberHen/WHU-OPT-SAR-dataset> (accessed on 30 January 2024).

Acknowledgments: The authors express their gratitude to the researchers who generously provided the AIR-PolSAR-Seg dataset and the WHU-OPT-SAR dataset at no cost.

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

References

- Letsoin, S.M.A.; Herak, D.; Purwestri, R.C. Evaluation Land Use Cover Changes over 29 Years in Papua Province of Indonesia Using Remote Sensing Data. *IOP Conf. Ser. Earth Environ. Sci.* **2022**, *1034*, 012013. [CrossRef]
- Dahhani, S.; Raji, M.; Hakdaoui, M.; Lhissou, R. Land cover mapping using sentinel-1 time-series data and machine-learning classifiers in agricultural sub-saharan landscape. *Remote Sens.* **2022**, *15*, 65. [CrossRef]
- Gómez, C.; White, J.C.; Wulder, M.A. Optical remotely sensed time series data for land cover classification: A review. *ISPRS J. Photogramm. Remote Sens.* **2016**, *116*, 55–72. [CrossRef]
- Xu, S.; Qi, Z.; Li, X.; Yeh, A.G.O. Investigation of the effect of the incidence angle on land cover classification using fully polarimetric SAR images. *Int. J. Remote Sens.* **2019**, *40*, 1576–1593. [CrossRef]
- Xie, C.; Zhang, X.; Zhuang, L.; Han, W.; Zheng, Y.; Chen, K. Classification of polarimetric SAR imagery based on improved MRF model using Wishart distance and category confidence-degree. In Proceedings of the 2023 IEEE International Radar Conference (RADAR), Sydney, Australia, 6–10 November 2023; pp. 1–4.
- Chaudhari, N.; Mitra, S.K.; Mandal, S.; Chirakkal, S.; Putrevu, D.; Misra, A. Edge-Preserving classification of polarimetric SAR images using Wishart distribution and conditional random field. *Int. J. Remote Sens.* **2022**, *43*, 2134–2155. [CrossRef]
- Montanaro, A.; Valsesia, D.; Fracastoro, G.; Magli, E. Semi-supervised learning for joint SAR and multispectral land cover classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 2506305. [CrossRef]
- Kang, W.; Xiang, Y.; Wang, F.; You, H. CFNet: A cross fusion network for joint land cover classification using optical and SAR images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1562–1574. [CrossRef]
- Ghanbari, M.; Xu, L.; Clausi, D.A. Local and global spatial information for land cover semi-supervised classification of complex polarimetric SAR data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 3892–3904. [CrossRef]
- Wu, Y.; Ji, K.; Yu, W.; Su, Y. Region-based classification of polarimetric SAR images using Wishart MRF. *IEEE Geosci. Remote Sens. Lett.* **2008**, *5*, 668–672. [CrossRef]
- Mishra, P.; Singh, D.; Yamaguchi, Y. Land cover classification of PALSAR images by knowledge based decision tree classifier and supervised classifiers based on SAR observables. *Prog. Electromagn. Res. B* **2011**, *30*, 47–70. [CrossRef]
- Zhou, Y.; Wang, H.; Xu, F.; Jin, Y.Q. Polarimetric SAR Image Classification Using Deep Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1935–1939. [CrossRef]
- Mei, S.; Ji, J.; Hou, J.; Li, X.; Du, Q. Learning sensor-specific spatial-spectral features of hyperspectral images via convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4520–4533. [CrossRef]
- Kavran, D.; Mongus, D.; Žalik, B.; Lukač, N. Graph Neural Network-Based Method of Spatiotemporal Land Cover Mapping Using Satellite Imagery. *Sensors* **2023**, *23*, 6648. [CrossRef] [PubMed]
- Zhao, W.; Peng, S.; Chen, J.; Peng, R. Contextual-Aware Land Cover Classification with U-Shaped Object Graph Neural Network. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6510705. [CrossRef]
- Fang, Z.; Zhang, G.; Dai, Q.; Xue, B.; Wang, P. Hybrid Attention-Based Encoder–Decoder Fully Convolutional Network for PolSAR Image Classification. *Remote Sens.* **2023**, *15*, 526. [CrossRef]
- Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.

18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
19. Dong, H.; Zhang, L.; Zou, B. Exploring Vision Transformers for Polarimetric SAR Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5219715. [CrossRef]
20. Wang, H.; Xing, C.; Yin, J.; Yang, J. Land cover classification for polarimetric SAR images based on vision transformer. *Remote Sens.* **2022**, *14*, 4656. [CrossRef]
21. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.
22. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.
23. 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. *arXiv* **2023**, arXiv:2304.02643.
24. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. *arXiv* **2021**, arXiv:2106.09685.
25. Chen, X.; Lin, K.Y.; Wang, J.; Wu, W.; Qian, C.; Li, H.; Zeng, G. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In Proceedings of the European Conference on Computer Vision, Online, 23–28 August 2020; pp. 561–577.
26. Xu, F.; Shi, Y.; Ebel, P.; Yu, L.; Xia, G.S.; Yang, W.; Zhu, X.X. GLF-CR: SAR-enhanced cloud removal with global–local fusion. *ISPRS J. Photogramm. Remote Sens.* **2022**, *192*, 268–278. [CrossRef]
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
28. Chollet, F. Deep learning with depthwise separable convolutions. In Proceedings of the CVPR, Honolulu, HI, USA, 21–26 July 2017.
29. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
30. Wu, Z.; Jiang, X. Extraction of pine wilt disease regions using UAV RGB imagery and improved mask R-CNN models fused with ConvNeXt. *Forests* **2023**, *14*, 1672. [CrossRef]
31. Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv* **2016**, arXiv:1606.08415.
32. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the ICML, Haifa, Israel, 21–24 June 2010.
33. Zhang, K.; Liu, D. Customized segment anything model for medical image segmentation. *arXiv* **2023**, arXiv:2304.13785.
34. Dozat, T. Incorporating Nesterov Momentum into Adam. 2016. Available online: <https://openreview.net/forum?id=OM0jvwB8jIp57ZJjtNEZ> (accessed on 7 February 2024).
35. Wang, Z.; Zeng, X.; Yan, Z.; Kang, J.; Sun, X. AIR-PolSAR-Seg: A large-scale data set for terrain segmentation in complex-scene PolSAR images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 3830–3841. [CrossRef]
36. Li, X.; Zhang, G.; Cui, H.; Hou, S.; Wang, S.; Li, X.; Chen, Y.; Li, Z.; Zhang, L. MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *106*, 102638. [CrossRef]
37. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 205–218.
38. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
39. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
40. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
41. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7151–7160.
42. Chen, L.; 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.
43. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
44. Li, Y.; Hou, Q.; Zheng, Z.; Cheng, M.-M.; Yang, J.; Li, X. Large selective kernel network for remote sensing object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 16794–16805.

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.