# Robust Land Cover Classification With Local–Global Information Decoupling to Address Remote Sensing Anomalous Data

Jianbo Xiao, Taotao Cheng, Deliang Chen, Hui Chen, Ning Li, Yanyan Lu, and Liang Cheng

***Abstract*—Remote sensing images play a critical role in urban planning, land resources, and environmental monitoring. Land cover classification is one of the straightforward applications of remote sensing. However, the anomalous remote sensing data challenges the reliability of land cover classification results. Deep learning has been widely used in remote sensing image analysis, but it remains sensitive to anomalous data. To address this issue, we reevaluate a land cover classification map in high-noise scenarios with anomalous data and propose a novel network architecture to solve the problem. A new network architecture is proposed to solve this problem. Our proposed network architecture focuses on decoupling the extraction of global information and local information. Through three global–local feature fusion modules, we output features emphasizing global information, features emphasizing local information, and consistency evaluation scores, respectively. A specially designed decoder integrates these three features. Our method performs better compared to mainstream models on the public datasets the Wuhan high-definition landscape dataset with obvious anomaly data, with a mean intersection over union (MIoU) of 63.58% and a mean pixel accuracy (Mpa) of 74.32%. Compared to the suboptimal method, our method improves MIoU by 1.29% and Mpa by 3.05%.**

***Index Terms*—Deep learning, land cover classification, remote sensing anomalous data, remote sensing imagery.**

## I. Introduction

REMOTE sensing plays an important role in many areas of geoscience, including urban planning, natural resource management, and environmental monitoring. Among these applications, the classification of land cover types is of paramount importance [1], [2], [3]. The interpretation of remote sensing imagery allows to relate pixels in the image to land surface features, providing both quantitative and qualitative geographic information. This process is of immense value in various fields requiring detailed land cover information, including resource management [4], [5], environmental monitoring [6], [7], and urban planning [8], [9]. However, in the real-world applications, remote sensing images not only contain abundant and complex terrain information [10] but also suffer from challenges due to factors, such as atmospheric errors [11], terrain distortions [12], and sensor malfunctions [13]. These problems result in image spectral inaccuracies and random noise, causing various anomalous data problems and posing challenges to the reliability and applicability of land cover classification in remote sensing [14]. Fig. 1 shows the color differences in images collected at the same location and at similar times due to different platforms, which is a typical anomalous data.

In recent years, there have been significant advances in remote sensing land cover classification. Traditional methods include learning-based approaches, such as random forests, as well as measurement-based thresholding, clustering, and other techniques [15], [16], [17] have been extensively explored in land cover classification. These methods either integrate expert knowledge to create features and establish feature-land cover type relationships, or combine expert knowledge to develop a limited set of supervised learning samples and learning models for land cover type classification. For example, Wulder et al. [18] introduced a novel clustering method that improves land cover type classification results by merging and splitting clusters. Chen et al. [19] based their work on an improved feature space transformation that allows clustering methods to process a large amount of data more quickly. Pal et al. [20] summarized the advantages of random forest, including fast computation and absence of statistical assumptions, and enhanced its performance in remote sensing land cover type classification using boosting techniques. However, the effectiveness of these conventional methods is highly dependent on the prior knowledge derived from expert experience. Therefore, they are typically suitable for analyzing small-scale data, which facilitates human summarization of prior knowledge. This specificity makes it difficult for these methods to handle data with anomalies.

With the advancement of deep learning technology, it has been widely applied to remote sensing-based investigations [21], [22], [23], [24], [25]. For example, Kussul et al. [26] used unsupervised neural networks to classify land cover and crop types using multitemporal and multisource satellite imagery. Zhao et al. [27] used a deep learning fusion network in

Manuscript received 27 October 2023; revised 16 January 2024 and 26 January 2024; accepted 28 January 2024. Date of publication 31 January 2024; date of current version 8 March 2024. This work was supported in part by the Jiangsu Province Postgraduate Research and Practice Innovation Plan under Grant KYCX22_0981. *(Corresponding author: Deliang Chen.)*

Jianbo Xiao, Taotao Cheng, and Deliang Chen are with the School of Geography and Bioinformatics, Nanjing University of Posts and Telecommunications, Nanjing 210023, China (e-mail: 1807180656@qq.com; 1022173206@njupt.edu.cn; 122592146@qq.com).

Hui Chen, Ning Li, and Liang Cheng are with the School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China (e-mail: 742543787@qq.com; lining@smail.nju.edu.cn; lcheng@nju.edu.cn).

Yanyan Lu is with the Institute of Natural Resources and Environment Audit, Nanjing Audit University, Nanjing 211815, China (e-mail: cdlyy@nau.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3360458

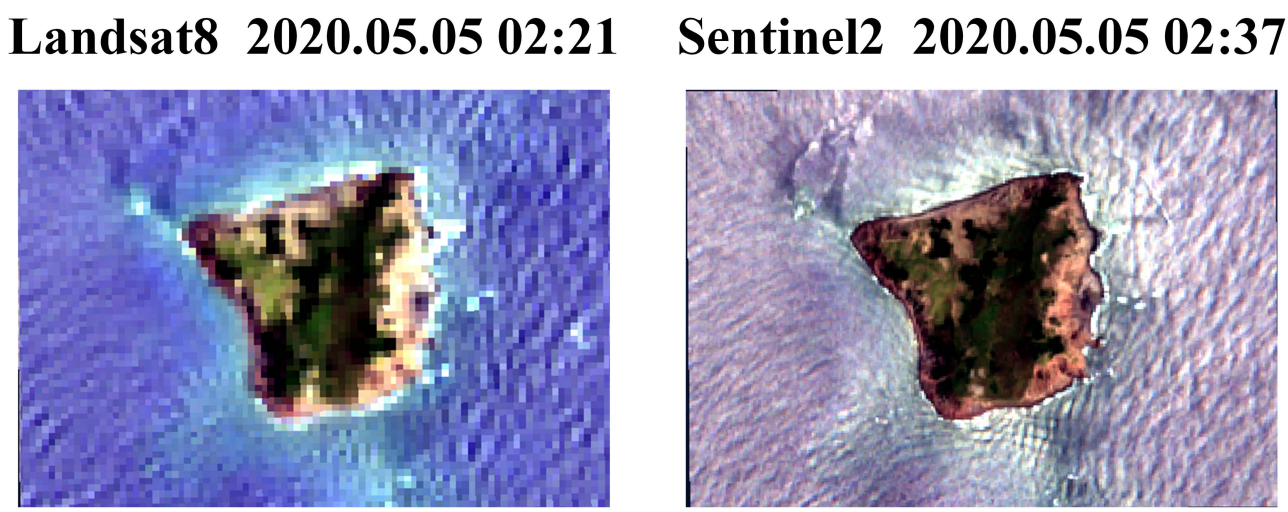


Fig. 1. Differences between different platforms within the same time period. The acquisition times of both images are referenced to coordinated universal time (UTC).

conjunction with normalized difference vegetation index (NDVI) for high-resolution remote sensing land cover classification. Jin et al. [28] modified the structure of the convolutional neural network (CNN) model to improve the classification performance. Deep learning methods are data-driven and automatically learn features and relationships between land cover types from extensive data, resulting in improved robustness to large datasets compared to methods that rely on manual prior knowledge. In deep learning scenarios, Zhao et al. [29] pointed out that the criterion for judging anomalous data is whether it deviates from predefined normality. This predefinition is typically determined by selecting training data. However, traditional deep learning models tend to deeply couple the extraction of global and local features when handling different types of features. This leads to two main problems: first, when the training set contains a large amount of anomalous data, it increases the training complexity and risks overfitting [30], [31]; second, there is a discrepancy in the noise distribution between the training samples and the actual scene, which challenges the generalization performance of the model [32], [33]. As a result, traditional deep learning methods can only handle a certain amount of anomalous data and have difficulty in dealing with the diverse noise in remote sensing images.

To address the problem of performance degradation of current deep learning methods in high-noise scenarios with anomalous data, researchers have worked hard to improve the robustness of these models. For example, Sun et al. [34] introduced a medical image segmentation algorithm that automatically corrects labels based on salient regions. Makantasis et al. [35], [36] proposed a tensor-based learning model to deal with noise scenarios; based on this, Tzortzis et al. [37] efficiently applied the model to detect abnormalities in digital mammograms. Hang et al. [38], [39] proposed a multitask generative adversarial network (MTGAN) to get more comprehensive training data by taking advantage of the rich information from unlabeled samples, thus indirectly improving the generalization ability of the classification task. Tanaka et al. [40] corrected labels during training by iteratively updating network parameters and labels. Zhu et al. [41] adjusted network parameters using well-defined labels and controlled overfitting during training on anomalous data using an overfitting control module. However, these approaches often address the problem through data augmentation, incorporating prior knowledge to artificially model noisy data and integrate it into the training process. This approach tends to increase training complexity and faces challenges in efficiently and flexibly handling remote sensing land cover classification, especially when dealing with diverse noise distributions. This can ultimately lead to reduced efficiency.

To address these challenges, we rethought the task of classifying land cover in high-noise scenarios. Consequently, we developed an innovative network architecture tailored to this objective. Specifically, we observe that noise in image spectral information affects local features and global features to different extents. For example, local features are stable in the face of color inaccuracies, while global features are more stable in the face of random noise. Therefore, when different kinds of anomalies appear in the image, the global and local features do not change to the same extent at the same time. Therefore, when different kinds of anomalies occur in the image, the global and local features will not change to the same extent simultaneously. Traditional deep neural networks typically extract both global and local features simultaneously. This makes independent optimization of global features difficult and reduces the efficiency of land cover type classification. To address this problem, we separate global features from local features. This allows local features to focus on classifying individual pixels, while prioritizing the optimization of global features that support semantic classification and handle different noise distributions, ultimately achieving a broader effect. Based on this rationale, we divide the land cover classification task into two subtasks: 1) a consistency assessment based on global features and 2) a semantic segmentation based on local features. We have developed a new model tailored for land cover classification in high-noise scenarios. This model facilitates the separate extraction of global and local feature

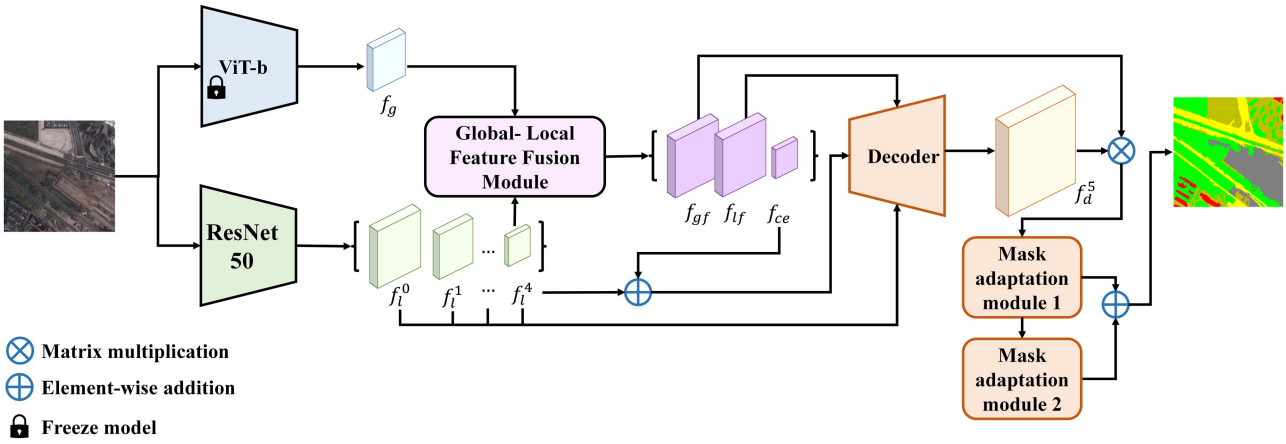


Fig. 2. Model architecture consists of two independent encoders, a global–local feature fusion module, and a decoder.

information, allowing the model to focus on the extraction of valuable information. As a result, it increases the flexibility and efficiency of the model, which ultimately strengthens its robustness.

In summary, our main contributions can be summarized as follows.

1) We reevaluated land cover classification tasks in high-noise images and introduced a new network architecture. In this design, we separated the extraction of global features from the extraction of local features, thereby improving the robustness of land cover classification tasks in high-noise remote sensing images.
2) To achieve effective feature embedding, we introduced an innovative global–local feature fusion module (GLFM). This module combines global and local features in three different ways, emphasizing local features, emphasizing global features, and emphasizing global feature similarity, thus enhancing the robustness of the model in handling high-noise scenarios.
3) Using the publicly available Wuhan high-definition landscape dataset (WHDLD) [42], which contains obviously anomalous data, we compared our proposed approach with several classical deep neural network models in terms of classification accuracy and computational complexity. The experimental results show that our method exhibits the best performance compared to the others.

The rest of this article is organized as follows. In Section II, we introduce the model architecture. Following that, Section III provides detailed information about the experiments, encompassing experimental details, metrics, and the datasets used. Moving on to Section IV, we present the specific results of the experiments. Subsequently, in Section V, we delve into the analysis of the influence of different fusion methods in the GLFM, the influence of anomalous data on deep learning models, and other discussions that demonstrate the validity of our model. Finally, Section VI concludes this article.

## II. Methodology

### A. Model Architecture

As shown in Fig. 2, our model architecture consists of two independent encoders, a GLFM, and a decoder. The two independent encoders, one focusing on the extraction of global features and the other on the extraction of local features, have different roles: global features are more concerned with data consistency evaluation, while local features focus on land-cover type classification. In the GLFM, global features and local features are combined in three different ways, with different emphasis on global and local information. In this process, the differences in global features are integrated into local features to produce comprehensive features that retain local information and are more discriminative. Finally, the decoder continuously combines the more discriminative deep features with shallow features until the original image size is restored, resulting in the final land-cover type classification results.

### B. Global and Local Feature Extraction Networks

Since we need to independently extract both global and local information from remote sensing images, in the model design, we utilize two separate feature extraction networks for the extraction of global and local features.

For local feature extraction, CNNs move convolutional kernels across the image, capturing information within the local receptive field at each location. This means that CNNs can focus on small regions of the image, allowing for better capture of local features and subtle textures. ResNet [43], a classic CNN, is an example of such a network that uses residual connections to solve the vanishing gradient problem and prevent model degradation, thus achieving excellent performance in local feature extraction. The computational process of ResNet is as follows:

$$f_l^i = \emptyset_{\text{res}}\ \left(x^i\right) = F\left(x^i, W\right) + x^i \tag{1}$$

$$F\ \left(x^i, W\right) = BN_2\left(\text{conv}_2(\text{ReLU}(BN_1\left(\text{conv}_1\left(x^i\right)\right)))\right). \tag{2}$$

Here, $\emptyset_{\text{res}}(.)$ is a residual block, $F(.)$ is the residual function, $x^i \in R^{(C,H,W)}$ is the input feature of the $i$th residual block, $W \in R^{(N,C,H,W)}$ is the weight of the residual block, $BN(.)$ is batch normalization, conv(.) is convolution, ReLU(.) is the activation function, and $f_{\text{res}}$ is the output feature of a residual block.

On the other hand, ViT [44] is a model based on an attention mechanism. It divides the input image into patches and uses self-attention mechanisms to capture relationships between these patches. Since the ViT model captures relationships between different positions in the image on a global scale by considering all positions simultaneously, it has a stronger ability to extract global features. We use ViT as a global feature extraction network. The computational process of ViT is as follows:

$$f_{\text{trans}} = \left[x_p^0 E_{\text{trans}}; x_p^1 E_{\text{trans}}; \ldots; x_p^M E_{\text{trans}}\right] \tag{3}$$

$$f_g = LN\left(\text{flatten}\left(f_{\text{trans}} + E_{\text{pos}} E_{\text{trans}}\right)\right). \tag{4}$$

In this equation, $E_{\text{trans}} \in R^{((P^2 * C)\times 512)}$ is the weights of the vision transformer embedding module, $P$ is the patch size, $E_{\text{pos}} \in R^{((M+1)\times 512)}$ is the positional embedding, $M$ is the number of patches, flatten(.) is the flattening operation, $LN(.)$ is a linear layer, and $f_g$ is the output feature of a ViT block.

Furthermore, since it is necessary to use global features for data consistency assessment, we used pretrained ViT weights trained on the WIT [45] dataset and freeze this part of the model to obtain a more stable prior for global feature judgments.

### C. Global–Local Feature Fusion Module

After extracting global and local features using the respective global and local feature extraction networks, we need to fuse these features, considering that we use two independent feature extraction networks. In the GLFM, we design three different structures: GLFM1, which generates global–local fusion features that emphasize global information; GLFM2, which generates global–local fusion features that emphasize local information; and GLFM3, which generates consistency evaluation scores. The output of GLFM1 primarily provides detailed information on land use classification results. The output of GLFM2 is incorporated into the features output by the decoder through dot multiplication, providing the main category information for land-use classification results. The consistency evaluation score comprehensively assesses differences in ViT features among different samples, representing variations in features that focus on global information between different samples.

In GLFM1, we employed the cross-attention mechanism, using global features as queries and the deepest local information as key/value to combine features. Finally, the obtained result is input to a feedforward neural network (FFN) for adjustment, yielding the global–local fusion feature $f_{gf} \in R^{\frac{HW}{256}\times D}$. This process is expressed using the following equation:

$$f_{gf}^0 = \text{FFN}\left(\gamma_g \text{Attention}\left(\text{norm}\left(f_g\right), \text{norm}\left(f_l^4\right)\right) + f_g\right) \tag{5}$$

$$f_{gf} = f_{gf}^0 + f_g. \tag{6}$$

Here, $f_g \in R^{\frac{HW}{256}\times D}$ is the global feature; $f_l^4 \in R^{\frac{HW}{1024}\times D}$ is the local feature; $\gamma_g \in R^D$, where $\gamma_g$ is learnable, is used to balance the output of the attention layer and the input features, Attention(.) is the attention layer, and norm(.) is the LayerNorm.

In GLFM2, we also utilized the cross-attention mechanism. However, in this case, the global features serve as key/values, while the deepest local information serves as the query for feature fusion. This process yields the global–local fusion feature $f_{lf} \in R^{\frac{HW}{1024}\times D}$, emphasizing the preservation of local information. This process is expressed using the following equation:

$$f_{lf} = \gamma_l \text{ Attention}\left(\text{norm}\left(f_l^4\right), \text{norm}\left(f_g\right)\right) + f_l^4. \tag{7}$$

Here, $\gamma_l \in R^D$ is learnable, and is also used to balance the output of the attention layer and the input features.

In GLFM3, we used a multilayer perceptron (MLP) to project the global features into a parameter, yielding a consistency evaluation score $f_{\text{ce}}$ for each image. This process is expressed using the following equation:

$$f_{\text{ce}} = \text{MLP}\left(f_g\right). \tag{8}$$

Here, $f_{\text{ce}} \in R^1$, and MLP(.) is a multilayer perceptron.

These three generated features are fed into the fusion feature decoder for further decoding to obtain the final results. Fig. 3 shows the detailed structure of the GLFM.

### D. Fusion Feature Decoder and Mask Adaptation Module (FDMAM)

In order to better preserve low-level features, we refer to the U-net network structure and use skip connections to directly connect the corresponding levels of the encoder and decoder, the decoder obtains more detailed low-level information through these connections, improving the segmentation model's ability to perceive details. The fusion feature decoder consists of four identically structured decoder blocks, gradually restoring the fusion features into per-pixel classification labels. This process can be represented using the following equation:

$$f_d^0 = \text{FFN}\left(\text{Concat}\left(\left(f_l^4 + f_{ce}\right), f_{lf}\right)\right) \tag{9}$$

$$f_d^1 = DB_1\left(f_d^0, f_l^3\right) \tag{10}$$

$$f_d^2 = DB_2\left(f_d^1, f_l^2\right) \tag{11}$$

$$f_d^3 = DB_3\left(f_d^2, f_l^1\right) \tag{12}$$

$$f_d^4 = DB_4\left(f_d^3, f_l^0\right). \tag{13}$$

Here, Concat(.) is the concatenation operation, $DB_i(.)$ is the $i$th decoder block. Fig. 4 shows the detailed structure of the FDMAM.

### E. Mask Adaptation Module

To generate masks, we perform matrix multiplication between the globally focused fusion feature $f_{\text{gf}}$ and the decoder output feature $f_d^4$ for each category, using the output feature $f_d^4$ from the fusion feature decoder. To obtain the final output, we need a mask adaptation module to adjust the decoder output feature

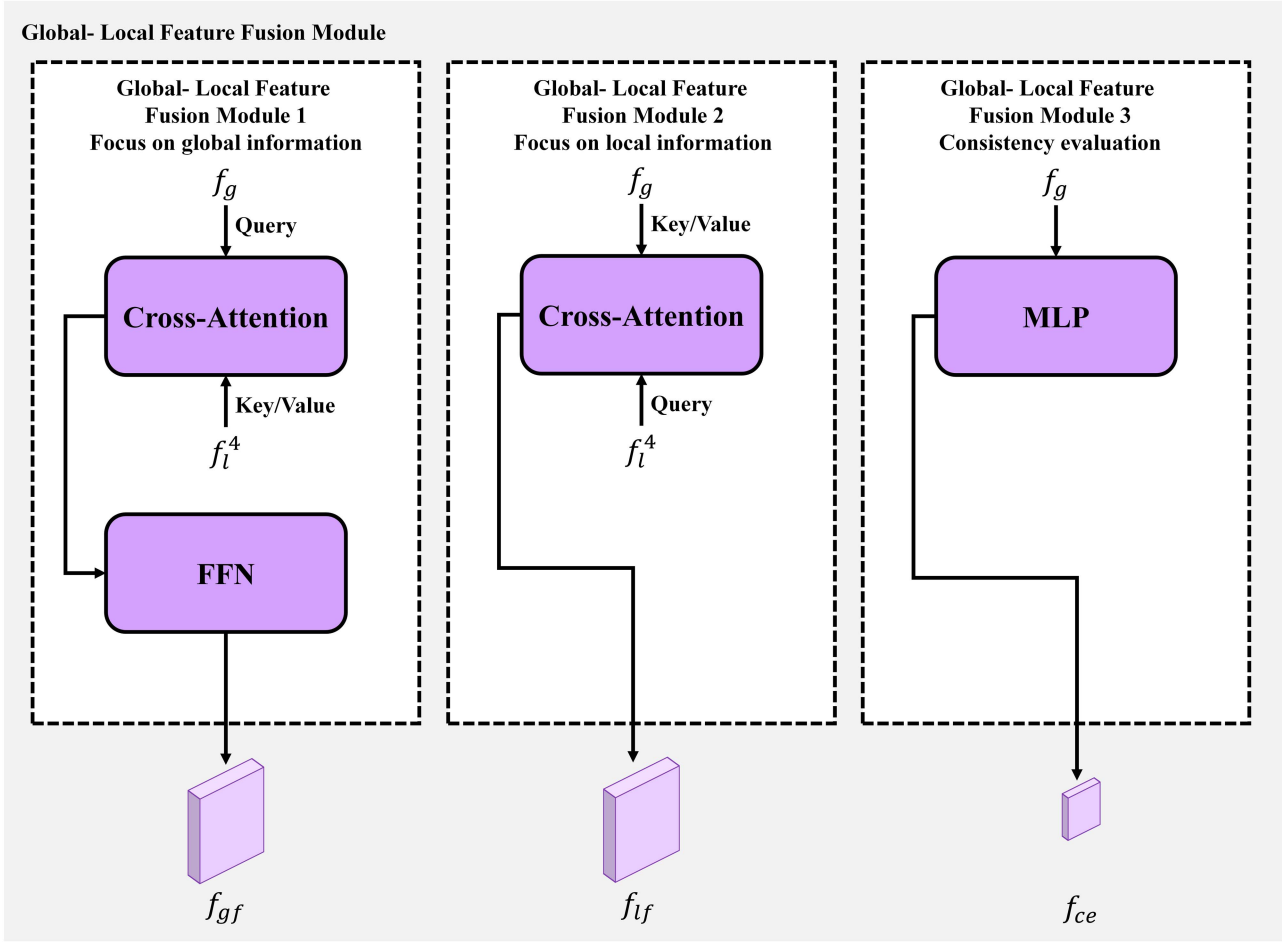


Fig. 3. Structure of the global–local feature fusion module.

by emphasizing the global–local feature mask generated by the combination of the fusion feature and the decoder output feature. Both mask adaptation module 1 and mask adaptation module 2 are independent MLPs. This process is expressed by the following equations:

$$f_d^5 = \mathrm{MLP}_1\ \left(f_d^4 \times f_{\mathrm{gf}}\right) \tag{14}$$

$$\hat{y}brk = \mathrm{MLP}_2\ \left(f_d^5\right) + f_d^5. \tag{15}$$

Here, $\mathrm{MLP}_1(.)$ is the mask adaptation module 1, and $\mathrm{MLP}_2(.)$ is the mask adaptation module 2.

## III. Experiment

### A. Experimental Details

Our method is implemented based on the PyTorch framework. According to the network training requirements, the network is trained for 50 epochs with a learning rate set to 1e-4. We use the Adam optimizer for training, with a batch size of 8 and weight decay set to 0. In addition, we do not perform any data augmentation. For the loss function, we use the cross-entropy loss, and the calculation formula is as follows:

$$\mathrm{Loss}_{\mathrm{CE}}\left(y, \hat{y}\right) = -\sum_{i=1}^{N} y_i \log\left(\hat{y}_i\right). \tag{16}$$

Here, $y$ is the true label, $\hat{y}$ is the model's output, and $N$ is the number of classes.

### B. Experimental Metrics

To evaluate the accuracy and efficiency of the image segmentation model, we used three commonly used metrics: 1) mean pixel accuracy (Mpa), mean intersection over union (MIoU), and floating-point operations(FLOPs). Mpa measures the accuracy of the model's predictions for each pixel and averages the accuracy over all pixels. The calculation formula is as follows:

$$\mathrm{MPA} = \frac{1}{k} \sum_{i=1}^{k} \frac{p_{ii}}{\sum_{j=1}^{k} p_{ij}}. \tag{17}$$

Here, $k$ is the number of classes, $p_{ii}$ is the number of pixels belonging to class $i$ and predicted as class $i$, and $p_{ij}$ is the number of pixels belonging to class $i$, but predicted as class $j$.

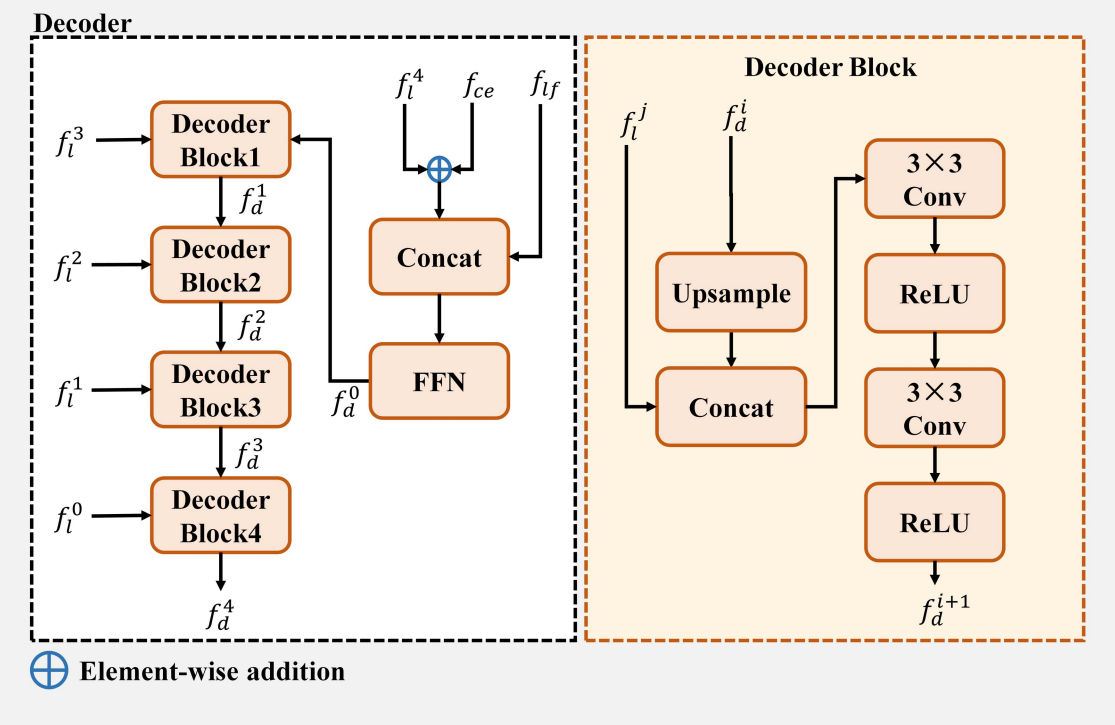


Fig. 4. Structure of the fusion feature decoder.

MIoU is another metric commonly used to evaluate the performance of land cover classification methods. It averages the intersection over union for each category to measure prediction accuracy. The formula is as follows:

$$\text{MIoU} = \frac{1}{k}\sum_{i=1}^{k}\frac{p_{ii}}{\sum_{j=1}^{k}p_{ij}+\sum_{j=1}^{k}p_{ji}-p_{ii}}. \quad (18)$$

Here, $p_{ji}$ is the number of pixels belonging to class $j$, but predicted as class $i$.

FLOPs, which can be understood as the amount of floating-point calculations required to completely execute a model, can be used to measure the complexity of an algorithm or model.

### C. Datasets

1) The WHDLD is a comprehensive dataset extracted and clipped from high-resolution remote sensing images covering the urban area of Wuhan. The dataset includes six classes: 1) buildings, 2) roads, 3) sidewalks, 4) vegetation, 5) bare soil, and 6) water bodies. In total, WHDLD consists of 4940 color images, each 256 × 256 pixels, with a spatial resolution of 2 m. Some images in the dataset contain noticeable color distortions indicating the presence of anomalous data. To evaluate the performance of our method in the presence of anomalous data, we divided the dataset into training, validation, and test sets in a 60:20:20 ratio, ensuring a balanced distribution of categories and anomalous data in each set. For a clearer illustration, Fig. 5 shows examples of training data from the WHDLD dataset, demonstrating the visual diversity, complexity, and inclusion of anomalous data within the dataset.

In addition, to validate that our method remains stable in the presence of different types of noise and to further assess the robustness of our method to anomalous data, we removed the preexisting anomalous data from the WHDLD test dataset that we had previously partitioned. We then introduced Gaussian noise with a mean of 0 and a standard deviation of 0.02 to the remaining data, simulating anomalous data. Fig. 6 shows a comparison before and after adding noise to some of the data.

2) The land cover from aerial imagery (LandCover.ai) dataset is used to automatically draw land object types from aerial images. The dataset includes four categories of features: 1) buildings, 2) roads, 3) woodland, and 4) water, plus a background category. LandCover.ai consists of a total of 10 674 color images, each image is 512 × 512 pixels and has a spatial resolution of 0.25 m. We further downsample the image size to 256 × 256. Same as WHDLD, we divide the dataset into training set, validation set, and test set in the ratio of 60:20:20. Fig. 7 shows an example of training data from the LandCover.ai dataset.

In addition, like WHDLD, we also added Gaussian noise with a mean of 0 and a standard deviation of 0.02 to the LandCover.ai test dataset to simulate anomalous data.

## IV. Results

Using the original WHDLD dataset, we conducted experiments with several mainstream and state-of-the-art models using the same experimental setup described in Section III-A. These

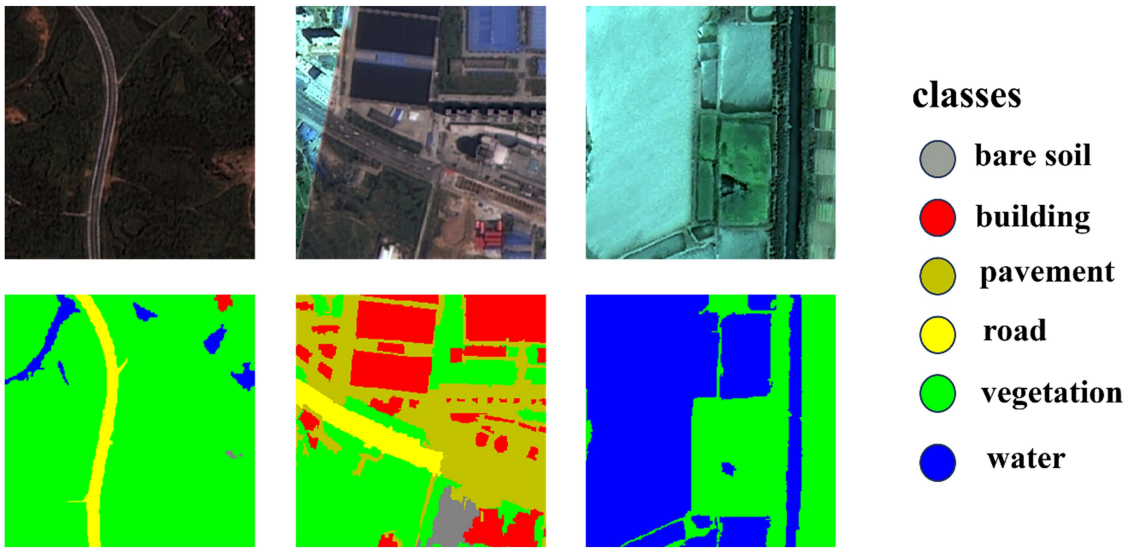


Fig. 5. Examples of training data from the WHDLD dataset, including some anomalous data.

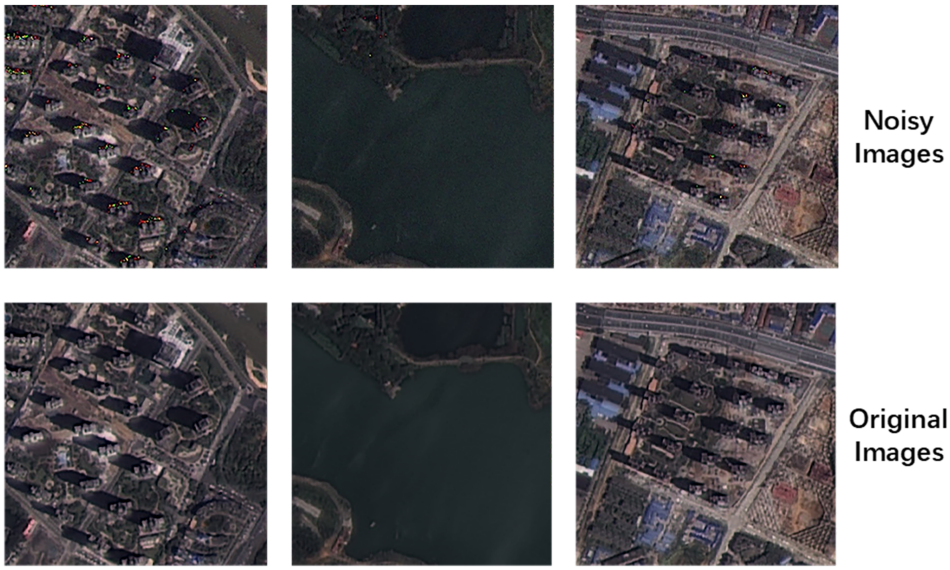


Fig. 6. Comparison of training dataset before and after adding noise.

models include DEEPLabV3+ [46], UNet [47], UNet++ [48], PSPNet [49], and DABNet [50] based on convolutional structures, as well as attention-based models, such as Segformer [51], CCNet [52], and AttUNet [53] that fuse attention mechanisms with convolutions. In addition, Fig. 8 is the loss curve of our model during the training process on the WHDLD dataset, which can be observed. In the last 5 epochs of training, the loss has been maintained near 0.2, proving that our model has converged within 50 epochs.

Table I presents the results of these models on the test set, focusing on three metrics: 1) Mpa, 2) MIoU, and 3) FLOPs. In assessing model accuracy in land cover classification, our approach achieves the highest scores in the Mpa and MIoU metrics, with scores of 74.32% and 63.58%, respectively. The

TABLE I
COMPARING THE ACCURACY OF EXPLORED MODELS FOR LAND COVER CLASSIFICATION

| Method | Mpa(%) | Miou(%) | FLOPs(G) |
|---|---|---|---|
| PSPNet | 46.19 | 46.19 | **0.56** |
| CCNet | 65.53 | 54.95 | 42.13 |
| DABNet | 66.08 | 57.27 | 1.02 |
| DEEPLabV3+ | 59.48 | 52.13 | 15.98 |
| UNet | 70.51 | 60.26 | 41.97 |
| AttUNet | 72.75 | 60.81 | 103.68 |
| UNet++ | 71.27 | 62.29 | 153.79 |
| Segformer | 65.06 | 53.09 | 13.71 |
| **Ours*** | **74.32** | **63.58** | **32.17** |

The significance of bold entities is the optimal performance indicator.

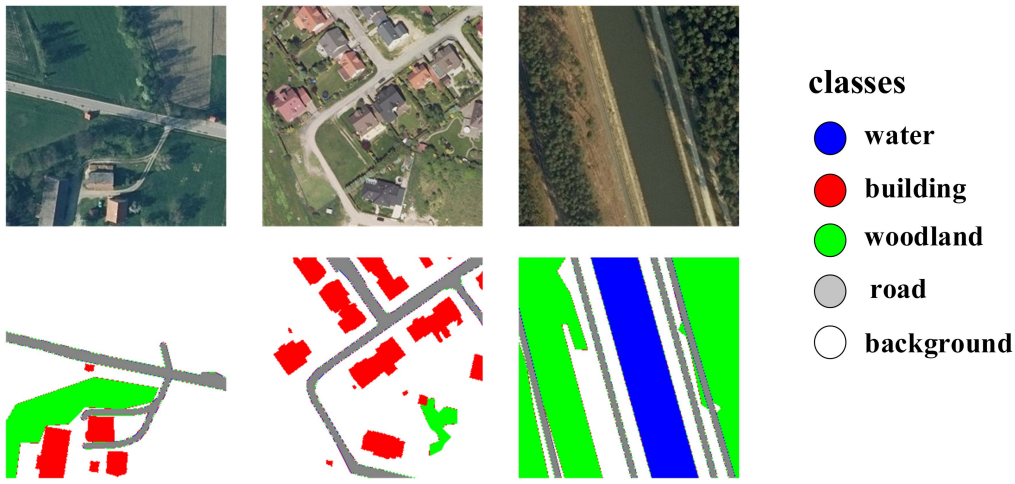


Fig. 7. Illustrates examples of training data from the LandCover.ai dataset.

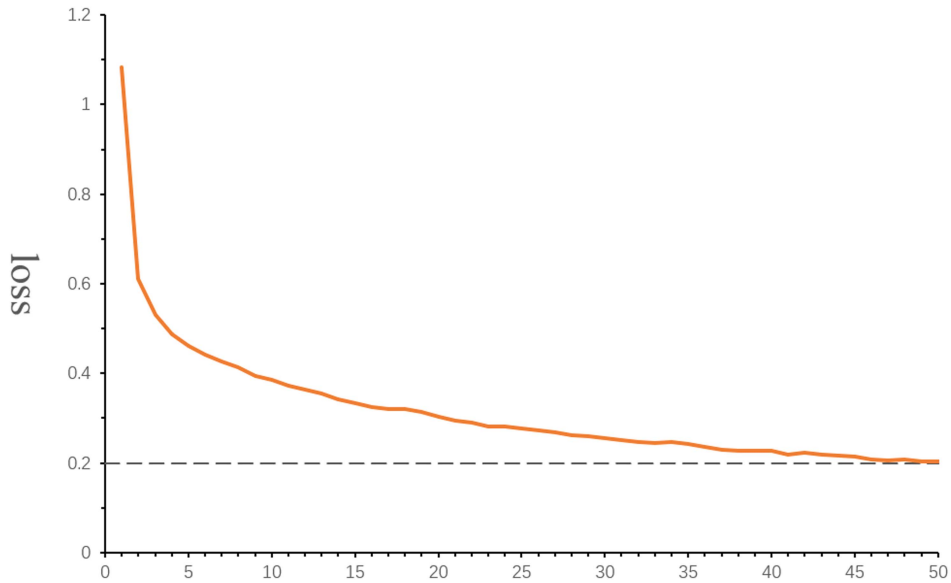


Fig. 8. Loss curve trained on WHDLD data set, can see that it has converged within 50 epochs.

relatively suboptimal UNet++ improves by 3.05% and 1.29% in these metrics, respectively. Regarding the computational cost measured by the FLOPs metric, our method incurs significantly lower computational cost compared to UNet++ and AttUNet, and slightly less than UNet and CCNet.

Furthermore, as shown in Fig. 9, we present some representative visualization results of different networks on the test sets, including the anomalous data and complex scenes present in the dataset. From the results, it is clear that our model excels in distinguishing land cover categories. In particular, for road type identification, our approach achieves a better balance by preserving more coherent road shapes while improving classification accuracy.

To validate the robustness of our method to different types of anomalous data, we tested the same model on the test set with added Gaussian noise. Our method achieved optimal results, and the performance degradation due to noise was minimal. Table II presents our results on the WHDLD dataset after the addition of Gaussian noise, showing the performance comparison.

TABLE II
PERFORMANCE COMPARISON ON THE WHDLD DATASET AFTER ADDING GAUSSIAN NOISE

| Method | Mpa(%) | Miou(%) |
|---|---|---|
| PSPNet | 41.65 | 29.77 |
| CCNet | 58.94 | 46.75 |
| DABNet | 55.00 | 42.35 |
| DEEPLabV3+ | 51.05 | 39.06 |
| UNet | 64.26 | 50.23 |
| AttUNet | 66.56 | 55.38 |
| UNet++ | 66.06 | 55.07 |
| Segformer | 48.42 | 34.54 |
| **Ours*** | **70.51** | **59.05** |

The significance of bold entities is the optimal performance indicator.

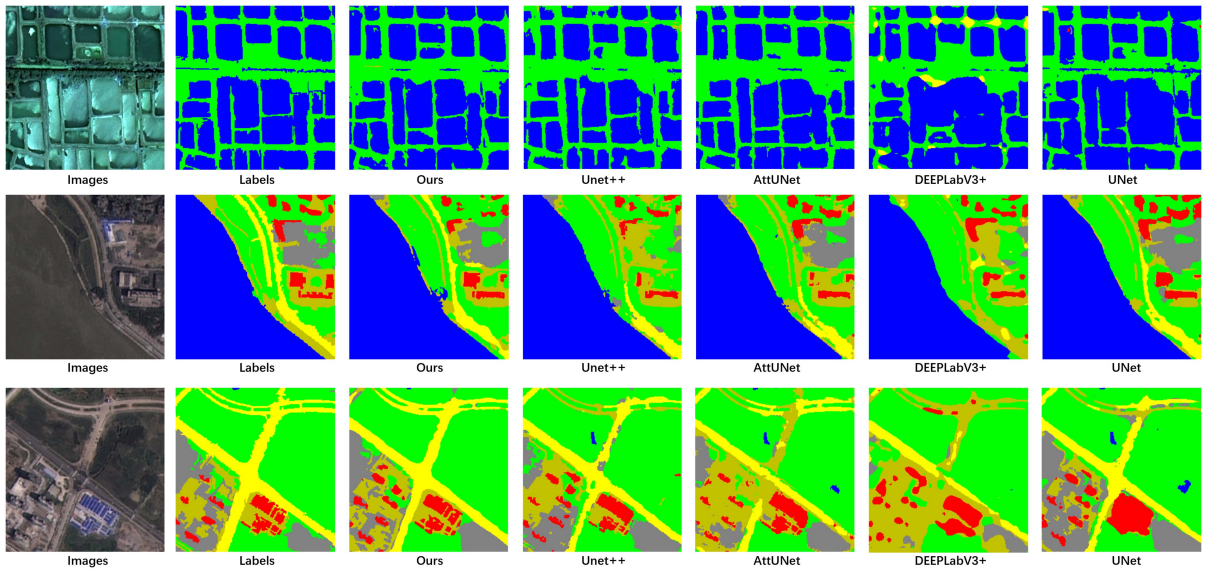


Fig. 9. Results of different models' performances using the original WHDLD dataset.

TABLE III
PERFORMANCE COMPARISON ON THE ORIGINAL LANDCOVER.AI DATASET

| Method | Mpa(%) | Miou(%) |
|---|---|---|
| PSPNet | 55.82 | 49.71 |
| CCNet | 67.57 | 61.22 |
| DABNet | 69.29 | 61.98 |
| DEEPLabV3+ | 61.36 | 53.85 |
| UNet | 72.82 | 66.30 |
| AttUNet | 76.67 | 69.87 |
| UNet++ | 76.72 | 69.79 |
| Segformer | 69.57 | 61.16 |
| **Ours*** | **78.41** | **71.31** |

The significance of bold entities is the optimal performance indicator.

TABLE IV
PERFORMANCE COMPARISON ON THE LANDCOVER.AI DATASET AFTER ADDING GAUSSIAN NOISE

| Method | Mpa(%) | Miou(%) |
|---|---|---|
| PSPNet | 49.13 | 42.79 |
| CCNet | 63.15 | 57.04 |
| DABNet | 63.20 | 56.38 |
| DEEPLabV3+ | 56.32 | 48.57 |
| UNet | 68.84 | 62.36 |
| AttUNet | 70.78 | 64.51 |
| UNet++ | 70.59 | 64.36 |
| Segformer | 64.88 | 56.99 |
| **Ours*** | **75.89** | **69.54** |

The significance of bold entities is the optimal performance indicator.

The performance of the models on noisy and noise-free data is shown in Fig. 10, which shows that our model can still accurately determine land-cover categories on noisy data, demonstrating superior robustness.

In addition, we extended our evaluation to LandCover.ai dataset. Following the same experimental setup outlined in Section III-A, we tested our model along with the previously mentioned models on this new dataset.

Table III outlines the performance comparison on the LandCover.ai test set, utilizing the same metrics (Mpa and MIoU). Our model achieves competitive scores in terms of Mpa (78.41%) and MIoU (71.31%). In addition, our method maintained its superiority over UNet++ and AttUNet, showcasing its effectiveness across diverse datasets.

Furthermore, as shown in Fig. 11, we also show some representative visualization results of different networks on the LandCover.ai test set. It can be observed that our results surpass other models in terms of classification accuracy.

To assess the model's adaptability to varying conditions, we also added Gaussian noise to the LandCover.ai test set. Surprisingly, our model exhibited robustness similar to its performance on the WHDLD dataset, with minimal degradation in accuracy due to the added noise. The results are summarized in Table IV.

Fig. 12 illustrates the comparative performance of different models on both noisy and no noise data in the LandCover.ai dataset.

In conclusion, our model not only performs exceptionally well on the WHDLD dataset but also showcases its robustness on the LandCover.ai dataset, affirming its potential for broad applicability in diverse scenarios.

## V. DISCUSSION

### A. *Influence of Different Fusion Methods in the Global–Local Feature Fusion Module*

To investigate the influence of different fusion methods in the GLFM and to further validate the effectiveness of our approach, we performed ablation experiments specifically targeting the GLFM. We sequentially removed the submodules corresponding to different fusion methods within this module and compared them to the backbone where the entire GLFM was removed.

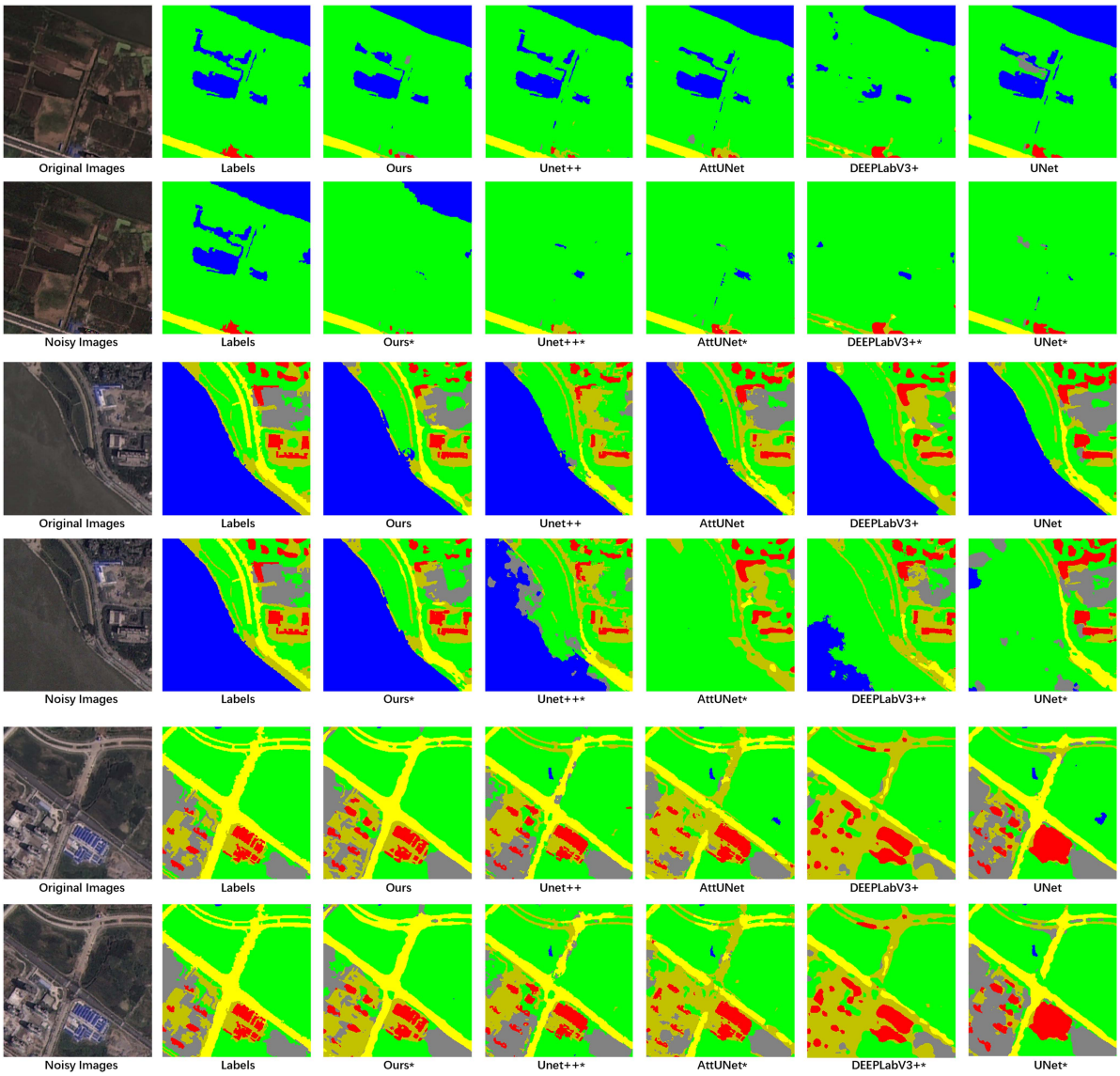


Fig. 10. Results of different models' performances using the Gaussian noise-added WHDLD dataset.

The experimental results are presented in Table V. The results show that GLFM1 and GLFM2 alone are helpful for improving MIoU, but have a negative impact on Mpa; GLFM3 is helpful for improving Mpa, and GLFM3 has no obvious negative impact on MIoU. The two combinations of GLFM1+ GLFM3 and GLFM2+ GLFM3 integrate their respective advantages and significantly improve Mpa and MIoU. The combination of GLFM1+ GLFM2 further improves MIoU and reduces Mpa. Finally, the combination of GLFM1+ GLFM2+ GLFM3 obtained the highest Mpa and MIoU scores. In summary, we believe that GLFM1 and GLFM2 may pay more attention to geometric information, such as boundaries, while GLFM3 improves the classification performance of the model. Therefore, the three different fusion methods complement each other to maximize the improvement of the comprehensive performance of the model.

### B. *Influence of Anomalous Data on Deep Learning Models*

While our training set does contain a small amount of anomalous data, the proportion is extremely small. Through our manual evaluation, we found that in the entire WHDLD dataset, there are only 105 images with obvious hue anomalies, representing only 2.13% of the data. The anomalous data in WHDLD are shown in the middle and right side of Fig. 5. We want to verify whether such a small amount of anomalous data would profoundly affect the results of model training. Therefore, we exclude the anomalous data from the WHDLD dataset, divide it into training and test sets, and train models based on the dataset without anomalous data. We then compare these models with those trained using the same model architecture but with noisy data to assess the influence of anomalous data on model training. We test four models, including our model and the suboptimal model UNet++ (Table VI).

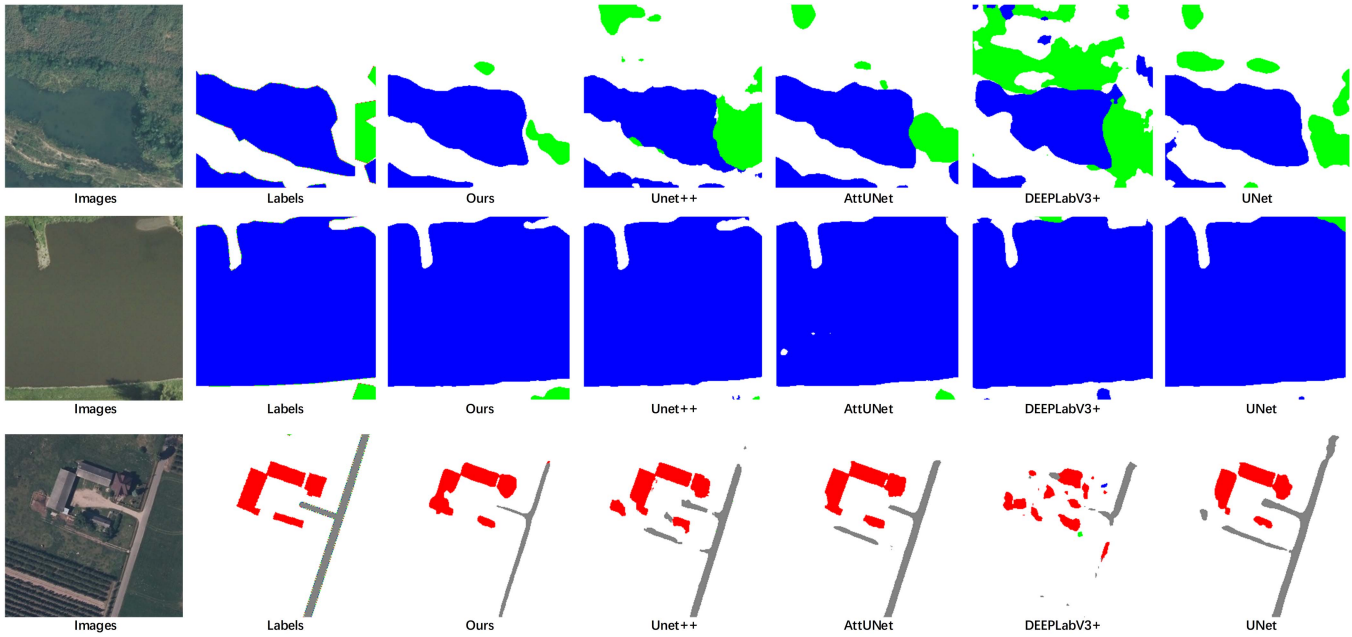


Fig. 11. Comparison of partial visualization results on the original LandCover.ai dataset.

TABLE V
COMPARE THE EFFECTS OF ADDING DIFFERENT MODULES

| Method | Mpa(%) | Miou(%) |
|---|---|---|
| Backbone | 72.74 | 61.68 |
| Backbone+GLFM1 | 72.71 | 61.87 |
| Backbone+GLFM2 | 72.54 | 62.85 |
| Backbone+GLFM3 | 73.45 | 61.74 |
| Backbone+ GLFM1+ GLFM3 | 73.71 | 61.86 |
| Backbone+ GLFM1+ GLFM2 | 72.32 | 63.36 |
| Backbone+ GLFM2+ GLFM3 | 73.59 | 63.01 |
| Backbone+ GLFM1+ GLFM2+GLFM3(Ours) | **74.32** | **63.56** |

"Backbone" refers to the model that removes the global–local feature fusion module.
The significance of bold entities is the optimal performance indicator.

TABLE VI
EFFECTS OF USING AN ANOMALOUS TRAINING SET ON THE MODELS

| Method | Abnormal training set | Mpa(%) | Miou(%) |
|---|---|---|---|
| UNet | Yes | 64.26 | 50.23 |
| UNet | No | 71.09 | 61.44 |
| DEEPLabV3+ | Yes | 59.48 | 52.13 |
| DEEPLabV3+ | No | 65.97 | 56.67 |
| UNet++ | Yes | 71.27 | 62.29 |
| UNet++ | No | 74.14 | 63.64 |
| Ours | Yes | 74.32 | 63.56 |
| Ours | No | 74.57 | 63.85 |

Our results show that the training of UNet, DEEPLabV3+, and UNet++ is significantly affected by the dataset containing anomalous data. Compared to models trained without anomalous data, the UNet model trained with anomalous data shows a decrease of 6.83% in Mpa and a decrease of 11.21% in MIoU on the test set. The DEEPLabV3+ model trained on anomalous data shows a decrease of 6.49% in Mpa and a decrease of 4.54% in MIoU on the test set. The UNet++ model trained on anomalous data shows a decrease of 2.87% in Mpa and a decrease of 1.35% in MIoU on the test set. On the other hand, our model trained with anomalous data shows a small decrease of 0.29% in Mpa and a negligible decrease of 0.15% in MIoU on the test set, with a very small range of variation. Therefore, we can draw two conclusions: first, for traditional deep learning models, even a small amount of noise can have a significant impact on the training results, greatly reducing the effectiveness of land cover classification tasks. Second, compared to other models, our model is less sensitive to anomalous data, which shows greater robustness.

### C. *Verification of Global and Local Feature Change Patterns of Anomalous Data*

The basic assumption of this study is that when an abnormality occurs in an image, the degree of change of global and local features of anomalous images is different. This feature can be used to mitigate performance degradation caused by data anomalies

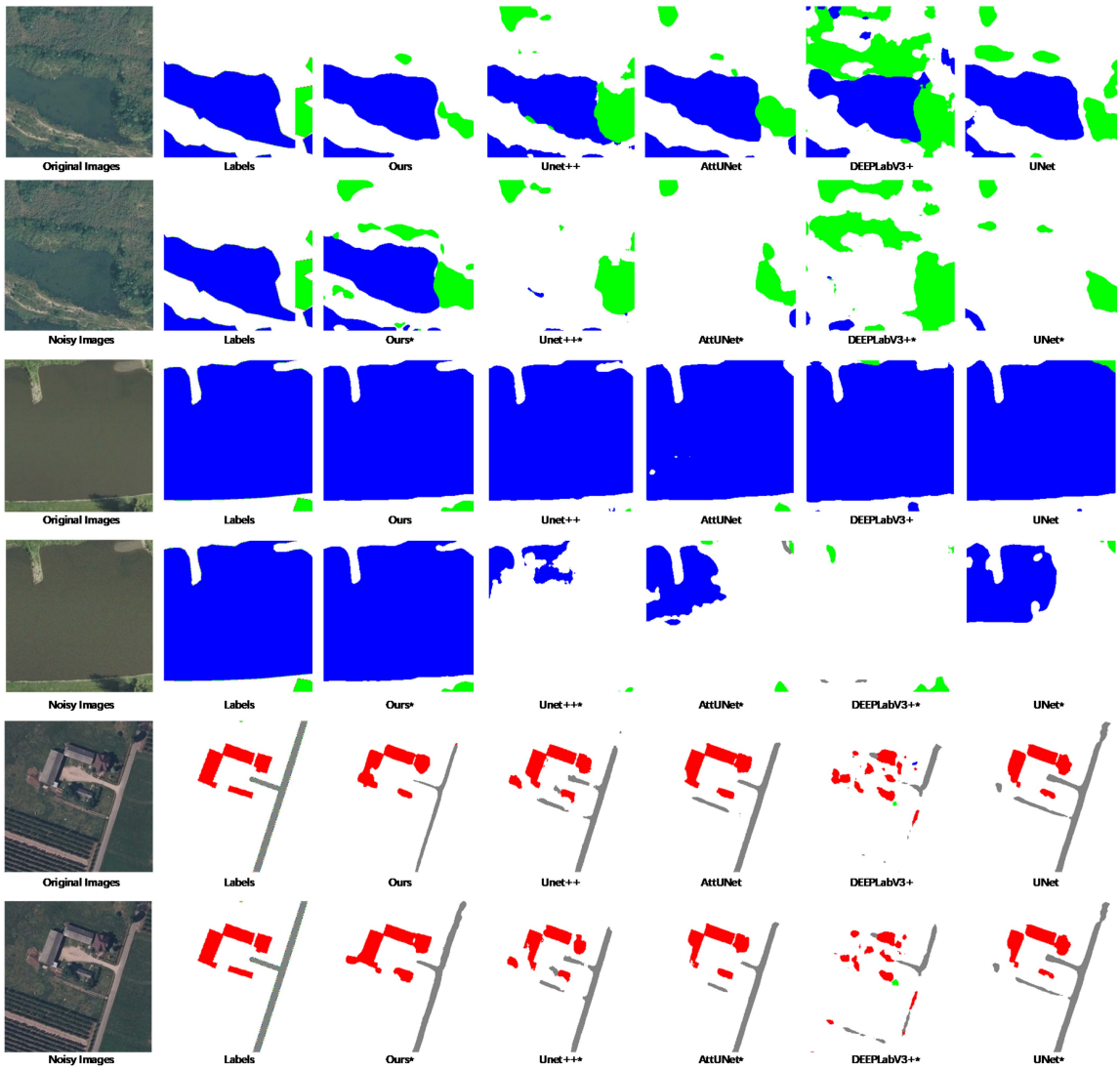


Fig. 12. Partial visualization results comparison on the LandCover.ai dataset after adding Gaussian noise.

within a certain range. In fact, our model does achieve excellent performance in the presence of anomalous data. In order to further verify whether the global and local features of anomalous data change to the same extent, we conducted the following experiments: Perform different types of transformations on the original image, and then evaluate the difference between the global features and local features of the transformed image and the original image.

As shown in Fig. 13, we performed two types of transformations on the original image: 1) modifying the intensity of different color channels (R channel attenuation 50%; G channel attenuation 50%; B channel attenuation 50%.), and 2) adding varying degrees of Gaussian noise to the image (0.1 mean, 0.1 standard deviation; 0.1 mean, 0.05 standard deviation; 0.1 mean, 0.02 standard deviation.).

We use MPEG-7's [54] color layout descriptor (CLD) and histogram of oriented gradients (HOG) to describe images. The former is more concerned with color information, and the latter is more concerned with edge information. Subsequently, we measured the differences between the transformed image and the original image using Euclidean distance. The results are shown in Tables VII and VIII. We observed that color transformation had a greater impact on CLD than HOG, while Gaussian noise had a greater impact on HOG than CLD.

TABLE VII
INFLUENCE OF DIFFERENT TYPES OF IMAGE TRANSFORMATIONS ON CLD

| Image | Transform type | Difference from original image (Euclidean distance) |
|---|---|---|
| (a) | None | 0.0 |
| (b) | Color | 1298.10 |
| (c) | Color | 902.22 |
| (d) | Color | 1065.69 |
| (e) | Gaussian noise | 993.14 |
| (f) | Gaussian noise | 982.97 |
| (g) | Gaussian noise | 999.16 |

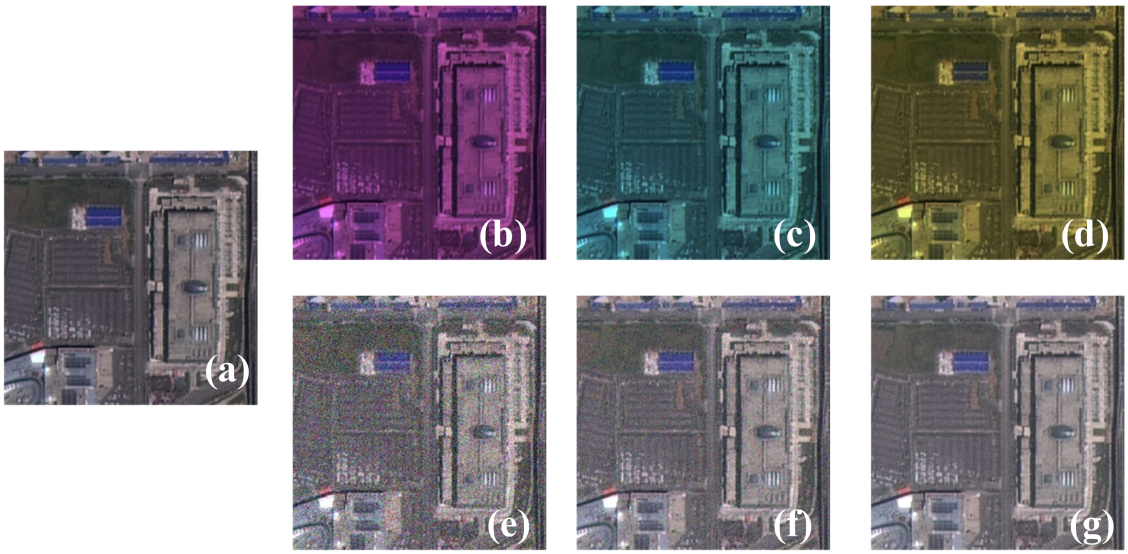


Fig. 13. Images that have been transformed by different types. (a) Original image. (b) G channel attenuation 50%. (c) R channel attenuation 50%. (d) B channel attenuation 50%. (e) Add Gaussian noise with 0.1 mean, 0.1 standard deviation. (f) Add Gaussian noise with 0.1 mean, 0.05 standard deviation. (g) Add Gaussian noise with 0.1 mean, 0.02 standard deviation.

TABLE VIII
INFLUENCE OF DIFFERENT TYPES OF IMAGE TRANSFORMATIONS ON HOG

| Image | Transform type | Difference from original image (Euclidean distance) |
|---|---|---|
| (a) | None | 0.0 |
| (b) | Color | 4.53 |
| (c) | Color | 3.79 |
| (d) | Color | 3.75 |
| (e) | Gaussian noise | 8.21 |
| (f) | Gaussian noise | 13.20 |
| (g) | Gaussian noise | 18.10 |

TABLE IX
INFLUENCE OF DIFFERENT TYPES OF IMAGE TRANSFORMATIONS ON SIFT

| image | Transform type | Number of matching points |
|---|---|---|
| (a) | None | 400 |
| (b) | Color | 367 |
| (c) | Color | 367 |
| (d) | Color | 385 |
| (e) | Gaussian noise | 172 |
| (f) | Gaussian noise | 270 |
| (g) | Gaussian noise | 318 |

Simultaneously, we conducted a comparison of scale-invariant feature transform (SIFT) local features for images with different transformations. The results are shown in Table IX. We performed brute-force matching of SIFT points between the transformed and original images, where a higher number of matches indicates fewer changes. We found that the Gaussian noise had a greater impact on SIFT local features than color transform.

In summary, different types of anomalous data have different effects on different types of features.

### *D. Advantages of Using CNN and ViT Feature Extractors Respectively*

In order to verify the rationality of our model structure design, we will analyze the CNN and ViT feature extractor separately. We assume that CNN pays more attention to local information and ViT can pay more attention to global information. Our hypothesis can be verified by using one of these feature extractors alone. Specifically, if we remove ViT, we only need to remove GLFM at the same time. Table V shows the effect of the backbone network. It can be found that the ViT feature extractor contributes greatly to network performance. ViT features provide more global information.

We will next verify the effectiveness of CNN and verify that the performance improvement of our model is not all due to ViT. However, if we want to remove the CNN, must make substantial modifications to the network structure, which will seriously interfere with the effectiveness of the analysis. Therefore, we use hand-crafted local feature SIFT to compare with CNN features to verify how the CNN feature extractor behaves. If our CNN feature extractor is effective and pays more attention to local information. Then the effect of the model using CNN features should be better than that of SIFT features and the focus points should be similar to SIFT local feature points.

First, we visualize the CNN features and SIFT features of some examples. Specifically, we convert CNN features into heat maps; map the positions and directions of SIFT feature points to the original image. The visualization results are shown in Fig. 14. It can be observed from the figure that the CNN features focus on some areas with obvious boundaries, proving that CNN can well capture the local features of the image. Similar to but not

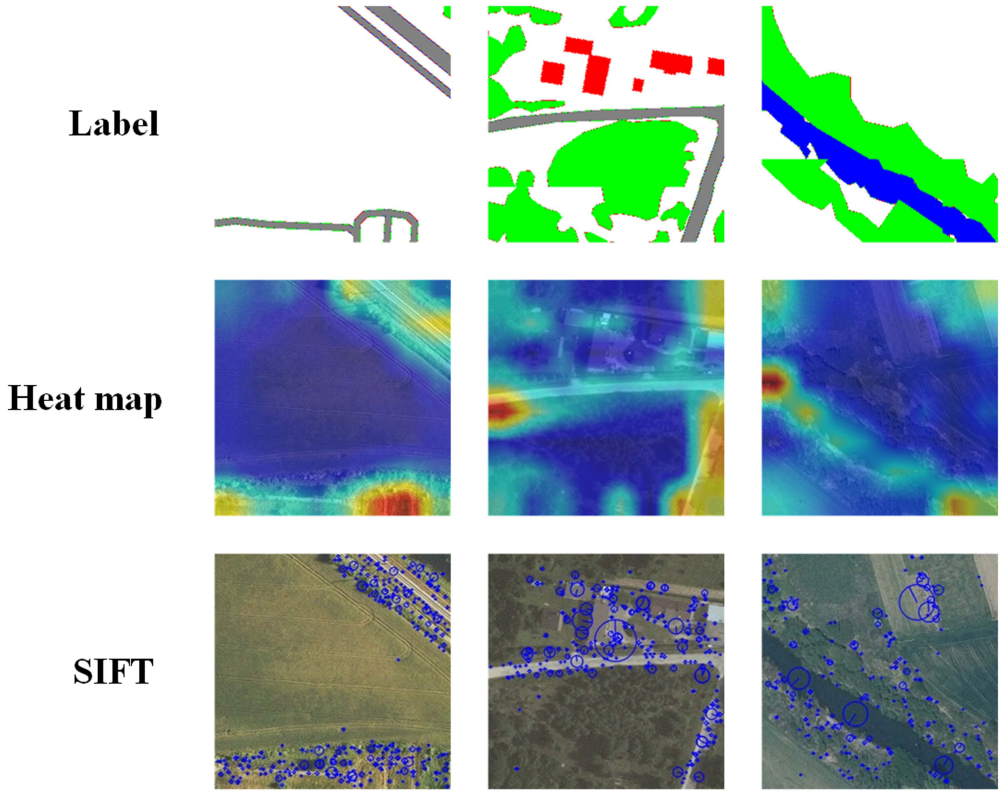


Fig. 14. Difference between heatmap of CNN features and SIFT features.

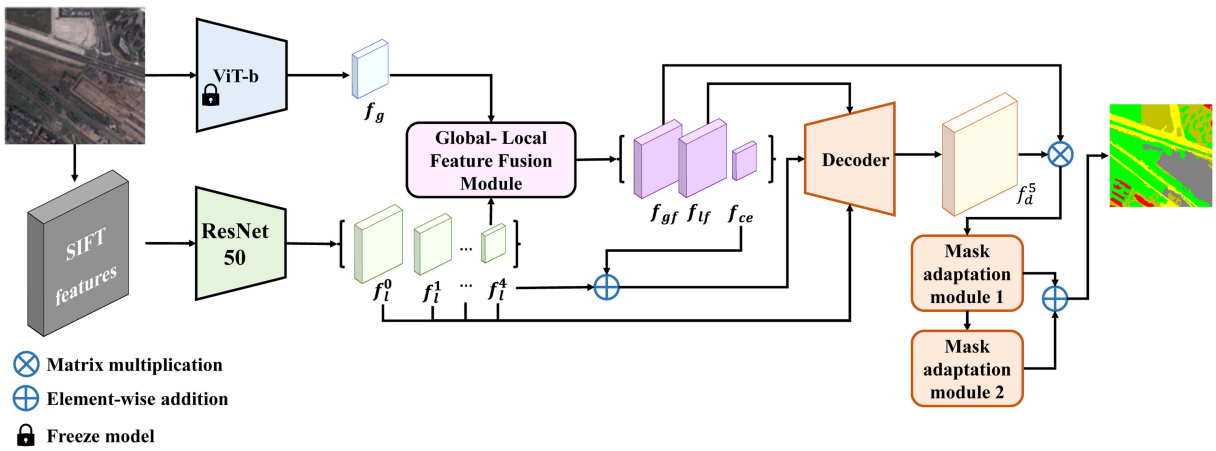


Fig. 15. Overall structure of the SIFT +Vit model.

identical to SIFT local features, the focus of CNN features is more closely related to the label.

Next, in order to quantitatively verify the difference between SIFT local features and CNN features, we modified the structure of the network. Extract SIFT features from the image and use them as input to the CNN. In addition, input the image separately into ViT to extract features, and then fuse the two features in the same way as the original model. The new model structure is shown in Fig. 15. On the LandCover.ai dataset, the performance of using SFIT features instead of CNN features dropped significantly compared to the original structure. Shown in Table X, on the LandCover.ai dataset, the SIFT+Vit structure achieved results of 36.93% for Mpa and 30.47% for MIoU. It is significantly lower than the results of Mpa78.41% and MIoU 71.31% obtained by the original structure.

TABLE X
PERFORMANCE OF MODEL USING SFIT FEATURES COMPARED TO ORIGINAL STRUCTURE ON LANDCOVER.AI DATASET

| Dataset | Features type | Mpa(%) | Miou(%) |
|---|---|---|---|
| LandCover.ai | SIFT | 36.93 | 30.47 |
| LandCover.ai | CNN features | 78.41 | 71.31 |

TABLE XI
INFLUENCE OF THE SIZE OF THE TRAINING SET

| Training set size | Mpa(%) | Miou(%) |
|---|---|---|
| 2964 (base) | 74.32 | 63.56 |
| 2223 | 72.72 | 61.22 |
| 1482 | 69.98 | 59.78 |
| 741 | 67.07 | 55.28 |

### E. Influence of the Size of the Training Set

To verify the *influence* of changes in training set size on model training, we reduced the training set size of the WHDLD dataset to 25%, 50%, and 75% of the original size, respectively, and trained and tested on the same test set. The experimental results are shown in Table XI.

The results show that as the amount of training data decreases, the model performance also decreases.

## VI. CONCLUSION

This study revisits the task of remote sensing land cover classification in high-noise scenarios and introduces a novel network architecture. We argue that in real remote sensing scenarios, common noise affects only certain parts of the data. For example, noise in spectral information typically does not significantly affect local features, such as textures and edges. However, noise resulting from common differences in spectral characteristics and variations in atmospheric states typically affects the overall global features. Therefore, to deal with typical color distortions and noise anomalies in remote sensing images, we distinguish between global and local features. This allows local features to focus on pixel-level classification, while global features focus on classification with different noise distributions, leading to a broader generalization effect. Based on this concept, we divided the land cover classification task into two subtasks: 1) a consistency assessment based on global features and 2) a semantic segmentation based on local features. We have developed a novel model that is capable of extracting global and local feature information independently. By employing a feature fusion module that emphasizes local, global, and global consistency, we optimize global and local features independently, thus reducing optimization complexity. This approach allows the model to focus on the extraction of valuable information, which ultimately improves the robustness of the model. The current work is still based on the referenced dataset, and data bias still exists when applying the model in more complex real-world scenarios. The impact of this data bias on model performance remains uncertain. Therefore, investigating the model's performance in high-noise real-world scenarios has become a compelling research direction. Our objective is to validate the model's performance in such scenarios and address the challenges posed by more complex noise. This will be the focus of our future research.

## REFERENCES


[1] C. Gómez, J. C. White, and M. A. Wulder, "Optical remotely sensed time series data for land cover classification: A review," *ISPRS J. Photogrammetry Remote Sens.*, vol. 116, pp. 55–72, 2016.

[2] K. Thyagharajan and T. Vignesh, "Soft computing techniques for land use and land cover monitoring with multispectral remote sensing images: A review," *Arch. Comput. Methods Eng.*, vol. 26, no. 2, pp. 275–301, 2019.

[3] Z. Ma, M. Xia, H. Lin, M. Qian, and Y. Zhang, "FENet: Feature enhancement network for land cover classification," *Int. J. Remote Sens.*, vol. 44, no. 5, pp. 1702–1725, 2023.

[4] J. Sheffield et al., "Satellite remote sensing for water resources management: Potential for supporting sustainable development in data-poor regions," *Water Resour. Res.*, vol. 54, no. 12, pp. 9724–9758, 2018.

[5] C. Giardino, M. Bresciani, P. Villa, and A. Martinelli, "Application of remote sensing in water resource management: The case study of Lake Trasimeno, Italy," *Water Resour. Manage.*, vol. 24, pp. 3885–3899, 2010.

[6] J. Li, Y. Pei, S. Zhao, R. Xiao, X. Sang, and C. Zhang, "A review of remote sensing for environmental monitoring in China," *Remote Sens.*, vol. 12, no. 7, 2020, Art. no. 1130.

[7] Q. Yuan et al., "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, 2020, Art. no. 111716.

[8] T. Wellmann et al., "Remote sensing in urban planning: Contributions towards ecologically sound policies?," *Landscape Urban Plan.*, vol. 204, 2020, Art. no. 103921.

[9] A. M. Coutts, R. J. Harris, T. Phan, S. J. Livesley, N. S. Williams, and N. J. Tapper, "Thermal infrared remote sensing of urban heat: Hotspots, vegetation, and an assessment of techniques for use in urban planning," *Remote Sens. Environ.*, vol. 186, pp. 637–651, 2016.

[10] X. Dai, M. Xia, L. Weng, K. Hu, H. Lin, and M. Qian, "Multi-scale location attention network for building and water segmentation of remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5609519.

[11] F. N. Kogan, "Remote sensing of weather impacts on vegetation in non-homogeneous areas," *Int. J. Remote Sens.*, vol. 11, no. 8, pp. 1405–1419, 1990.

[12] E. R. Vivoni, M. Gebremichael, C. J. Watts, R. Bindlish, and T. J. Jackson, "Comparison of ground-based and remotely-sensed surface soil moisture estimates over complex terrain during SMEX04," *Remote Sens. Environ.*, vol. 112, no. 2, pp. 314–325, 2008.

[13] A. Sedaghat and H. Ebadi, "Distinctive order based self-similarity descriptor for multi-sensor remote sensing image matching," *ISPRS J. Photogrammetry Remote Sens.*, vol. 108, pp. 62–71, 2015.

[14] S. Salem, N. Kalyankar, and S. Khamitkar, "A comparative study of removal noise from remote sensing image," *Int. J. Comput. Sci.*, vol. 7, no. 1, pp. 32–36, 2010.

[15] L. Bruzzone and B. Demir, "A review of modern approaches to classification of remote sensing data," in *Land Use and Land Cover Mapping in Europe: Practices & Trends*, Berlin, Germany: Springer-Verlag, 2014, pp. 127–143.

[16] L. S. Macarringue, É. L. Bolfe, and P. R. M. Pereira, "Developments in land use and land cover classification techniques in remote sensing: A review," *J. Geographic Inf. Syst.*, vol. 14, no. 1, pp. 1–28, 2022.

[17] G. M. Foody, "Status of land cover classification accuracy assessment," *Remote Sens. Environ.*, vol. 80, no. 1, pp. 185–201, 2002.

[18] T. R. Loveland et al., "Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data," *Int. J. Remote Sens.*, vol. 21, no. 6/7, pp. 1303–1330, 2000.

[19] Y. Chen and P. Gong, "Clustering based on Eigenspace transformation–CBEST for efficient classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 83, pp. 64–80, 2013.

[20] M. Pal and P. M. Mather, "An assessment of the effectiveness of decision tree methods for land cover classification," *Remote Sens. Environ.*, vol. 86, no. 4, pp. 554–565, 2003.

[21] S. Chu, P. Li, and M. Xia, "MFGAN: Multi feature guided aggregation network for remote sensing image," *Neural Comput. Appl.*, vol. 34, no. 12, pp. 10157–10173, 2022.

[22] K. Hu, E. Zhang, M. Xia, L. Weng, and H. Lin, "Mcanet: A multi-branch network for cloud/snow segmentation in high-resolution remote sensing images," *Remote Sens.*, vol. 15, no. 4, 2023, Art. no. 1055.

[23] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogrammetry Remote Sens.*, vol. 152, pp. 166–177, 2019.

[24] S. Miao, M. Xia, M. Qian, Y. Zhang, J. Liu, and H. Lin, "Cloud/shadow segmentation based on multi-level feature enhanced network for remote sensing imagery," *Int. J. Remote Sens.*, vol. 43, no. 15/16, pp. 5940–5960, 2022.

[25] K. Hu et al., "MCSGNet: A encoder–decoder architecture network for land cover classification," *Remote Sens.*, vol. 15, no. 11, 2023, Art. no. 2810.

[26] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, May 2017.
[27] J. Zhao et al., "A land cover classification method for high-resolution remote sensing images based on NDVI deep learning fusion network," *Remote Sens.*, vol. 14, no. 21, 2022, Art. no. 5455.
[28] B. Jin, P. Ye, X. Zhang, W. Song, and S. Li, "Object-oriented method combined with deep convolutional neural networks for land-use-type classification of remote sensing images," *J. Indian Soc. Remote Sens.*, vol. 47, pp. 951–965, 2019.
[29] Y. Liang, J. Zhang, S. Zhao, R. Wu, Y. Liu, and S. Pan, "Omni-frequency channel-selection representations for unsupervised anomaly detection," *IEEE Trans. Image Process.*, vol. 32, pp. 4327–4340, 2023.
[30] S. Liu, K. Liu, W. Zhu, Y. Shen, and C. Fernandez-Granda, "Adaptive early-learning correction for segmentation from noisy annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2606–2616.
[31] G. Wang et al., "A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2653–2663, Aug. 2020.
[32] X. Luo, J. Zhang, K. Yang, A. Roitberg, K. Peng, and R. Stiefelhagen, "Towards robust semantic segmentation of accident scenes via multi-source mixed sampling and meta-learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4429–4439.
[33] J. Zhang, K. Yang, and R. Stiefelhagen, "ISSAFE: Improving semantic segmentation in accidents by fusing event-based data," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 1132–1139.
[34] Y. Shu, X. Wu, and W. Li, "LVC-Net: Medical image segmentation with noisy label based on local visual cues," in *Proc. 22nd Int. Conf. Med. Image Comput. Comput. Assist. Intervention*, 2019, pp. 558–566.
[35] K. Makantasis, A. D. Doulamis, N. D. Doulamis, and A. Nikitakis, "Tensor-based classification models for hyperspectral data analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 6884–6898, Dec. 2018.
[36] K. Makantasis, K. Karantzalos, A. Doulamis, and N. Doulamis, "Deep supervised learning for hyperspectral data classification through convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 4959–4962.
[37] I. N. Tzortzis et al., "Tensor-based learning for detecting abnormalities on digital mammograms," *Diagnostics*, vol. 12, no. 10, 2022, Art. no. 2389.
[38] R. Hang, F. Zhou, Q. Liu, and P. Ghamisi, "Classification of hyperspectral images via multitask generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1424–1436, Feb. 2021.
[39] R. Hang, P. Yang, F. Zhou, and Q. Liu, "Multiscale progressive segmentation network for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5412012.
[40] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa, "Joint optimization framework for learning with noisy labels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5552–5560.
[41] H. Zhu, J. Shi, and J. Wu, "Pick-and-learn: Automatic quality evaluation for noisy-labeled image segmentation," in *Proc. 22nd Int. Conf. Med. Image Comput. Comput. Assist. Intervention*, 2019, pp. 576–584.
[42] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, "Multilabel remote sensing image retrieval based on fully convolutional network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 318–328, 2020.
[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[44] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.
[45] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork, "Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 2443–2449.
[46] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
[47] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc.18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.
[48] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. 4th Int. Workshop Deep Learn. 8th Int. Workshop Med. Image Anal. Multimodal Learn. Clin. Decis. Support, Held Conjunction MICCAI*, 2018, pp. 3–11.
[49] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.
[50] G. Li, I. Yun, J. Kim, and J. Kim, "DABNet: Depth-wise asymmetric bottleneck for real-time semantic segmentation," in *Proc. 30th Brit. Mach. Vis. Conf.*, 2019.
[51] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 12077–12090, 2021.
[52] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 603–612.
[53] O. Oktay et al., "Attention U-net: Learning where to look for the pancreas," *Med. Imag. Deep Learn.*, 2018.
[54] T. Sikora, "The MPEG-7 visual standard for content description—An overview," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 696–702, Jun. 2001.

**Jianbo Xiao** received the B.Sc. degree in remote sensing science and technology from Chang'an University, Xi'an, China, in 2021. He is currently working toward the master's degree in surveying and mapping science and technology with the Nanjing University of Posts and Telecommunications, Nanjing, China.

His research interests include deep learning and remote sensing applications.

**Taotao Cheng** received the B.Eng. degree in surveying and mapping engineering from Chengdu University, Chengdu, China, in 2022. He is currently working toward the M.E. degree in surveying and mapping from the Nanjing University of Posts and Telecommunications, Nanjing, China.

His research interests include deep learning, and remote sensing image application.

**Deliang Chen** received the Ph.D. degree in geodesy and survey engineering from He'Hai University, Nanjing, China, in 2014.

Since 2014, he has been with the School of Geography and Bioinformatics, Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include deep learning, GIS, and remote sensing applications.

**Hui Chen** received the M.A.Sc. degree in photogrammetry and remote sensing from Nanjing Normal University, Nanjing, China, in 2020. He is currently working toward the Ph.D. degree in geography with Nanjing University, Nanjing, China.

His research interests include deep learning and remote sensing processing.

**Ning Li** received the B.Sc. degree in geographic information science from Chang'an University, Xi'an, China, in 2014. He is currently working toward the Ph.D. degree in geography with Nanjing University, Nanjing, China.

**Yanyan Lu** received the M.A.Sc. degree in geodesy and survey engineering from He'Hai University, Nanjing, China, in 2017.

Since 2014, he has been with Nanjing Audit University, Nanjing, China. His research interests include InSAR, GIS, and remote sensing applications.

**Liang Cheng** received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2008.

He is currently a Professor with the Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Nanjing University, Nanjing, China. His research interests include integration of multisource spatial data and remote sensing image processing.