Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.064441

ARTICLE

Remote Sensing Image Information Granulation Transformer for Semantic Segmentation

Haoyang Tang[1,2] **and Kai Zeng**[1,2,*]

[1]Faculty of Information Engineering Automation, Kunming University of Science and Technology, Kunming, 650500, China
[2]Yunnan Key Laboratory of Computer Technologies Application, Kunming, 650500, China

*Corresponding Author: Kai Zeng. Email: zengkai@kust.edu.cn

Received: 20 January 2025; Accepted: 17 April 2025; Published: 09 June 2025

ABSTRACT: Semantic segmentation provides important technical support for Land cover/land use (LCLU) research. By calculating the cosine similarity between feature vectors, transformer-based models can effectively capture the global information of high-resolution remote sensing images. However, the diversity of detailed and edge features within the same class of ground objects in high-resolution remote sensing images leads to a dispersed embedding distribution. The dispersed feature distribution enlarges feature vector angles and reduces cosine similarity, weakening the attention mechanism's ability to identify the same class of ground objects. To address this challenge, remote sensing image information granulation transformer for semantic segmentation is proposed. The model employs adaptive granulation to extract common semantic features among objects of the same class, constructing an information granule to replace the detailed feature representation of these objects. Then, the Laplacian operator of the information granule is applied to extract the edge features of the object as represented by the information granule. In the experiments, the proposed model was validated on the Beijing Land-Use (BLU), Gaofen Image Dataset (GID), and Potsdam Dataset (PD). In particular, the model achieves 88.81% for mOA, 82.64% for mF1, and 71.50% for mIoU metrics on the GID dataset. Experimental results show that the model effectively handles high-resolution remote sensing images. Our code is available at https://github.com/sjmp525/RSIGT (accessed on 16 April 2025).

KEYWORDS: Land-cover/land-use; high-resolution remote sensing images; transformer; adaptive granulation

1 Introduction

LCLU serves as a critical foundation for understanding earth system changes and human activities [1]. It provides a scientific foundation for the optimal allocation of land resources, ecosystem conservation, climate change assessment, and disaster management. In LCLU research, semantic segmentation techniques based on high-resolution remote sensing imagery have been widely applied. Semantic segmentation technology precisely labels and classifies each pixel within a high-resolution image. It effectively enhances the accuracy and efficiency of researchers in identifying and analyzing surface features. Thus, semantic segmentation technology of high-resolution remote sensing images has been instrumental in the advancement of LCLU research.

In recent years, deep learning-based high-resolution image segmentation networks have been widely applied in LCLU research. Current deep learning network approaches primarily encompass CNN-based models [2] and transformer-based models [3]. CNN-based models primarily focus on spatial and detailed feature extraction by stacking multiple convolutional layers [4]. This has enabled CNN-based models

to achieve promising performance in semantic segmentation tasks for high-resolution remote sensing. Nevertheless, CNN-based models still face inherent limitations stemming from the local receptive fields of convolutional neural networks. To address this issue, researchers have focused on transformer-based networks with global context modeling capabilities [5]. The core of transformer-based models lies in the attention mechanism [6]. Through calculating the cosine similarity between feature vectors, the attention mechanism can obtain the similarity between the geographical objects embedded therein. This enables the model to effectively pay attention to all geographical objects with high similarity in high-resolution remote sensing images. The characteristic of the attention mechanism has propelled transformer-based models to become a major research focus in the domain of high-resolution remote sensing image semantic segmentation.

However, there remains a significant challenge for transformer-based approaches when dealing with high-resolution remote sensing images. High-resolution remote sensing images exhibit two distinctive characteristics. Firstly, in high-resolution remote sensing images, objects of the same class exhibit varying sizes and details [7]. Secondly, the complex and diverse backgrounds in high-resolution remote sensing images can blur the boundaries of ground objects [8]. These characteristics lead to markedly different feature representations for ground objects belonging to the same category. As shown in Fig. 1a, the diversified feature representations lead to an expansion and dispersion of the embedding distributions of objects belonging to the same class in the feature space [9]. Dispersed feature distributions imply significant angles between feature vectors. The larger the angles between feature vectors, the smaller the values obtained from cosine similarity calculations. This reveals that attention scores among objects of the same class are relatively low and consequently the attention mechanism might overlook some objects from that class, as shown in Fig. 1b. The semantic segmentation performance of transformer-based models is limited by their incomplete focus on information. Therefore, it is necessary to address the significant differences in feature representations among objects of the same class.

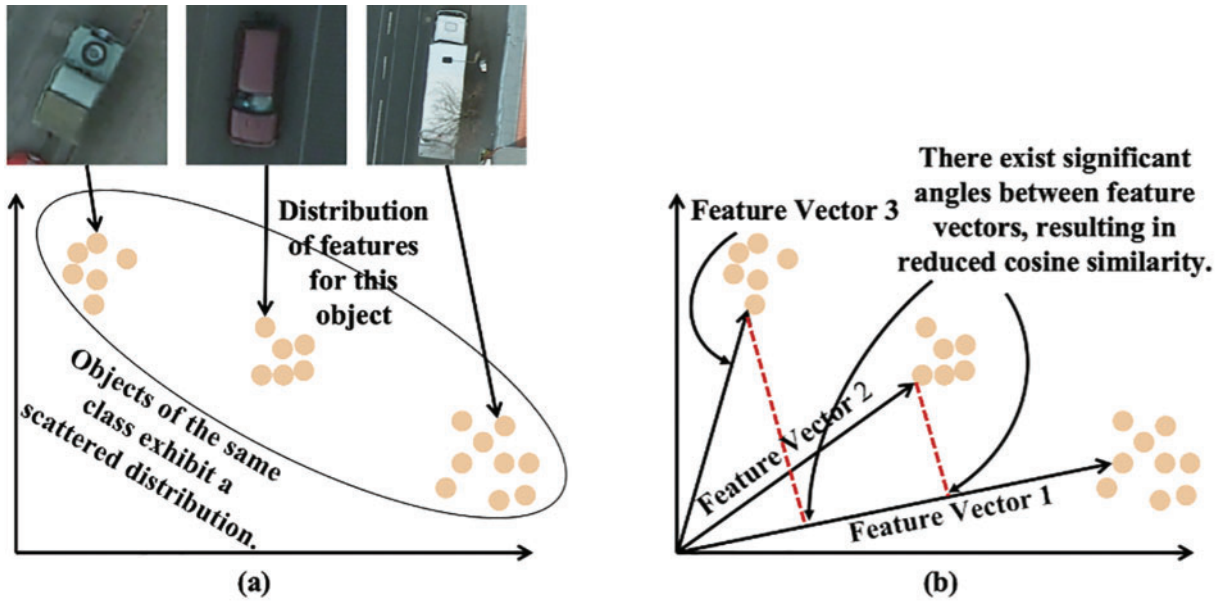

Figure 1: Schematic representation of the challenges of high resolution remote sensing images on transformer performance (a) Features distribution of objects within the same class is dispersed. (b) The cosine similarity between the eigenvectors where the decentralized feature distribution is located decreases

Existing approaches to this problem have primarily focused on leveraging multi-scale information [10] or incorporating supplementary data. But these methods often encounter challenges related to data inseparability. As a result, it is essential to seek new solutions. Notably, although objects of the same class in a high-resolution remote sensing image show significant differences in feature representations, intrinsic relationships still exist among their feature points [11]. This inspired us to leverage intrinsic relationships to reconstruct feature representations of the same class objects in high-resolution remote sensing images.

Information granulation is a commonly used method for reconstructing feature representations based on intrinsic relationships [12]. The process of information granulation involves calculating the relationships between feature points and encapsulating related feature points into an information granule. An information granule is a novel information entity that represents the shared semantic features of its internal feature points [11]. The shared semantic features of information granules ensure consistent feature representation for the same type of objects. In the feature space a unified representation forms a more compact distribution. This yields smaller angles between feature vectors and higher cosine similarity. Such a structure enhances the effectiveness of attention mechanisms in focusing on critical information. This highlights the necessity of utilizing information granulation to address the significant feature representation differences of same class objects in high-resolution remote sensing images.

Numerous information granulation methods are currently utilized in deep learning [13]. These methods establish the correlation between feature points by calculating the spatial distance between feature points and a manually specified fixed reference point. Nevertheless, since the distribution of feature points within the same category in the feature space is typically stochastic and dispersed. The presence of varying optimal reference points for different feature distributions presents a substantial challenge in manually identifying the most suitable reference point. Therefore, conventional granulation methods are not well-suited for high-resolution remote sensing images. Developing a granulation method capable of automatic reference point selection is essential.

In deep learning, the Laplacian operator is frequently employed to extract the edges of geographic objects. The Laplacian operator captures second-order variations in the input image, thereby serving as a parameter-free method to extract high-frequency details such as edges and contours. These high-frequency features play a pivotal role in delineating the boundaries of geospatial objects, particularly by discriminating between the boundaries and background in scenarios where the background information is complex. It is conceivable that applying the Laplacian operator to information granules could also extract the edge features of objects. However, there is currently no research that applies the Laplacian operator to information granules.

Based on the above analysis, the adaptive granulation method and the Laplacian operator of information granule applicable to high-resolution remote sensing image segmentation are proposed. Subsequently, remote sensing image information granulation transformer for semantic segmentation is developed. The contributions of the article are as follows:

1. An adaptive granulation method suitable for high-resolution remote sensing images is proposed. Adaptive granulation captures relationships between same-class feature points and constructs information granules. These granules are then used to reconstruct unified feature representations for objects of the same class.
2. The Laplacian operator for information granule is defined. The Laplacian operator for information granule is specifically designed to process the information granule. The core function of the Laplacian operator of information granule is to extract the edge information of the features represented by the information granule.

3. Remote sensing image information granulation transformer for semantic segmentation is presented. The model includes Swin with adaptive granulation, a feature transformation structure, and a dilated convolutional block, among other components.

The remainder of this paper is organized as follows: Section 2 provides the related works of this paper. Section 3 provides a detailed description and explanation of the remote sensing image information granulation transformer for semantic segmentation. Then, Section 4 demonstrates the effectiveness of the proposed method through experiments, and Section 5 concludes the study.

2 Related Works

This section begins with a concise review of the transformers for semantic segmentation of high-resolution remote sensing images in LCLU research. Following this, it will discuss research on information granulation within deep learning.

2.1 Transformers for Semantic Segmentation of High-Resolution Remote Sensing Images in LCLU Research

In recent years, scholars have shown increasing interest in using transformers for high-resolution remote sensing image semantic segmentation in LCLU analysis.

Zhang et al. introduce high-order transformer blocks to model global dependencies and a global enhancement attention module (GEAM) to enhance global feature representation, addressing the challenge of segmenting complex ground objects [14]. Yang et al. propose a multi-scale Transformer (MSTrans) with a plug-and-play multi-scale transformer module based on atrous spatial pyramid pooling (ASPP) to enhance multi-scale feature extraction for building extraction from high-resolution remote sensing images [15]. Yu et al. employ the Swin transformer as the backbone to model global information interaction, effectively capturing long range dependencies and overcoming the limitations of CNNs in remote sensing image super-resolution [16].

Although transformers have achieved significant success in the semantic segmentation of high-resolution remote sensing images, the challenges posed by high intra-class variance in these images limit their segmentation performance.

2.2 Resolving the Impact of High-Resolution Remote Sensing Images on Transformers Performance Research

Currently, numerous scholars have proposed various methods to mitigate the impact of high intra-class variance on the performance of transformer. Yang et al. introduced a spatial-frequency multiscale transformer framework that effectively captures global multiscale features of the target by employing spatial multiscale modeling and a frequency-domain texture enhancement encoder. The framework further integrates spatial and frequency information using an adaptive feature fusion module [17]. Li et al. addressed high intra-class variance in remote sensing segmentation by proposing the Synergistic Attention Module (SAM), which jointly models spatial and channel affinities in a unified attention map. This approach enhances feature consistency and reduces attention bias, leading to improved segmentation accuracy when integrated into SAPNet [10]. Yang et al. proposed a method combining global spatial features and Fourier frequency domain learning to reduce intra-class variance. High-frequency components enhance boundaries, while low-frequency components improve internal consistency, resulting in clearer road features [18].

Despite promising results achieved by existing methods, these approaches primarily focus on enhancing data description from a multi-scale perspective, which can lead to issues with data separability. Therefore, it is

necessary to explore new methods to overcome the shortcomings of existing approaches. Notably, although objects of the same class in a high-resolution remote sensing image show significant differences in feature representations, intrinsic relationships still exist among their feature points. This inspired us to leverage intrinsic relationships to reconstruct feature representations of the same class objects in high-resolution remote sensing images.

2.3 Information Granulation in Deep Learning

Information granulation is a widely used approach to analyze similarity relationships, emphasizing the creation of information granules by grouping feature points with similar properties. Numerous scholars have incorporated information granulation into deep learning research.

Yu et al. proposed the GGI-DDI model, which decomposes drugs into information particles of key substructures through information granulation. This approach enhances feature expression and model interpretability, enabling more effective prediction of drug-drug interactions [19]. Behzadidoost et al. enhance the accuracy of classifiers in natural language processing tasks by decomposing data into multiple granular matrices and introducing combinatorial algorithms along with regularized numerical information granules to improve feature representation [20]. Chen et al. enhanced feature expression through information granulation and employed a granular convolutional neural network for feature extraction from information granules, thereby improving the recognition efficiency of the convolutional neural network [21].

Current granulation methods assess feature relationships by calculating distances from individual pixels to a fixed reference point. However, the feature points of the same category are often dispersed in the feature space, with different distributions requiring distinct optimal reference points. This presents a substantial challenge for manually identifying the most suitable reference points. Therefore, it is essential to develop a granulation method capable of automatically selecting reference points.

3 Methodology

This section begins by introducing the adaptive granulation method and the feature transformation method. Following this, a comprehensive analysis is presented on the integration and application of these methods within the transformer. Finally, this section defines the remote sensing image information granulation transformer for semantic segmentation.

3.1 Overview of the Model Architecture

As shown in Fig. 2a, the model is divided into a feature extraction phase and an up-sampling phase. In the feature extraction phase, the deep semantic features of the image are extracted using both the Swin with AG and ResNet-50. The semantic features extracted by ResNet 50 will be further processed by a dilated convolutional block for feature extraction. The features extracted by Swin with AG are then concatenated with the semantic features processed by the dilated convolutional block. In the up-sampling phase, the feature resolution is progressively restored through up-sampling and the residual block. Finally, the segmentation result is obtained through a 1×1 convolution.

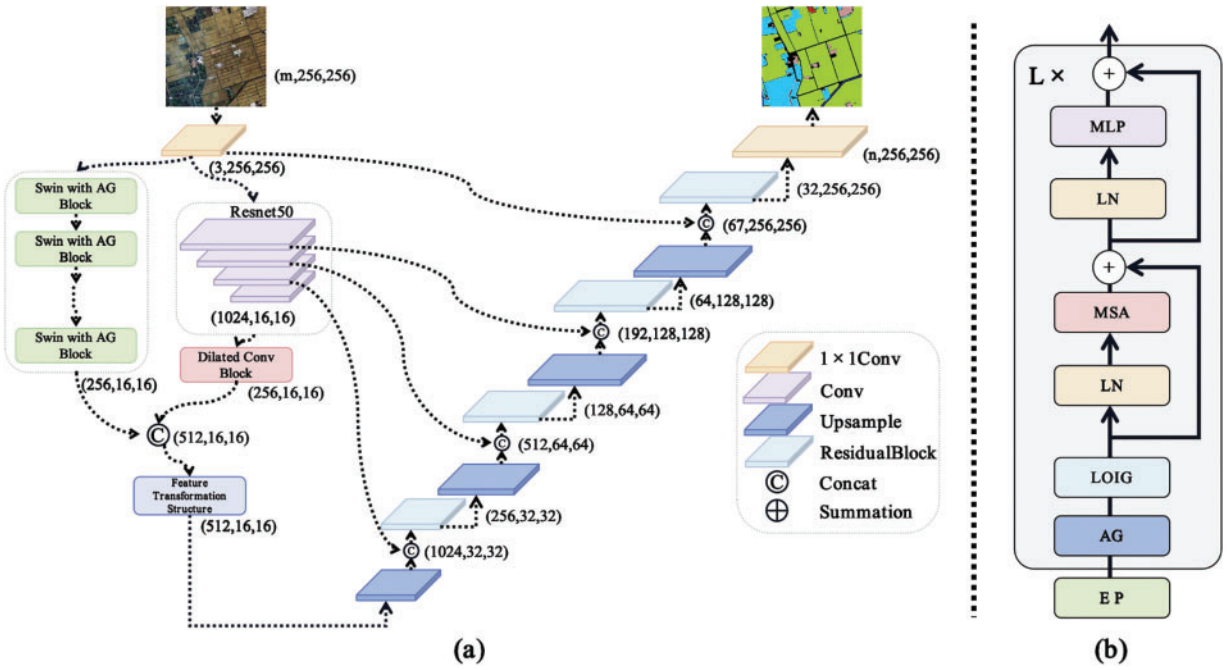

Figure 2: Schematic structure of remote sensing image information granulation transformer for semantic segmentation and the transformer block with adaptive granulation (AG)

There has been much interest in how to solve the problem of gradient vanishing in deep neural networks. Residual connections [22] and the oriented stochastic loss descent proposed by Abuqaddom et al. [23] provide different perspectives for addressing this problem. The transformer architecture leverages residual connections to address this issue. Fig. 2b shows the structure of the transformer block with adaptive granulation. The input represented as embedded patches (EP) is processed through adaptive granulation and the Laplacian operator of information granule (LOIG). The outputs of these modules are then passed to the core components of the transformer block, including layer normalization (LN), multi-head self-attention mechanism (MSA), and multilayer perceptron (MLP). Residual connections are applied after the MSA and MLP using summation operations. This structure is iteratively stacked L times.

3.2 Adaptive Granulation Methods

This chapter provides a comprehensive introduction to adaptive granulation methods. First, Section 3.2.1 defines the concept of adaptive granulation. Subsequently, Section 3.2.2 introduces the matrix representation of information granules. Section 3.2.3 presents the application of information granulation to image patches.

3.2.1 Definition of Adaptive Granulation

For the information system $U = \{X, C\}$. $X = \{x_1, x_2, x_3, ..., x_n\}$ is a sample set, where n represents the number of samples. $C = \{c_1, c_2, c_3, ..., c_m\}$ is the attribute set, where m denotes the number of attributes. Each sample possesses m attributes. $p = \{p_1, p_2, p_3, ..., p_m\}$ is an adaptive reference sample set. p_c denotes the reference sample under attribute c. The granulation method for the sample set X with respect to any

property c is defined as follows:

$$g(X, c) = \{r_i\}_{i=1}^{n} = \{r_1, r_2, ..., r_n\}, \tag{1}$$

where $r_i = \frac{1}{1+exp(x_i+p_c)}$.

$g(X, c)$ denotes the information granule of the sample set X under attribute c. r_i denotes the information kernels constituting the information granule. The calculation of r_i is based on the relationship between the sample x_i and the reference point corresponding to attribute c.

The collection of information granules for the sample set X paired with the attribute set $C = \{c_1, c_2, ..., c_m\}$ can be expressed as:

$$g(X, C) = \{g(X, c_1), g(X, c_2), ..., g(X, c_m)\}. \tag{2}$$

$g(X, C)$ represents the set of all information granules of the sample set X under the attribute set C. Each information granule is defined as a fundamental feature unit that provides a uniform representation of features. It is essentially characterized as a collection, where each value explicitly describes the spatial relationship between a sample x and an adaptive reference point p. Additionally, each value within the information granule quantitatively represents the degree of correlation with the feature encapsulated by the granule.

3.2.2 Matrix Representation for Information Granules

Since information granules are inherently collections, they cannot be directly processed by neural networks. To address this limitation, we define a matrix representation for the collection of information granules.

For the information system $U = \{X, C\}$, the matrix representation of the information granule set $g(X, C)$ is defined as follows:

$$g(X, C) = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix}. \tag{3}$$

Eq. (3) provides the matrix representation of $g(X, C)$. In the matrix representation of information granules, each column corresponds to a distinct information granule, while each row represents a specific attribute within the attribute set $C = \{c_1, c_2, c_3, \ldots, c_m\}$. r_{ij} represents the i-th granular kenerl in the j-th information granule. Once the set of information granules $g(X, C)$ is transformed into a feature matrix, various matrix operations such as addition, subtraction, multiplication, division, and decomposition can be systematically performed.

3.2.3 Adaptive Granulation Method Based on Image Patch

Based on the adaptive granulation method, we take the input image patches of the transformer as the objects of information granulation. Its information granulation method is as follows:

Each patch is considered an attribute in the attribute set $C = \{c_1, c_2, \ldots, c_m\}$. The feature points in each patch are considered as a single sample in the sample set $X = \{x_1, x_2, \ldots, x_n\}$. In our work, sample set $p = \{p_1, p_2, \ldots, p_m\}$ is a learnable parameter. The reference set p effectively identifies similar boundary elements across different classes by determining appropriate reference points during the model's training process. Subsequently, the information granulation operation is applied to the image patches using Eq. (1).

In our work, each patch is granulated into information granules. The information granules generated from all the patches collectively form a set of information granules, which are represented using the matrix representation defined in Section 3.2.2. The collection of information granules, represented in matrix form, enables the execution of any basic matrix-related operations.

3.3 The Laplacian Operator of Information Granule

The Laplacian operator captures second-order variations in the input image, serving as a parameter-free and harmonious method for extracting high-frequency details such as edges and contours. To improve computational efficiency, we approximate the Laplacian operator using a difference operator. Assume that the set of information granules of sample set X over attribute set C is denoted as $g(X, C)$. Based on the matrix representation of information granules, the Laplacian operator of information granule is defined as follows:

$$g(X, C) = g(X, C) - u(d(g(X, C))). \tag{4}$$

As defined in Eq. (4), d represents downsampling with a stride of 2, and u represents upsampling. The downsampling operation primarily reduces the spatial dimensionality of the information granule. This effectively decreases the maximum representable frequency within the granule. As a result, information near the high-frequency region in the frequency domain is lost, leaving primarily low-frequency information. The upsampling operation is primarily used to restore the original matrix dimensions of the information granule. The high-frequency information, including edge details within the information granule, can be obtained by calculating the difference between the granule and its low-frequency components.

3.4 Feature Transformation Architecture

Within the field of deep learning, feature enhancement is a technique that expands the feature representation of input data, thereby effectively improving the performance of the model [24]. This paper designs a feature transformation architecture for feature augmentation. The feature transformation architecture comprises 1×1 convolutions followed by $LeakyReLU$ activation functions, which processes the input features through multiple branching augmentations.

Fig. 3 illustrates the schematic diagram of the feature transformation architecture. Within the feature transformation architecture the process begins with an input X. This input is then partitioned into two distinct subsets labeled X_1 and X_2 using the split operation. Each branch independently undergoes 1×1 convolutions, followed by activation with LeakyReLU and Sigmoid. The outputs from these transformations are then fused through summation to generate intermediate features Y_1 and Y_2. Finally, Y_1 and Y_2 are concatenated and further aggregated through an additional summation operation to produce the final output Y.

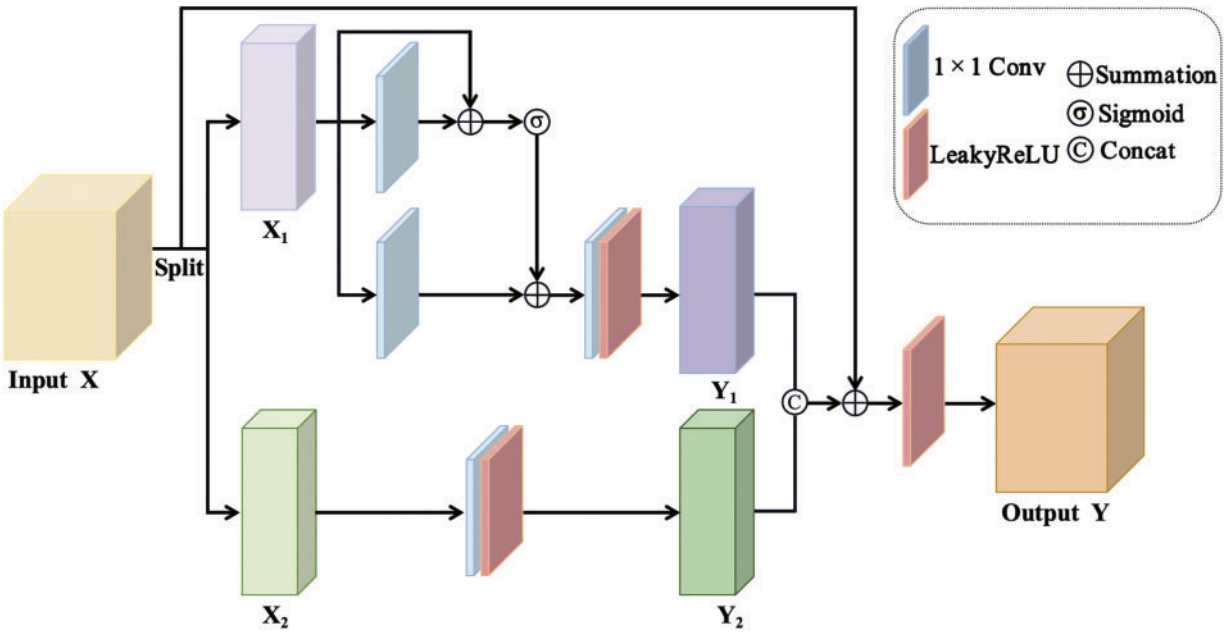

Figure 3: Illustrative schematic of the feature transformation architecture

3.5 Dilated Convolutional Block

Fig. 4 illustrates the structure of a dilated convolutional block, designed to efficiently extract multi-scale features through the use of dilated convolutions with varying dilation rates. The input is processed through four parallel branches, each involving a 3 × 3 convolution with dilation rates of 1, 3, 6, and 9, respectively. Each convolutional layer is followed by a batch normalization (BatchNorm) layer and a ReLU activation function. The outputs of the four branches are concatenated to aggregate features from multiple receptive fields. The concatenated result is subsequently processed by a 1 × 1 convolution, followed by another batch normalization layer and a ReLU activation. The final output represents the enhanced multi-scale feature map.

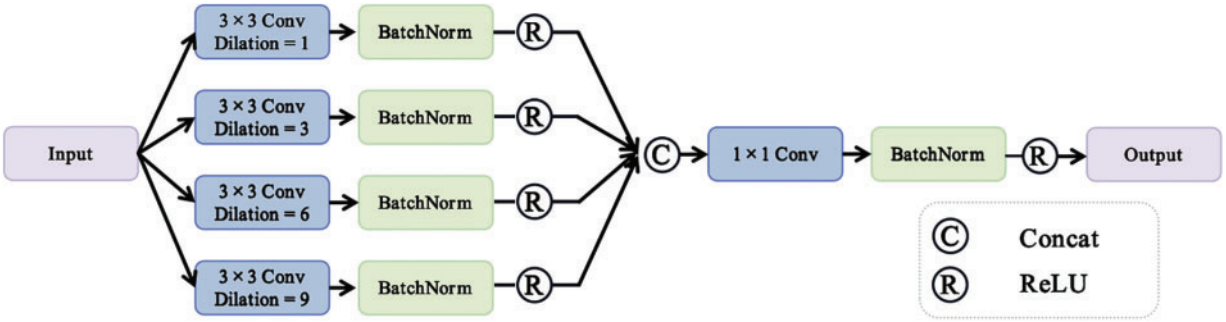

Figure 4: Schematic diagram of the dilated convolutional block

4 Experiments and Analysis

This chapter presents the experiments and analysis. Section 4.1 introduces the datasets and configurations used in the experiments. Section 4.2 describes and analyzes the results of the comparative experiments. Section 4.3 details the ablation experiments, which validate the effectiveness of the proposed components.

4.1 Datasets and Experiment Setting

This study selects the BLU, GID, and PD as the experimental datasets. A detailed introduction to these datasets is provided below:

(1) The BLU dataset is a high-resolution satellite benchmark dataset developed to support multi-class semantic segmentation research in remote sensing. Collected in June 2018 by the Beijing-2 satellite from 21st Century Aerospace Technology Co., Ltd., the dataset comprises RGB optical images with a ground sampling distance of 0.8 meters. It features fine-grained annotations across six land-use categories: background, building, vegetation, water, farmland, and road.

(2) The GID dataset is a large-scale land-cover dataset created from high-resolution images of the Gaofen-2 (GF-2) satellite. It consists of two components: a large-scale classification set with 150 GF-2 images and pixel-level annotations, and a fine classification set with 30,000 multi-scale image patches and 10 annotated GF-2 images. The GlD dataset includes 15 categories: industrial land (IDL), urban residential (UR), rural residential (RR), traffic land (TL), paddy field (PF), irrigated land (IL), dry cropland (DC), garden plot (GP), arbor woodland (AW), shrub land (SL), natural grassland (NG), artificial grassland (AG),river (RV), lake (LK) and pond (PN).

(3) The PD dataset is a high-resolution aerial image dataset used for semantic segmentation in urban scenes. It contains 38 ortho-rectified aerial images, each with a resolution of 6000 × 6000 pixels and a ground sampling distance of 5 cm. The dataset includes six labeled classes: buildings, trees, low vegetation, impervious surfaces, cars, and clutter/background.

In this experiment, a Tesla V100 GPU with 16 GB video memory is used for training, and the deep learning framework is Pytorch. The visual tasks selected in the experiment are segmentation tasks. In this experiment, the input size is set to 256 ∗ 256 pixels, with a learning rate of 1e−4, using Adam as the optimizer and employing a cosine annealing strategy for learning rate optimization. The epoch for training is 100 rounds. All image processing methods used in this experiment are referenced from the methods described in [25].

4.2 Comparison Experiments

To evaluate the efficacy of our method, we conduct comparative analyses with predominant semantic segmentation networks, we utilize data from the original paper (results not provided in the original papers are denoted as "-"). In all tables, bolded numbers signify optimal results.

4.2.1 Comparison with State-of-the-Art Methods on the BLU

To evaluate the performance of the proposed model, experiments were conducted on the BLU dataset. Fig. 5 presents a visual comparison of segmentation results among the proposed model, WicoNet, LANet, Deeplab v3+, and DANet. Table 1 compares the quantitative metrics of the proposed model with those of state-of-the-art models in recent years. The experimental results demonstrate that the proposed model achieves the best segmentation performance. Although the F1 score for the road class is slightly lower, all other evaluation metrics surpass those of the comparison models.

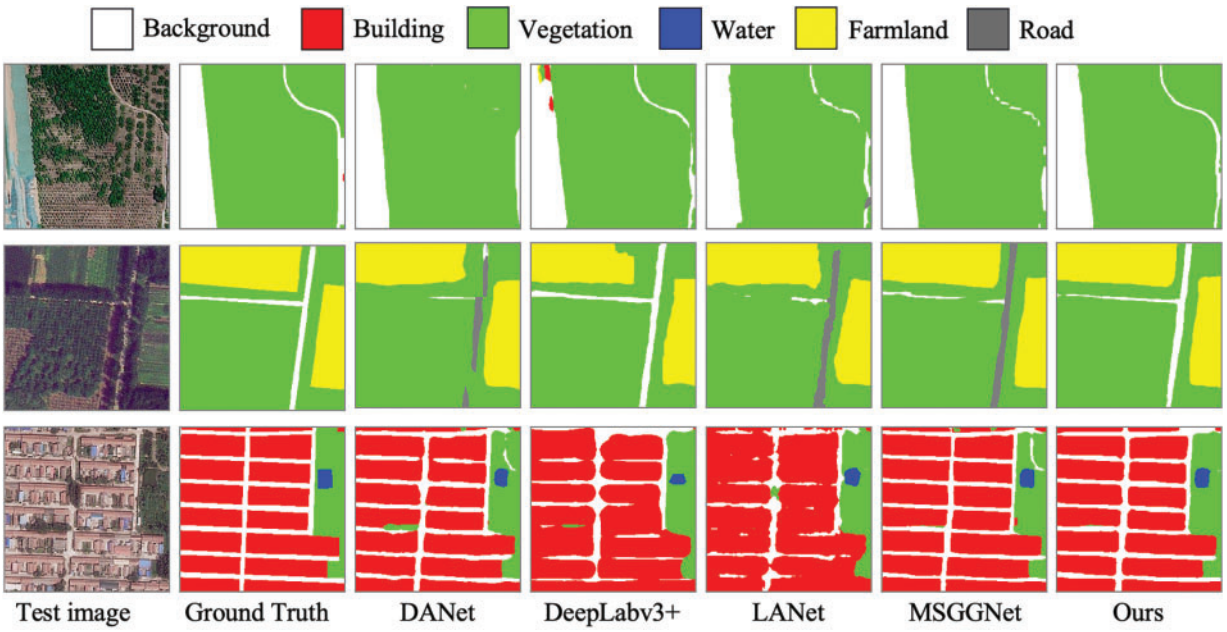

Figure 5: Visualization of segmentation outcomes for various methods applied to the BLU

Table 1: Comparative analysis of segmentation indices across various segmentation methods applied to BLU

Method	Per-class F1 (%)						mOA (%)	mF1 (%)	mIoU (%)
	Background	Building	Vegetation	Water	Agriculture	Road			
PSPNet [26]	72.66	87.40	90.41	85.15	86.42	68.88	86.59	82.05	70.35
DeepLabv3+ [27]	73.99	87.93	90.76	86.46	87.32	68.85	87.08	82.55	71.07
DANet [28]	73.06	87.73	90.55	85.45	86.77	69.07	86.76	82.10	70.40
SCAttNet [29]	73.21	87.62	90.54	86.26	86.87	69.32	86.77	82.30	70.68
MSCA-Net [30]	73.71	88.34	90.74	85.92	86.86	70.31	87.17	82.64	71.21
LANet [31]	73.81	87.48	90.60	85.99	87.02	68.49	86.89	82.28	70.60
UnetFormer [32]	–	–	–	–	–	–	86.04	80.66	68.56
WiCoNet [25]	57.91	78.83	83.13	75.91	77.65	52.58	86.99	82.51	71.02
ST_Unet [33]	54.15	76.47	81.87	80.89	86.33	67.32	86.43	80.86	68.68
MSGCNet [34]	–	–	–	–	–	–	87.16	82.34	70.85
TCNet [35]	–	–	–	–	–	–	87.42	82.93	71.58
OURS	75.23	88.59	91.12	87.30	87.86	68.95	87.61	83.11	71.90

Fig. 5 illustrates the segmentation results of our model alongside other models, including DANet, DeepLabv3+, LANet, and WiCoNet, on the BLU dataset. The comparison highlights the superior segmentation performance of our model, particularly in preserving fine-grained details and accurately identifying boundaries between different classes. Our model demonstrates exceptional capability in processing complex spatial structures, such as small buildings, road networks, and farmland patches, which are often misclassified or poorly segmented by other methods. This performance improvement is evident in the sharper boundaries, reduced noise, and higher consistency with the ground truth, as compared to other models. These results validate the effectiveness of our proposed approach in achieving more precise and robust segmentation outcomes.

Table 1 reports the quantitative comparison results on the BLU dataset. SCAttNet integrates both spatial and channel attention mechanisms, demonstrating superior performance in key metrics when compared to DANet and PSPNet. WiCoNet leverages a contextual transformer to surpass UnetFormer and LANet, achieving higher performance. The integrated multi-scale interaction module of MSGCNet outperforms DeepLabv3+, which incorporates the ASPP module and a distinctive decoder structure. By incorporating relevant category semantic enhancement modules, TCNet respectively surpasses MSGCNet by 0.26%, 0.59%, and 0.73% in the mOA, mF1, and mIoU metrics. The proposed model. The proposed model achieves improvements of 0.19%, 0.18%, and 0.32% over TCNet in terms of mOA, mF1, and mIoU, respectively. This illustrates that TCNet facing high intra-class variance is still insufficient. In contrast, the proposed model enhances the accuracy of information processing by computing intrinsic data relationships through AG to form information granules.

4.2.2 Comparison with State-of-the-Art Methods on the GID

We further validated the effectiveness of the proposed model on the GID dataset. Fig. 6 illustrates the visual segmentation results of representative models. Table 2 presents the performance of each semantic segmentation method on the GID dataset. Overall, the proposed model achieves the best comprehensive performance, although it does not reach the optimal level in terms of mIoU.

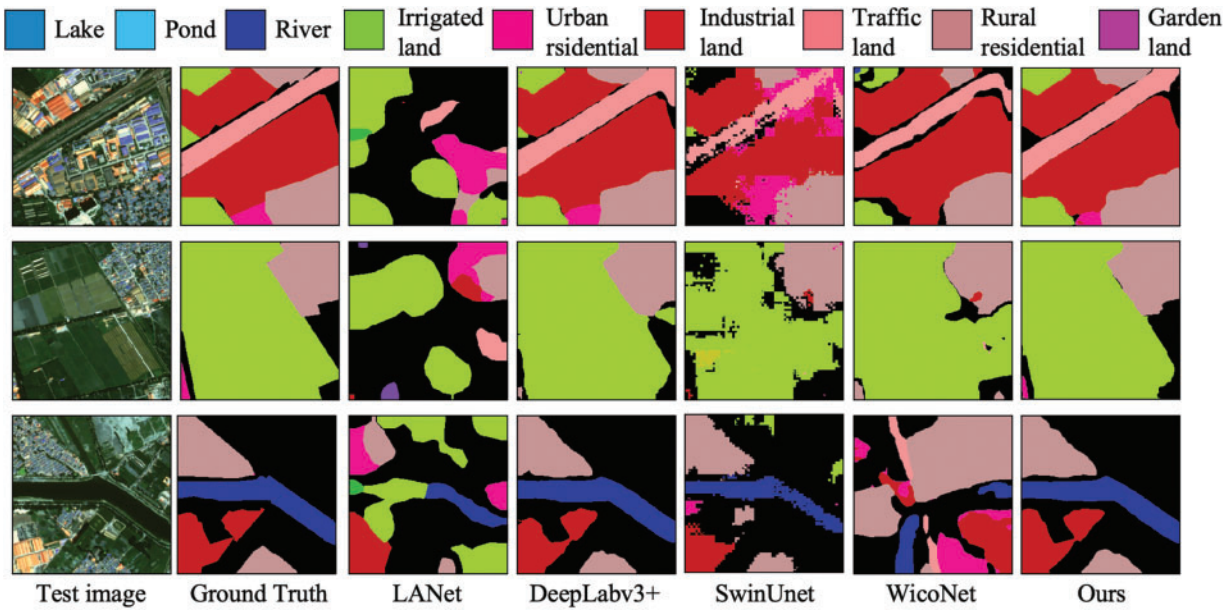

Figure 6: Visualization of segmentation outcomes for various methods applied to the GID

Table 2: Comparative evaluation of segmentation metrics for different methods applied to GID

Method	Per-class F1 (%)															mOA (%)	mF1 (%)	mIoU (%)
	IDL	UR	RR	TL	PF	IL	DC	GP	AW	SL	NG	AG	RV	LK	PN			
PSPNet [26]	59.84	76.29	58.50	67.70	74.82	82.45	39.23	31.69	85.34	8.97	73.07	62.79	83.11	76.70	75.94	75.44	64.41	50.44
DeepLabv3+ [27]	60.44	76.67	58.49	67.67	75.25	82.50	38.62	33.03	84.39	7.58	71.12	64.83	83.17	74.60	74.93	75.38	64.27	50.21
DANet [28]	62.53	76.50	56.73	68.08	75.65	82.76	38.03	26.72	85.75	7.13	73.99	62.95	83.45	77.68	77.25	75.68	64.70	50.81
SCAttNet [29]	61.87	77.32	59.19	68.75	75.29	82.29	35.75	33.32	86.31	12.62	71.53	74.26	81.72	80.96	80.67	76.05	65.59	52.01
MSCA-Net [30]	62.06	77.27	56.51	68.69	74.36	82.46	35.99	24.51	87.08	5.66	72.75	70.65	83.78	78.61	79.09	76.10	65.33	51.60
LANet [31]	63.65	77.67	58.77	69.13	76.80	82.71	37.01	25.68	86.14	16.00	72.42	73.58	84.55	83.53	82.02	76.75	66.06	52.83
HRCNet [36]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	84.50	77.80	71.20
MSGGNet [34]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	82.08	65.85	53.58
C2F [37]	76.87	82.19	74.62	74.25	85.42	93.05	55.55	6.33	55.46	24.17	93.51	47.91	78.80	33.15	77.21	73.61	63.90	51.51
MATNet [38]	74.32	81.38	80.42	79.30	71.60	84.73	77.41	71.85	76.92	70.05	74.16	67.32	65.36	87.06	89.26	79.13	76.74	70.36
GFFNet [39]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	87.10	79.9	73.70
OURS	88.68	83.80	88.64	81.95	82.00	90.12	92.91	84.44	64.22	91.95	79.77	71.19	61.27	94.31	84.72	88.81	82.64	71.50

Fig. 6 presents the segmentation visualizations across different models. In the semantic segmentation results, our proposed model achieves the best boundary delineation performance between different LCLU classes. Our model demonstrates the highest segmentation quality, whereas the comparison models exhibit incomplete segmentation or misclassification in certain image regions. The road portion in the example image is heavily occluded by trees and the pixel feature representation of the road is severely weakened. Compared to the rest of the models, our segmentation is the best.

Table 2 provides a comparative evaluation of segmentation metrics for different methods applied to the GID dataset. DANET integrates the dual-attention mechanism and outperforms PSPNet, DeepLabv3+, and C2F in composite metrics. MANet integrates a novel kernel attention mechanism that more effectively captures the contextual information in remote sensing images. In the experiments, MANet outperforms SCAttNet and LANet across all performance metrics. GFFNet integrates a grouped Transformer structure with grouped convolutions to model spatial information, and achieves higher performance metrics than HRCNet and MSGGNet in the experiments. Although MDANet can fully obtain spatial information, its feature extraction ability for boundary details is poor, so there is still space for improvement. In the experiments, the mOA and mF1 metrics of MDANet are 1.71% and 2.74% lower than those of the proposed model. This is primarily because the LOIG can effectively extract boundary information. Notably, GFFNet achieves 2.2% higher mIoU than the proposed model.

4.2.3 Comparison with State-of-the-Art Methods on the PD

We further demonstrate the effectiveness of the proposed model on the PD dataset. Fig. 7 presents a visual comparison of the segmentation results obtained by the proposed model with those produced by CMLFormer, DeepLabv3+, SwinUnet, and UnetFormer. Table 3 reports the quantitative comparison results on the PD dataset. Based on the experimental results, our proposed model not only demonstrates superior segmentation performance but also achieves optimal results across all comprehensive evaluation metrics.

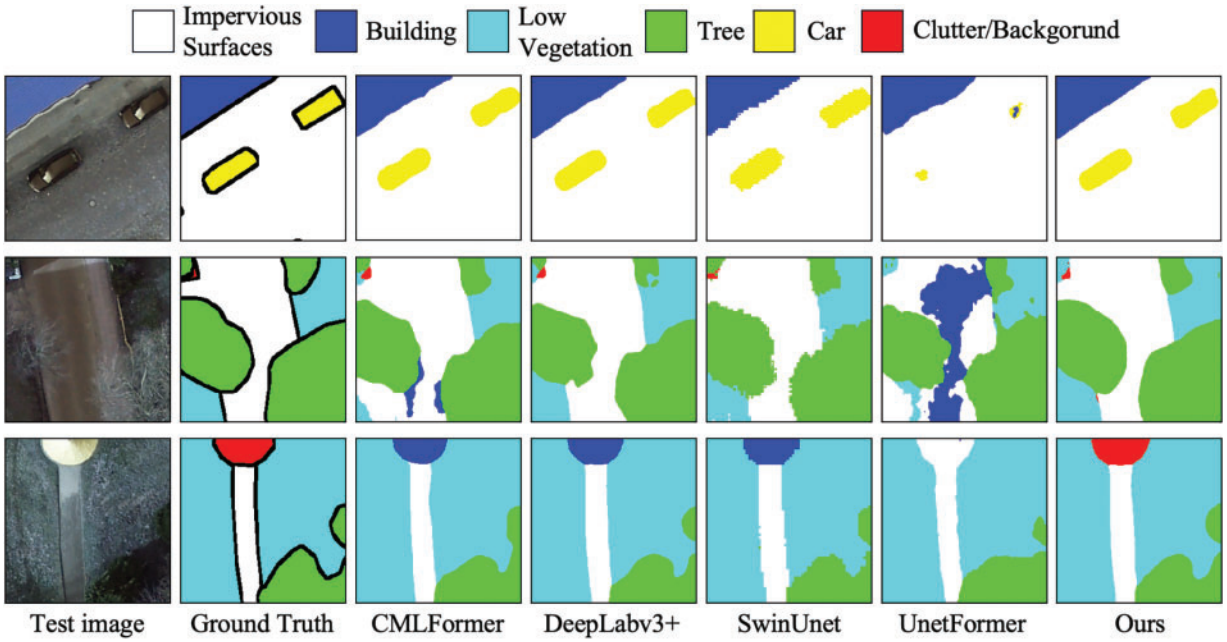

Figure 7: Comparison of segmentation results across different models on the PD

Table 3: Comparative assessment of segmentation metrics for various methods applied to the PD

Method	Per-class F1 (%)					mOA (%)	mF1 (%)	mIoU (%)
	Imp surf	Building	Low veg	Tree	Car			
PSPNet [26]	87.87	92.81	81.83	80.02	84.66	85.51	85.44	74.86
DeepLabv3+ [27]	87.71	92.69	81.72	79.57	84.03	87.23	85.14	74.42
TansFuse [40]	89.75	93.92	82.91	83.61	88.51	86.71	87.74	78.40
SCAttNet [29]	90.04	94.05	84.05	79.75	89.06	87.97	87.39	77.94
MsanlfNet [41]	89.08	93.64	82.37	80.49	86.14	86.41	86.34	76.28
SwinUnet [42]	90.30	94.41	83.05	82.83	88.84	–	87.89	78.68
CSBNet [43]	91.75	95.81	84.91	84.73	90.12	–	88.81	81.20
CTFNet [44]	91.48	96.30	86.04	86.00	91.70	89.38	90.70	83.20
UnetFormer [32]	90.14	94.44	83.15	83.15	83.16	–	88.19	79.16
CMLFormer [45]	90.70	94.96	84.01	83.91	90.31	–	88.79	80.06
EMRT [46]	90.87	94.86	84.12	85.23	90.32	88.12	83.59	73.62
MSINet [47]	91.34	94.88	87.10	84.72	95.15	88.88	90.64	83.14
OURS	92.19	96.46	85.58	86.92	95.57	89.49	91.34	84.37

Fig. 7 presents a comparative visualization of segmentation performance across different models on the Potsdam dataset, highlighting the effectiveness of our proposed method. The comparison includes models such as CMLFormer, DeepLabv3+, SwinUnet, and UnetFormer. In the third example image, the Cluster class is misclassified by all other models, whereas our method accurately identifies and classifies these pixels. This demonstrates the robustness of our approach in capturing subtle and fine-grained features that other models fail to distinguish, especially in complex regions with high intra-class variance or overlapping boundaries.

Table 3 presents quantitative comparison between the proposed model and leading approaches published in recent years. TransFuse integrates transformer and convolutional neural network to effectively fuse local and global features, surpassing PSPNet, DeepLabv3+, and MsanlfNet in performance. MSINet employs multi-scale interpolation to extract scale information, providing prior knowledge for segmentation networks and outperforming EMRT and CMLFormer in experiments. CTFNet introduces channel and spatial attention fusion modules, enabling adaptive fusion of deep semantic features with shallow detail features. However, it is prone to feature loss during the fusion process, which consequently leads to reductions of 0.11%, 0.70%, and 1.23% in mOA, mF1, and mIoU, respectively, compared with the proposed model. In contrast, AG and LOIG are able to extract the internal knowledge representations of the data more effectively, thus avoiding the problem of feature loss.

4.2.4 Complexity Comparison

To evaluate the computational resources required by the model, this experiment selected two commonly used metrics for assessment, the total number of parameters and floating point operations per second (FLOPs). Table 4 presents the results of comparisons between the complexities of several representative models.

Table 4: Comparison of FLOPs and Params for different models

Model	PSPNet	DANet	CMLFormer	MSGCNet	DeepLabV3+	SwinU-Net	MsanlfNet	Ours
FLOPs (Gbps)	49.24	49.52	46.39	28.64	31.77	28.06	32.24	28.08
Params (Mb)	51.31	47.44	56.17	27.02	26.12	33.73	69.53	50.56

DANet models the semantic features on both spatial and channel dimensions by introducing the position attention module and channel attention module, respectively. However, the dual attention mechanism requires more computational resources, which results in higher FLOPs for DANet compared to PSPNet and CMLFormer. MsanlfNet, which introduces a multi-scale attention mechanism, requires substantial computational resources. As a result, MsanlfNet has higher FLOPs compared to DeepLabv3+. In addition, the number of parameters in MsanlfNet is also higher than that of PSPNet and CMLFormer. By comparison, it can be observed that the proposed model is similar to MSGCNet and SwinUnet in terms of FLOPs, but its number of parameters is larger than both of these models. This is because the proposed model uses AG to process the image into information granules as objects, which reduces the number of objects the model needs to process to some extent, thereby lowering the required computational resources.

4.2.5 Comparative Experimental Analysis

In the comparative experiments, the proposed model attained the highest scores across the performance metrics. Our model demonstrates superior performance compared to the transformer network employing multi-scale features. This is mainly due to the fact that existing models are susceptible to data inseparability. In contrast, our model, by extracting internal knowledge representations, is better able to capture the intrinsic features of the data, thereby mitigating this issue. However, our model did not achieve the highest F1-score for certain classes. Through our analysis, we conclude that this is primarily due to two factors. Firstly, the imbalanced proportion of sample counts across different categories in the training dataset results in inconsistent training effectiveness for each class. Secondly, when preprocessing the training dataset by slicing large-scale images into smaller ones, the class imbalance within individual images was not taken into consideration. This demonstrates that our proposed model can better handle high-resolution remote sensing images. In the visualization comparison experiments, the model proposed in this study exhibits the best performance in segmenting high-resolution remote sensing images. Particularly in scenarios involving feature objects influenced by complex backgrounds, our model demonstrates superior performance in capturing fine details. This is because our proposed model is less susceptible to receiving high intra-class variance. The effectiveness of our model is validated through comparative experiments.

4.3 Ablation Experiment

In this section, we will evaluate the effectiveness of AG and LOIG in enhancing transformer performance. Section 4.3.1 visually analyzes the attention matrices, demonstrating that our model can capture more critical information. The validity of LOIG for edge feature extraction is verified in Section 4.3.2. Section 4.3.3 validates the performance improvement provided by different components.

4.3.1 Attention Matrix Visualization

The attention mechanism captures global dependencies by calculating the correlations between different positions [48]. In this experiment, by visualizing the attention matrix, it was validated that the proposed AG and LOIG enable the attention mechanism to capture more comprehensive information.

Fig. 8 provides a visualization of the attention matrix. Incorporating AG and the LOIG into Swin enables the model to capture more critical information. The attention mechanisms are averaged and visualized across three datasets, including BLU, GID, and PD, providing a comprehensive analysis of attention behavior. By comparison, the AG and the LOIG demonstrated notable advantages in attention allocation. Specifically, AG and LOIG correspond to matrices where the highlighted regions are clearer and more uniformly distributed, indicating their ability to capture more essential information. In addition, the features extracted by AG and

LOIG exhibit a more organized structure, reflecting their ability to accurately identify and retain the relevant key information.

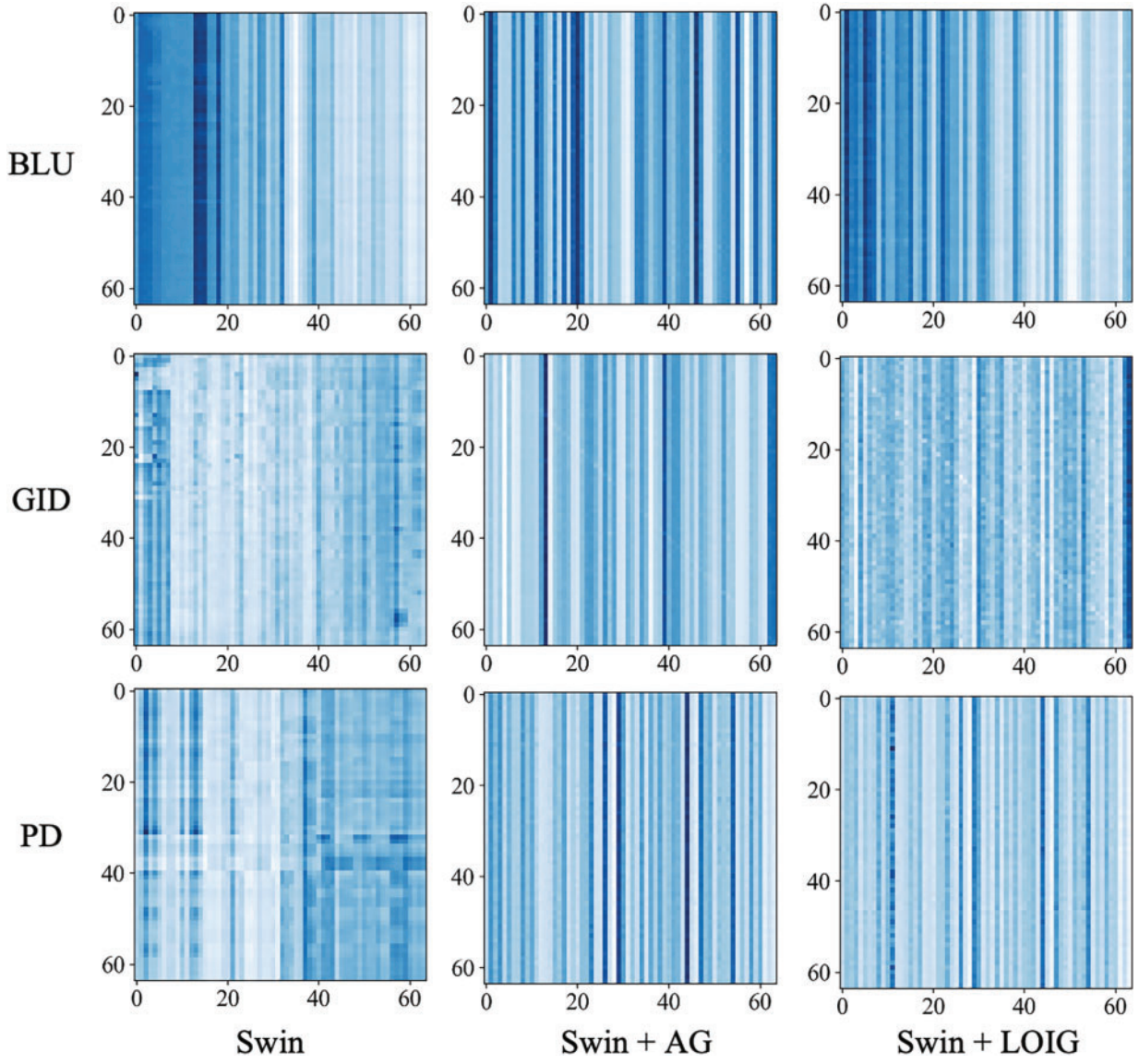

Figure 8: Attention matrix visualization of the AG and the LOIG used in the Swin

4.3.2 Visualization Experiment on the Effectiveness of LOIG

To evaluate the effectiveness of LOIG in boundary feature extraction. In this expriment, Grad-CAM was employed to visually analyze the segmentation results obtained with and without the LOIG model. In the resultant visualizations, the blue regions denote the areas that the model attends to, with a more intense blue hue corresponding to a higher level of attention. The experimental results indicate that LOIG demonstrates significant advantages in boundary feature extraction.

Fig. 9 shows the segmentation visualization results of the model without LOIG and the model with LOIG. In the BLU dataset, the model without LOIG demonstrated a relatively limited capacity for discerning

building and road boundaries, whereas the incorporation of LOIG markedly enhanced boundary detection. In the GID dataset, due to the similarity in object features, clearly delineating the boundary between irrigated land and the background proved challenging. The model without LOIG struggled to effectively address these boundary issues, whereas the introduction of LOIG enabled the model to better capture the boundary characteristics. In the PD dataset, due to the influence of lighting conditions, the edges between parked cars and the road became difficult to distinguish. The AG model's focus deviated from the object, whereas the LOIG model maintained more precise boundary segmentation. The experimental results show that the model with LOIG demonstrates significant advantages in handling complex boundary scenarios.

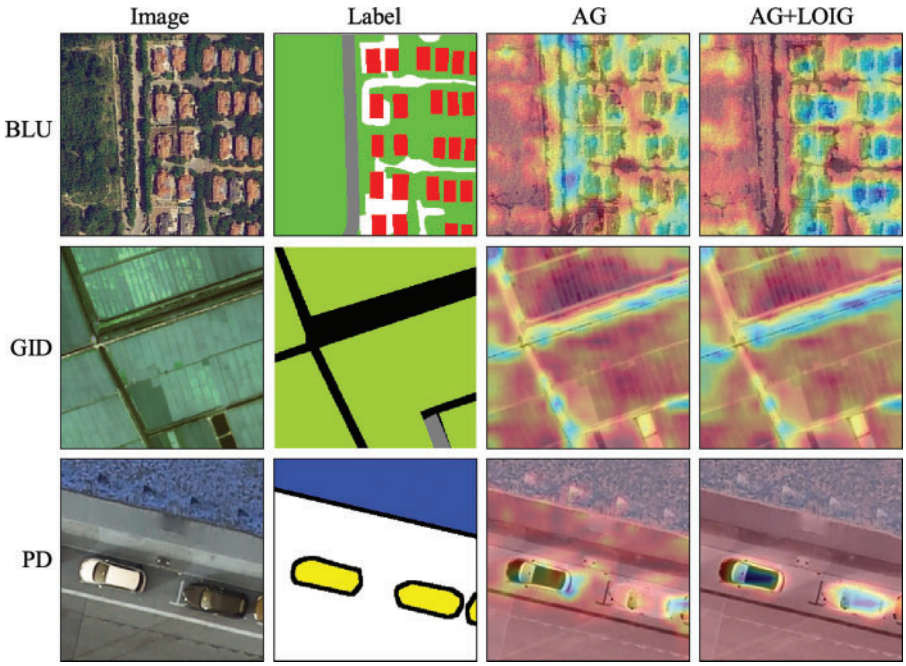

Figure 9: Segmentation visualization results comparing models with and without LOIG

4.3.3 Ablation Experiments with Different Components

In the ablation experiments, we assess the individual contributions of the AG and the LOIG. Table 5 presents the ablation study results across various datasets, providing a comparative analysis of performance.

Table 5: Comparison of segmentation metrics for different components with different datasets

Dataset	Swin	AG	LOIG	mOA (%)	mF1 (%)	mIoU (%)
BLU	✓			87.56	82.82	71.48
	✓	✓		87.46	82.90	71.60
	✓		✓	87.66	83.00	71.79
GID	✓			88.49	82.01	70.72
	✓	✓		88.75	82.80	71.91
	✓		✓	88.49	82.39	71.06
PD	✓			88.02	89.64	81.50
	✓	✓		89.87	91.10	84.01
	✓		✓	89.14	90.66	83.34

Since the model with AG can effectively represent the shared semantic features of data through information granules. This effectively mitigates the impact of high intra-class variance on the attention mechanism, enabling it to focus on more critical information. On the BLU dataset, the model with AG achieves mOA, mF1, and mIoU of 87.46%, 82.90%, and 71.60%, respectively. On the GID dataset, the model with AG improves mOA, mF1, and mIoU by 0.26%, 0.79%, and 1.19%, respectively. On the PD dataset, the model with AG achieves mOA, mF1, and mIoU scores of 89.87%, 91.10%, and 84.01% in order. Since LOIG can effectively extract edge features, it enables the segmentation model to achieve improved performance. As the LOIG can effectively extract edge features, the segmentation model achieves better performance. In the BLU dataset, the model with LOIG achieved mOA, mF1, and mIoU of 87.66%, 83.00%, and 71.79%, respectively. In the GID dataset, the model with LOIG achieved mOA, mF1, and mIoU of 88.49%, 82.39%, and 71.06%, respectively. In the PD dataset, compared with the model without LOIG, the model with LOIG improved mOA, mF1, and mIoU by 1.12%, 1.02%, and 1.84%, in that order.

The experimental metrics substantiate the efficacy of the components proposed in this paper in enhancing the model's performance. This is because adaptive granulation harmonizes the feature representation of terrestrial objects within the same category, while the Laplacian operator of information granule augments the expressive power of boundary features.

4.3.4 Analyze of Ablation Experiments

Ablation experiments include visualization experiments and segmentation metric comparison experiments. In the attention visualization experiments, the network incorporating the AG and the LOIG exhibits richer and more accurate feature attention. The visualization experiments on LOIG effectiveness indicate that models incorporating LOIG achieve superior edge feature perception. In the segmentation metric comparison experiment, the network employing AG and LOIG demonstrates significant improvements in segmentation performance. This is due to the fact that AG can effectively extract the common semantic features of land objects, while LOIG excels at capturing edge features. The ablation study systematically validates the effectiveness of each component proposed in this research.

5 Conclusion

In this paper, remote sensing image information granulation transformer for semantic segmentation is proposed. The model includes Swin with AG, a feature transformation structure, and a dilated convolutional block, among other components. Swin with AG applies adaptive granulation techniques and utilizes the Laplacian operator for information granules. The adaptive granulation method employs a dynamic learning strategy to ascertain the optimal reference point and produce information granules. Information granules can represent common semantic features for similar pixels. This attribute is leveraged to address the substantial discrepancies in feature representations among objects belonging to the same class. The core of Laplacian operator of the information granule lies in using the Laplace operator to extract the boundary features of the objects. This facilitates the expression of boundary features through information granules.

In the comparison experiments, our model achieves superior performance across segmentation metrics. In particular, the model achieves mOA of 88.81%, mF1 score of 82.64%, and mIoU of 17.50% on the GID dataset. Meanwhile, on the PD dataset, the model reaches mOA of 89.49%, mF1 score of 91.34%, and mIoU of 84.37%. It also demonstrates outstanding segmentation quality in the visualized segmentation results. In the complexity comparison experiment, the model's FLOPs are 28.08. This indicates that the model maintains higher inference performance while consuming fewer computational resources. The attention visualization experiment indicates that the model utilizing the proposed method captures key information

more comprehensively. In ablation studies, the effectiveness of the adaptive granulation and the Laplacian operator of the information granule was validated separately.

Although the proposed model exhibits high segmentation performance, there are still some limitations. For example, our model does not achieve optimal performance on certain types of ground objects. Upon analysis, we found that this is due to the differing proportions of various classes in the training dataset. Classes with a smaller proportion will not receive sufficient training. In the future, we will explore various methods for granulating information to describe data from multiple perspectives, such as the adaptive multi-granularity method.

Acknowledgement: All authors are thankful for the useful and constructive comments from the editors and reviewers.

Funding Statement: This research was supported by the National Natural Science Foundation of China (62462040), the Yunnan Fundamental Research Projects (202501AT070345), the Major Science and Technology Projects in Yunnan Province (202202AD080013), Sichuan Provincial Key Laboratory of Philosophy and Social Science Key Program on Language Intelligence Special Education (YYZN-2024-1) and the Photosynthesis Fund Class A (ghfund202407010460).

Author Contributions: Haoyang Tang: Conceptualization, Methodology, Software, Data curation validation, Writing—original draft, Resources, Writing—review editing. Kai Zeng: Supervision, Investigation, Formal analysis, Project administration, Funding acquisition. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data and materials supporting the findings of this study are openly available at https://github.com/sjmp525/RSIGT.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Liu C, Chen K, Zhang H, Qi Z, Zou Z, Shi Z. Change-agent: towards interactive comprehensive remote sensing change interpretation and analysis. IEEE Trans Geosci Remote Sens. 2024;62:5635616. doi:10.1109/TGRS.2024.3425815.
2. Shi Z, Fan J, Du Y, Zhou Y, Zhang Y. LULC-SegNet: enhancing land use and land cover semantic segmentation with denoising diffusion feature fusion. Remote Sens. 2024;16(23):4573. doi:10.3390/rs16234573.
3. Tan H, Sun S, Cheng T, Shu X. Transformer-based cloud detection method for high-resolution remote sensing imagery. Comput Mater Contin. 2024;80(1):661–78. doi:10.32604/cmc.2024.052208.
4. Zhang X, Jiang L, Wang L, Zhang T, Zhang F. A pruned-optimized weighted graph convolutional network for axial flow pump fault diagnosis with hydrophone signals. Adv Eng Inform. 2024;60:102365. doi:10.1016/j.aei.2024.102365.
5. Jamali A, Roy SK, Hong D, Atkinson PM, Ghamisi P. Spatial gated multi-layer perceptron for land use and land cover mapping. IEEE Geosci Remote Sens Lett. 2024;21:5502105.
6. Zhang X, Zhang X, Liu J, Wu B, Hu Y. Graph features dynamic fusion learning driven by multi-head attention for large rotating machinery fault diagnosis with multi-sensor data. Eng Appl Artif Intell. 2023;125:106601.
7. Zheng Z, Zhong Y, Wang J, Ma A. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020; Seattle, WA, USA.
8. Pang J, Li C, Shi J, Xu Z, Feng H. R^2-CNN: fast tiny object detection in large-scale remote sensing images. IEEE Trans Geosci Remote Sensing. 2019;57(8):5512–24.
9. Wang B, Wang Z, Sun X, Wang H, Fu K. DMML-Net: deep metametric learning for few-shot geographic object segmentation in remote sensing imagery. IEEE Trans Geosci Remote Sensing. 2021;60:1–18.

10. Li X, Xu F, Liu F, Lyu X, Tong Y, Xu Z, et al. A synergistical attention model for semantic segmentation of remote sensing images. IEEE Trans Geosci Remote Sens. 2023;61:1–16.
11. Yao JT, Vasilakos AV, Pedrycz W. Granular computing: perspectives and challenges. IEEE Trans Cybern. 2013;43(6):1977–89.
12. Liu H, Li W, Li R. A comparative analysis of granular computing clustering from the view of set. J Intell Fuzzy Syst. 2017;32(1):509–19. doi:10.3233/JIFS-152327.
13. Qin J, Martínez L, Pedrycz W, Ma X, Liang Y. An overview of granular computing in decision-making: extensions, applications, and challenges. Inf Fusion. 2023;98:101833. doi:10.1016/j.inffus.2023.101833.
14. Zhang Y, Zhu Z, Xia Z, Deng C, Tashi N, Cheng J. High-order transformer semantic segmentation network for high-resolution remote sensing images. In: IGARSS, 2024—2024 IEEE International Geoscience and Remote Sensing Symposium; 2024; Athens, Greece. p. 9774–7.
15. Yang F, Jiang F, Li J, Lu L. MSTrans: multi-scale transformer for building extraction from HR remote sensing images. Electronics. 2024;13(23):4610. doi:10.3390/electronics13234610.
16. Yu C, Hong L, Pan T, Li Y, Li T. ESTUGAN: enhanced swin transformer with U-Net discriminator for remote sensing image super-resolution. Electronics. 2023;12(20):4235. doi:10.3390/electronics12204235.
17. Yang Y, Jiao L, Liu F, Liu X, Li L, Chen P, et al. An explainable spatial–frequency multiscale transformer for remote sensing scene classification. IEEE Trans Geosci Remote Sens. 2023;61(6):1–15. doi:10.1109/TGRS.2023.3265361.
18. Yang H, Zhou C, Xing X, Wu Y, Wu Y. A high-resolution remote sensing road extraction method based on the coupling of global spatial features and fourier domain features. Remote Sens. 2024;16(20):3896. doi:10.3390/rs16203896.
19. Yu H, Wang J, Zhao SY, Silver O, Liu Z, Yao J, et al. GGI-DDI: identification for key molecular substructures by granule learning to interpret predicted drug–drug interactions. Expert Syst Appl. 2024;240:122500. doi:10.1016/j.eswa.2023.122500.
20. Behzadidoost R, Mahan F, Izadkhah H. Granular computing-based deep learning for text classification. Inf Sci. 2024;652(2):119746. doi:10.1016/j.ins.2023.119746.
21. Chen Y, Zhang X, Zhuang Y, Yao B, Lin B. Granular neural networks with a reference frame. Knowl Based Syst. 2023;260(2):110147. doi:10.1016/j.knosys.2022.110147.
22. Borawar L, Kaur R. ResNet: solving vanishing gradient in deep networks. In: Proceedings of International Conference on Recent Trends in Computing: ICRTC 2022. New Delhi, India: Springer; 2023. p. 235–47.
23. Abuqaddom I, Mahafzah BA, Faris H. Oriented stochastic loss descent algorithm to train very deep multi-layer neural networks without vanishing gradients. Knowl Based Syst. 2021;230(7553):107391. doi:10.1016/j.knosys.2021.107391.
24. Zhang X, Liu J, Zhang X, Lu Y. Self-supervised graph feature enhancement and scale attention for mechanical signal node-level representation and diagnosis. Adv Eng Inform. 2025;65:103197. doi:10.1016/j.aei.2025.103197.
25. Ding L, Lin D, Lin S, Zhang J, Cui X, Wang Y, et al. Looking outside the window: wide-context transformer for the semantic segmentation of high-resolution remote sensing images. IEEE Trans Geosci Remote Sens. 2022;60:1–13. doi:10.1109/TGRS.2022.3168697.
26. 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; 2017; Honolulu, Hawaii, USA. p. 2881–90.
27. Chen LC, 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); 2018; Munich, Germany. p. 801–18.
28. Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, et al. Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and pAttern Recognition; 2019; Long Beach, CA, USA. p. 3146–54.
29. Li H, Qiu K, Chen L, Mei X, Hong L, Tao C. SCAttNet: semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images. IEEE Geosci Remote Sens Lett. 2021;18(5):905–9. doi:10.1109/LGRS.2020.2988294.
30. Sun Y, Dai D, Zhang Q, Wang Y, Xu S, Lian C. MSCA-Net: multi-scale contextual attention network for skin lesion segmentation. Pattern Recognit. 2023;139(6):109524. doi:10.1016/j.patcog.2023.109524.

31. Ding L, Tang H, Bruzzone L. LANet: local attention embedding to improve the semantic segmentation of remote sensing images. IEEE Trans Geosci Remote Sensing. 2020;59(1):426–35. doi:10.1109/TGRS.2020.2994150.
32. Wang L, Li R, Zhang C, Fang S, Duan C, Meng X, et al. UNetFormer: a UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS J Photogramm Remote Sensing. 2022;190:196–214. doi:10.1016/j.isprsjprs.2022.06.008.
33. He X, Zhou Y, Zhao J, Zhang D, Yao R, Xue Y. Swin transformer embedding UNet for remote sensing image semantic segmentation. IEEE Trans Geosci Remote Sens. 2022;60:1–15. doi:10.1109/TGRS.2022.3230846.
34. Zeng Q, Zhou J, Tao J, Chen L, Niu X, Zhang Y. Multiscale global context network for semantic segmentation of high-resolution remote sensing images. IEEE Trans Geosci Remote Sens. 2024;62:1–13. doi:10.1109/TGRS.2024.3393489.
35. Zhang L, Tan Z, Zhang G, Zhang W, Li Z. Learn more and learn usefully: truncation compensation network for semantic segmentation of high-resolution remote sensing images. IEEE Trans Geosci Remote Sens. 2024;62:1–14. doi:10.1109/TGRS.2024.3510781.
36. Xu Z, Zhang W, Zhang T, Li J. HRCNet: high-resolution context extraction network for semantic segmentation of remote sensing images. Remote Sens. 2020;13(1):71. doi:10.3390/rs13010071.
37. Chen H, Yang W, Liu L, Xia GS. Coarse-to-fine semantic segmentation of satellite images. ISPRS J Photogramm Remote Sens. 2024;217(2):1–17. doi:10.1016/j.isprsjprs.2024.07.028.
38. Li R, Zheng S, Zhang C, Duan C, Su J, Wang L, et al. Multiattention network for semantic segmentation of fine-resolution remote sensing images. IEEE Trans Geosci Remote Sens. 2021;60:1–13.
39. Cao Y, Huo C, Xiang S, Pan C. GFFNet: global feature fusion network for semantic segmentation of large-scale remote sensing images. IEEE J Selected Topics Appl Earth Observ Remote Sens. 2024;17:4222–34. doi:10.1109/JSTARS.2024.3359656.
40. Zhang Y, Liu H, Hu Q. Transfuse: fusing transformers and cnns for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference; 2021 Sep 27–Oct 1; Strasbourg, France: Springer. p. 14–24.
41. Bai L, Lin X, Ye Z, Xue D, Yao C, Hui M. MsanlfNet: semantic segmentation network with multiscale attention and nonlocal filters for high-resolution remote sensing images. IEEE Geosci Remote Sens Lett. 2022;19:1–5. doi:10.1109/LGRS.2022.3185641.
42. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision. Tel Aviv, Israel: Springer; 2022. p. 205–18.
43. He X, Zhou Y, Liu B, Zhao J, Yao R. Remote sensing image semantic segmentation via class-guided structural interaction and boundary perception. Expert Syst Appl. 2024;252(4):124019. doi:10.1016/j.eswa.2024.124019.
44. Zhang Z, Chen Y, Zhang D, Qian Y, Wang H. CTFNet: long-sequence time-series forecasting based on convolution and time–frequency analysis. IEEE Trans Neural Netw Learn Syst. 2023;35(11):16368–82. doi:10.1109/TNNLS.2023.3294064.
45. Wu H, Zhang M, Huang P, Tang W. CMLFormer: CNN and multi-scale local-context transformer network for remote sensing images semantic segmentation. IEEE J Selected Topics Appl Earth Obs Remote Sens. 2024;17:7233–41. doi:10.1109/JSTARS.2024.3375313.
46. Xiao T, Liu Y, Huang Y, Li M, Yang G. Enhancing multiscale representations with transformer for remote sensing image semantic segmentation. IEEE Trans Geosci Remote Sens. 2023;61:1–16. doi:10.1109/TGRS.2023.3256064.
47. Peng C, Li H, Tao C, Li Y, Ma J. MSINet: mining scale information from digital surface models for semantic segmentation of aerial images. Pattern Recognit. 2023;143(3):109785. doi:10.1016/j.patcog.2023.109785.
48. Zhang X, Liu J, Zhang X, Lu Y. Multiscale channel attention-driven graph dynamic fusion learning method for robust fault diagnosis. IEEE Trans Ind Inform. 2024;20(9):11002–13. doi:10.1109/TII.2024.3397401.