





## Article

# MKF-NET: KAN-Enhanced Vision Transformer for Remote Sensing Image Segmentation

Ning Ye <sup>1</sup>, Yi-Han Xu <sup>1,\*</sup> , Wen Zhou <sup>2</sup> , Gang Yu <sup>3</sup>  and Ding Zhou <sup>4</sup> 

<sup>1</sup> College of Information Science and Technology & Artificial Intelligence, Nanjing Forestry University, Nanjing 210037, China; yn745862233@njfu.edu.cn

<sup>2</sup> School of Low Altitude Equipment and Intelligent Control, Guangzhou Maritime University, Guangzhou 510000, China; wenzhou@ustc.edu

<sup>3</sup> Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield S10 2TN, UK; gyu2@sheffield.ac.uk

<sup>4</sup> Faculty of Engineering, Universiti Malaya, Kuala Lumpur 50603, Malaysia; zhou ding@um.edu.my

\* Correspondence: xuyihan@njfu.edu.cn

## Abstract

Remote sensing images, which obtain surface information from aerial or satellite platforms, are of great significance in fields such as environmental monitoring, urban planning, agricultural management, and disaster response. However, due to the complex and diverse types of ground coverage and significant differences in spectral characteristics in remote sensing images, achieving high-quality semantic segmentation still faces many challenges, such as blurred target boundaries and difficulty in recognizing small-scale objects. To address these issues, this study proposes a novel deep learning model, MKF-NET. The fusion of KAN convolution and Vision Transformer (ViT), combined with the multi-scale feature extraction and dense connection mechanism, significantly improves the semantic segmentation performance of remote sensing images. Experiments were conducted on the LoveDA dataset to systematically evaluate the segmentation performance of MKF-NET and several existing traditional deep learning models (U-net, Unet++, Deeplabv3+, Transunet, and U-KAN). Experimental results show that MKF-NET performs best in many indicators: it achieved a pixel precision of 78.53%, a pixel accuracy of 79.19%, an average class accuracy of 76.50%, and an average intersection-over-union ratio of 64.31%; it provides efficient technical support for remote sensing image analysis.

**Keywords:** remote sensing imagery; deep learning; semantic segmentation; model evaluation



Academic Editor: Atsushi Mase

Received: 9 September 2025

Revised: 9 October 2025

Accepted: 10 October 2025

Published: 10 October 2025

**Citation:** Ye, N.; Xu, Y.-H.; Zhou, W.; Yu, G.; Zhou, D. MKF-NET: KAN-Enhanced Vision Transformer for Remote Sensing Image Segmentation. *Appl. Sci.* **2025**, *15*, 10905. <https://doi.org/10.3390/app152010905>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Remote sensing technology is a technology that uses air or space platforms (such as satellites) to conduct non-contact observation and data collection of the Earth's surface. It uses various sensors to capture electromagnetic radiation signals, including the spectrum, infrared band, and microwave, to reveal the spatiotemporal distribution of surface features [1]. With the continuous advancement of remote sensing technology, remote sensing images have become the core data source for obtaining surface information. The widespread use of high-spatial-resolution, multispectral, and even time series remote sensing data has enabled researchers to understand the evolution of the natural environment and the spatial pattern of human activities at an unprecedented scale and accuracy. Semantic segmentation technology can accurately identify farmland boundaries, types, and planting patterns by accurately classifying targets at the pixel level, providing

data support for precision agriculture management [2,3]. It is able to accurately extract building outlines to provide guarantees for urban space planning [4,5]. It can provide a scientific basis for ecosystem monitoring and protection [6], and the ability to quickly assess disaster impacts and provide real-time information for disaster relief and post-disaster recovery [7]. Therefore, achieving high-precision segmentation is an important research direction. However, remote sensing images have the characteristics of wide coverage, large data volume, complex object categories, and large scale differences, which pose challenges to the task of accurate semantic segmentation.

In the early traditional segmentation methods, fixed threshold segmentation is a commonly used technique [8]. The principle is to set a threshold. When the grayscale value is greater than the threshold, the pixel is set to 255, and when the grayscale value is less than or equal to the threshold, the pixel is set to 0. However, this method is only applicable to binary classification tasks. While multi-classification tasks can be achieved by setting multiple thresholds, the performance of threshold segmentation methods is highly dependent on the choice of threshold. Due to the complex nature of object classification in remote sensing images, thresholds are difficult to apply to all scenarios, resulting in unstable segmentation results. In the edge detection method based on the Sobel operator and Canny algorithm [9,10], through two  $3 \times 3$  convolution kernels, convolution calculations are performed in the horizontal and vertical directions of the image to extract the edge information of the image and detect the edges in the horizontal and vertical directions. However, it also relies on fixed operators and threshold settings, lacks understanding of complex target semantic information, has weak generalization capabilities, and cannot effectively handle the multi-scale and multi-category characteristics of remote sensing images. With the advancement of machine learning technology, the introduction of algorithms such as support vector machines and random forests has improved the performance of remote sensing image segmentation [11,12]. However, limited by the requirements of feature engineering, boundary accuracy, and generalization ability, the results are still unsatisfactory.

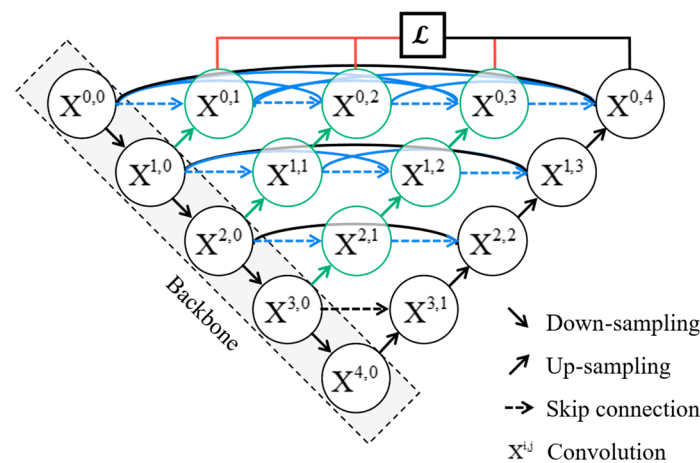
With the rapid development of artificial intelligence and the continuous improvement of computing power, deep learning has gradually emerged in image segmentation. Convolutional neural networks (CNNs) have shown significant advantages in image segmentation and provide an efficient solution for pixel-level classification tasks in complex scenes. The fully convolutional network (FCN) proposed by Shelhamer et al. ushered in a new era of image segmentation [13]. The essence of this method is to replace the traditional fully connected layer with the convolutional layer, realizing end-to-end pixel-level classification training and significantly improving segmentation efficiency and accuracy. Later, the introduction of the U-net series of models, the Deeplab series of models, and the Transformer architecture opened up new possibilities for remote sensing image segmentation tasks [14–18]. However, these models are used in a variety of different fields. For example, U-net was proposed for medical image segmentation: Medical image segmentation targets primarily microscopic and small-scale objects, and errors can directly impact medical decision-making. Remote sensing image segmentation, on the other hand, focuses on macroscopic and large-scale objects, with the core goal being to extract macroscopic geographic features. Simply applying the model to different fields will significantly reduce its effectiveness.

Therefore, in order to achieve high-precision segmentation of remote sensing images, this study absorbs and improves existing models and desirable modeling methods in other fields (such as medical imaging). A new deep learning model, MKF-NET, is proposed. It integrates KAN convolution into the Vision Transformer, enabling more flexible modeling of complex nonlinear relationships [19]. Combining multi-scale feature fusion and dense

connections, the model enhances its ability to capture objects of varying scales, compensates for spatial details that may be lost during downsampling, and addresses challenges such as blurred object boundaries and difficulty recognizing small-scale objects. Experimental results based on the LoveDA dataset demonstrate that MKF-NET outperforms several traditional deep learning models (U-net, Unet++, Deeplabv3+, Transunet, and U-KAN), thereby promoting the development of intelligent and automated remote sensing monitoring.

## 2. Related Work

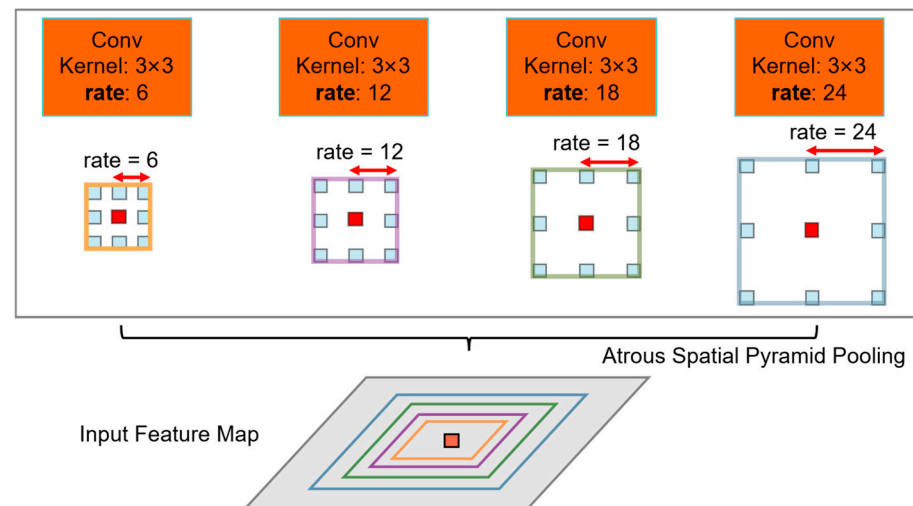
Ronneberger et al. proposed the U-net model for medical image segmentation. The model adopts a symmetrical codec structure. Its core idea is the concat operation, which transfers the multi-scale feature information extracted by the encoding layer to the corresponding decoding layer through the skip connection. This design allows the feature layer to retain more spatial details, enhance contextual information, and significantly improve segmentation accuracy. The success of U-net inspired a series of subsequent improvements, such as Unet++, which further enhanced the model's ability to capture complex boundaries and details by introducing dense skip connections and multi-scale feature fusion mechanisms (Figure 1).



**Figure 1.** Unet++ network architecture.

Zhao et al. proposed the Pyramid Scene Parsing Network (PSPNet) [20]. The model uses convolution kernels of various sizes for feature extraction. Multi-scale convolution can simultaneously capture the features of small-scale objects (such as buildings) and large-scale objects (such as forests). It extracts information from local details to global context through convolution kernels of different receptive fields, which is particularly important for processing images with significant scale differences.

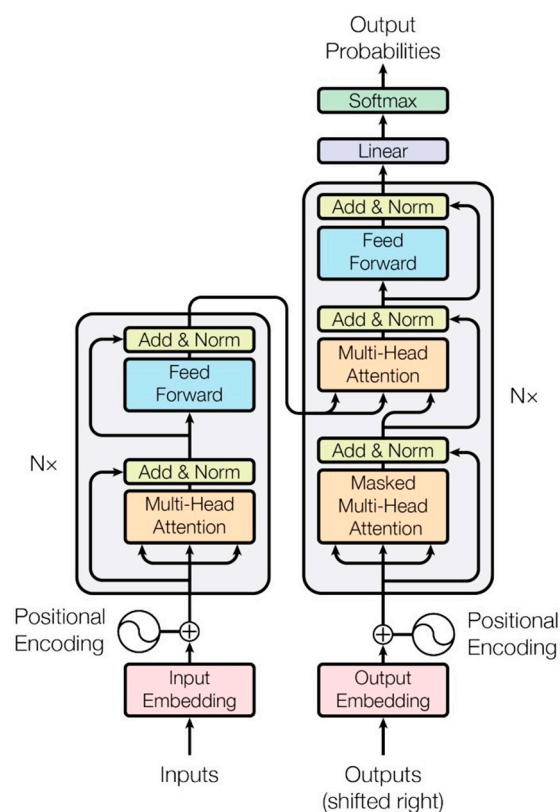
In addition, the Deeplab series of models also performs well in remote sensing image segmentation. Deeplab v1 first introduced dilated convolution to expand the receptive field. By padding the convolution kernel with zeros, this increases the receptive field of the convolution kernel without increasing computational effort, enabling it to capture a wider range of contextual information. The Deeplab v2 model proposed the ASPP module, which combines the multi-scale convolution in PSPNet and the dilated convolution in Deeplab v1, and proposes a learning rate change strategy. Deeplab v3 further improves the Atrous Spatial Pyramid Pooling (ASPP) module (Figure 2), capturing multi-scale features through serial and parallel atrous convolutions. Deeplab v3+ combines the Xception backbone network with an encoder–decoder architecture to further improve segmentation accuracy and the ability to handle complex object boundaries.



**Figure 2.** ASPP network architecture.

Inspired by the above models, Feng et al. proposed DANet. They used two ASPP modules in feature extraction to enable the model to learn feature information more effectively [21]. In the CA-UNet network proposed by Jia et al., coordinate information is incorporated into the attention mechanism module, avoiding the limitation of traditional attention mechanisms in ignoring position information and strengthening the network's ability to capture spatial positions. In the decoder, a content-aware feature reconstruction module is introduced to replace the traditional transposed convolution. This module can adaptively adjust the convolution kernel shape, thereby improving the processing of non-uniformly sampled data [22].

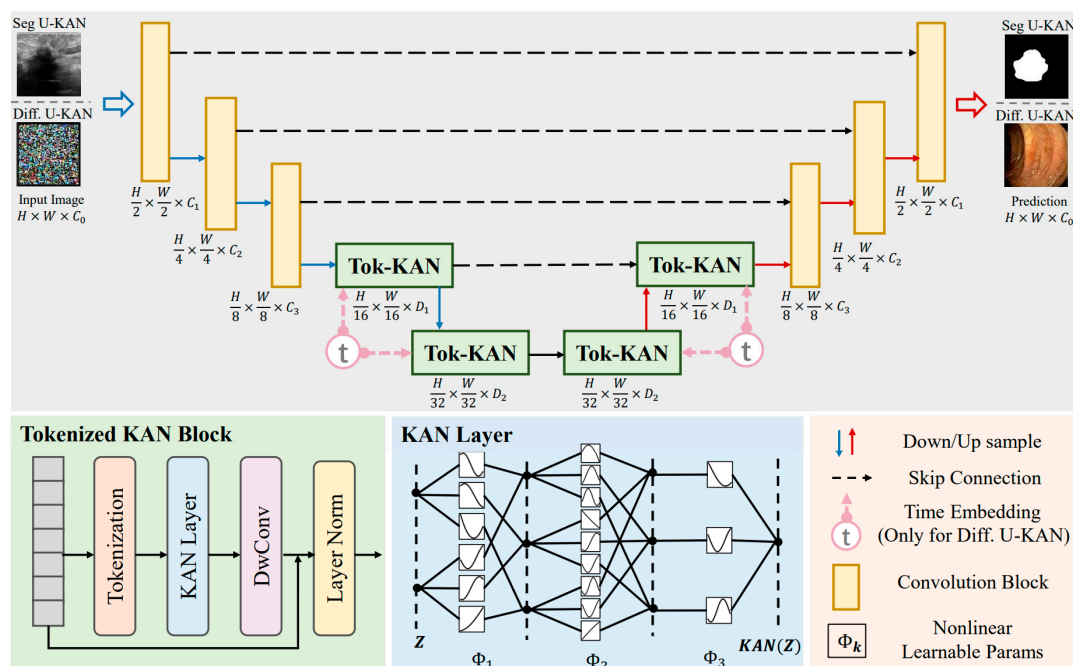
In recent years, with the explosive popularity of Transformer (Figure 3), applying Transformer to the field of semantic segmentation has also become a trend [23]:



**Figure 3.** Transformer network architecture.

Transunet enhances the model's ability to model long-distance dependencies by introducing a self-attention mechanism [24]. By adding the Transformer to the process of downsampling feature extraction, it has achieved great success in medical image segmentation. In the MBT-UNet proposed by Liu et al., they cropped images to different sizes before feeding them into the Transformer module, enabling the model to learn more diverse semantic information. They also proposed a new multi-scale fusion module (FFM) to incorporate richer semantic information. During upsampling, they proposed the MSUM architecture, which uses multiple convolutions of different sizes to extract features, fuses them, and then performs a transposed convolution. This process is like applying the ASPP module from the downsampling process to the upsampling process [25].

Last year, Liu et al. proposed the U-KAN network (Figure 4) [26]. The core idea of the KAN network is to train activation functions instead of weights. U-KAN integrates the KAN network into the U-net network and has also achieved good results in medical image segmentation tasks.



**Figure 4.** U-KAN network architecture.

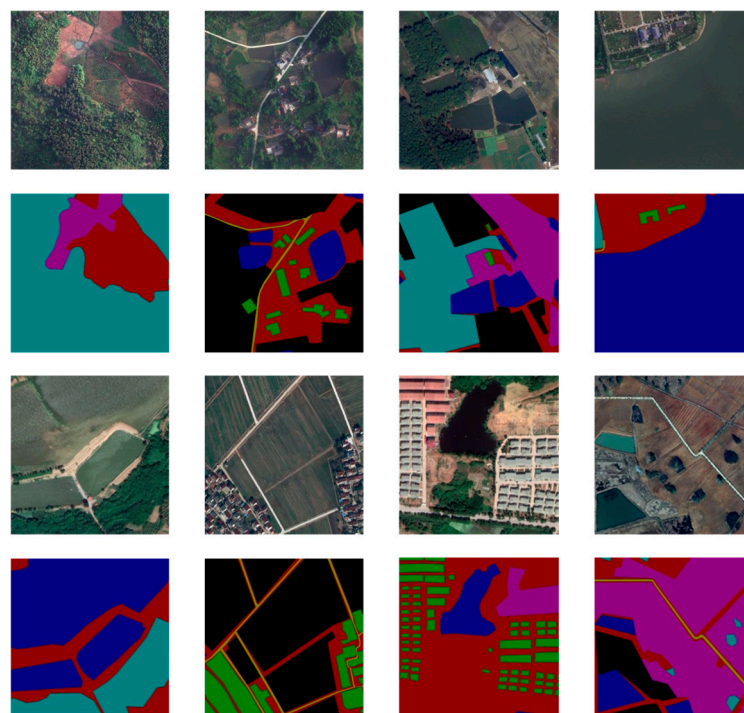
### 3. Materials and Methods

#### 3.1. Data Source and Notes

LoveDA (Land Cover Domain Adaptive Semantic Segmentation Dataset) is a benchmark dataset designed and released in 2021 by the RSDIA research team at Wuhan University [27]. It is specifically designed for semantic segmentation and unsupervised domain adaptation (UDA) tasks in high-resolution remote sensing imagery. This dataset aims to advance remote sensing image processing technology by providing high-quality remote sensing imagery data and detailed semantic annotations, particularly for handling the complexity and diversity of urban and rural land cover scenarios, providing strong data support for academic research and practical applications. The LoveDA dataset contains 5987 high-spatial-resolution optical remote sensing images covering Nanjing and Changzhou in Jiangsu Province, and Wuhan in Hubei Province. These cities, due to their diverse geographical locations, economic development levels, and land use patterns, provide rich scene diversity in the dataset. The dataset is divided into two main domains based on geographical characteristics: urban (2713 images) and rural (3274 images). Images in



the urban area mainly show dense buildings, complex road networks, and other artificial infrastructure, while images in the rural area are mainly farmland, woodland, water, and natural landscape. This division between urban and rural areas not only reflects the significant differences in land use and landscape characteristics between different geographical regions but also provides a unique experimental platform for studying cross-domain adaptation problems, especially suitable for exploring how to adapt models to the differences in spectral characteristics, texture patterns, and spatial structures between urban and rural scenes. In terms of annotation, the LoveDA dataset contains a total of 166,768 annotated objects, covering 7 semantic categories: background (1), buildings (2), roads (3), water bodies (4), barrens (5), woodlands (6), and farmland (7). The selection of these categories fully considers the common elements of remote sensing images in land cover and land use analysis, ensuring the wide applicability of the dataset. In addition, unlabeled or invalid areas are marked as 0 in the dataset and should be ignored during data preprocessing and model training to ensure the accuracy and consistency of the analysis results. The high resolution and fine annotation of the dataset make it of high research value in semantic segmentation tasks, especially in application scenarios that require accurate identification and classification of surface features. The design goal of the LoveDA dataset (Figure 5) is to promote the further development of semantic segmentation and domain adaptation technologies in the field of remote sensing by providing diverse and high-quality remote sensing image data. Its high spatial resolution and contrast characteristics of urban and rural scenes make it particularly suitable for studying cross-domain migration problems. In addition, the dataset has broad potential in practical applications and can support a variety of remote sensing-related tasks, including but not limited to forest cover monitoring, land use classification, urban planning, agricultural resource management, and ecological and environmental protection. By providing researchers with standardized data resources and a clear experimental framework, the LoveDA dataset provides a solid foundation for the development of more efficient and robust remote sensing image processing algorithms, helping to solve the challenges of remote sensing technology in complex real-world scenarios.



**Figure 5.** LoveDA dataset remote sensing images and their labels.

This study partitioned all 5987 remote sensing images in the LoveDA dataset into training, validation, and test sets in a ratio of 7:2:1. The training set was then augmented with color jittering, horizontal flipping, and vertical flipping [28]. Image enhancement is a key technique in deep learning. Deep learning models require large amounts of data to learn robust feature representations. Image enhancement compensates for data shortages by generating diverse training samples and enhancing the model's generalization capabilities. Horizontal and vertical flipping can simulate different image orientations, effectively enhancing the model's learning of directional invariance. Color dithering can simulate illumination variations in remote sensing images by randomly adjusting brightness and contrast. Zhang et al. analyzed the effects of flipping and color dithering on window state detection, analyzing the effects of traditional image enhancement methods on window state detection. They concluded that these methods enhance model robustness by simulating variations in viewpoint, illumination, and scale. The LoveDA dataset, which features complex scenes (mixed urban and rural areas) and environmental variations (illumination and seasonality), is well-suited for these image enhancement techniques [29]. This approach not only supplemented the limited training dataset but also effectively improved the model's generalization performance (Figure 6).



**Figure 6.** Augmented Images: From left to right, the images are the original image, the vertically flipped image, the color-jittered image, and the horizontally flipped image.

### 3.2. Models

The MKF-NET model utilizes an efficient encoder–decoder architecture specifically designed for the semantic segmentation of remote sensing imagery. Its multi-stage feature extraction and fusion mechanism effectively addresses the challenges of complex object distribution and multi-scale objects. The encoder layer, based on a four-stage convolutional downsampling structure, constructs feature representations from low-level to high-level layers by layer. Each stage uses convolution operations to gradually reduce the spatial resolution of the feature map while extracting deeper semantic information. To enhance the cross-scale representation of features, the encoder layer incorporates a dense connection mechanism. Upsampling concatenates the feature map of the current layer with the features of the previous layer, and then downsampling and fuses them with the features of the subsequent layers, forming a multi-layer feature aggregation network. This dense connection strategy effectively integrates contextual information at different scales, compensates for spatial details that may be lost during downsampling, and improves the model's ability to capture small-scale objects and complex boundaries.

When downsampling feature extraction, the calculation formula is as follows:

$$OH = \frac{H + 2P - FH}{S} + 1 \quad (1)$$

When upsampling, the input and output calculation formulas are as follows:

$$OH = (H - 1) \cdot S - 2P + FH \quad (2)$$

Among them, OH is the output feature size, H is the input feature size, P is the padding, FH is the convolution kernel size, and S is the step size. The overall densely connected network parameters are shown in Figure 7:

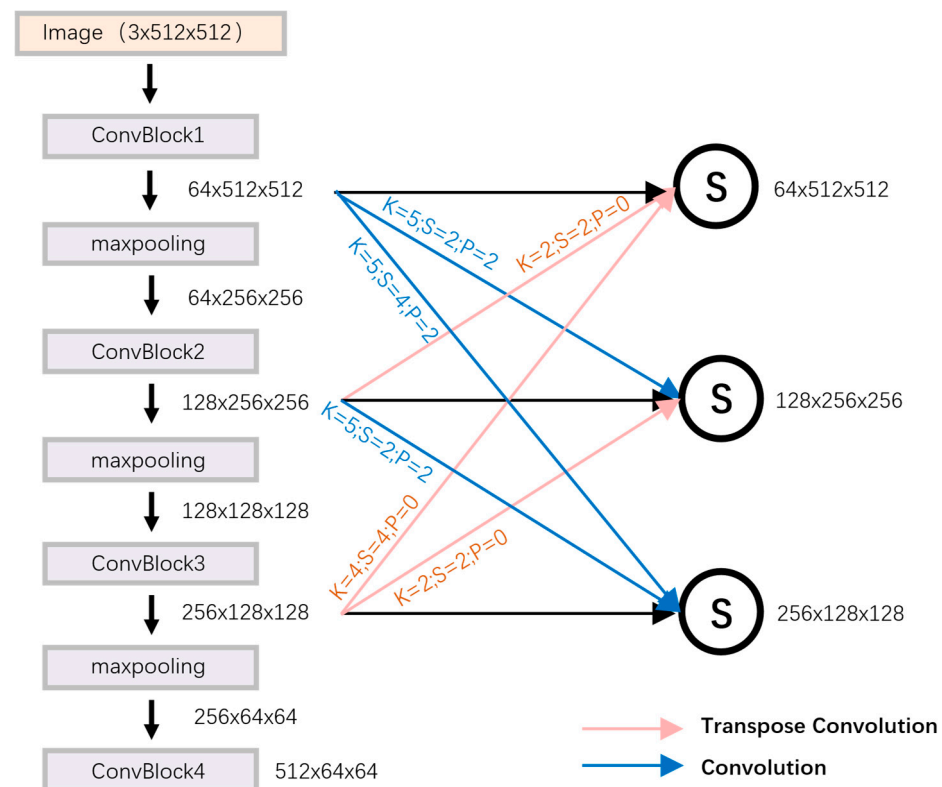


Figure 7. Densely connected network parameter diagram.

In the decoder layer, MKF-NET (Figure 8) utilizes the KAN convolution-based KiT module to extract deep features, further enhancing the robustness of the feature representation. The KiT module, combined with a self-attention mechanism, effectively integrates global semantic information with local details, optimizing the model's segmentation accuracy for complex object categories in remote sensing imagery. Through the synergy of multi-scale feature fusion and attention mechanism, MKF-NET significantly improves segmentation performance while maintaining computational efficiency. It is particularly suitable for semantic segmentation tasks of remote sensing images with significant differences between urban and rural scenes.

ConvBlock (Figure 9a) adopts a residual connection structure. By embedding a residual learning mechanism in the convolutional layer, this module can effectively improve the gradient vanishing and gradient exploding phenomena that are prone to occur during the training of deep neural networks, while greatly improving the convergence speed and efficiency of the model. The multi-scale module (Figure 9b) employs multiple sets of convolution kernels with varying dilation rates for parallel computation, effectively capturing multi-scale contextual information. This expands the model's receptive field without increasing the number of parameters or compromising feature map resolution. This allows the model to fully extract feature information containing both spatial details and semantic associations, significantly improving the perception and recognition of objects at different scales. In remote sensing image semantic segmentation tasks, multi-scale feature extraction is crucial for accurately classifying complex land feature categories (such as small objects and boundary details). It significantly improves the model's adaptability to urban and rural scene differences and complex backgrounds, thereby improving segmentation accuracy and robustness.



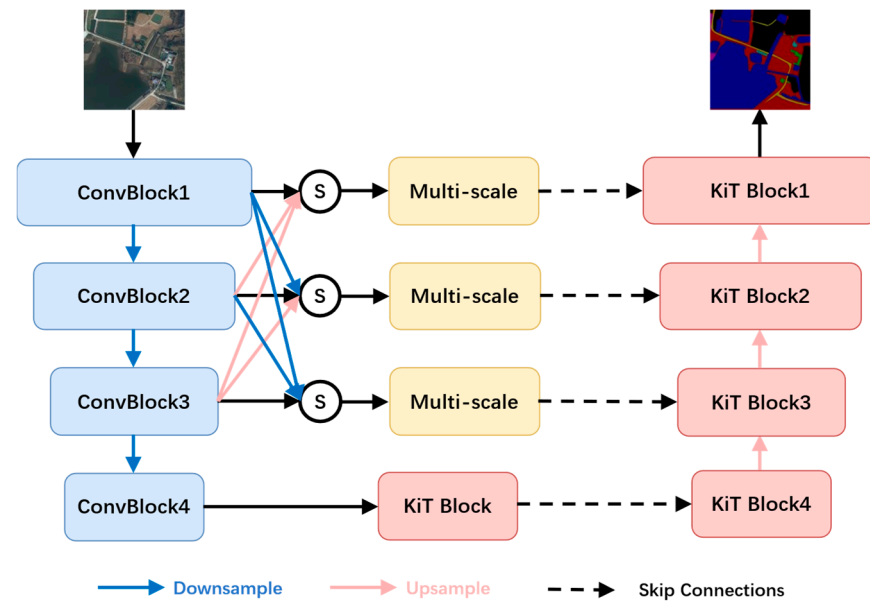


Figure 8. MKF-NET network structure.

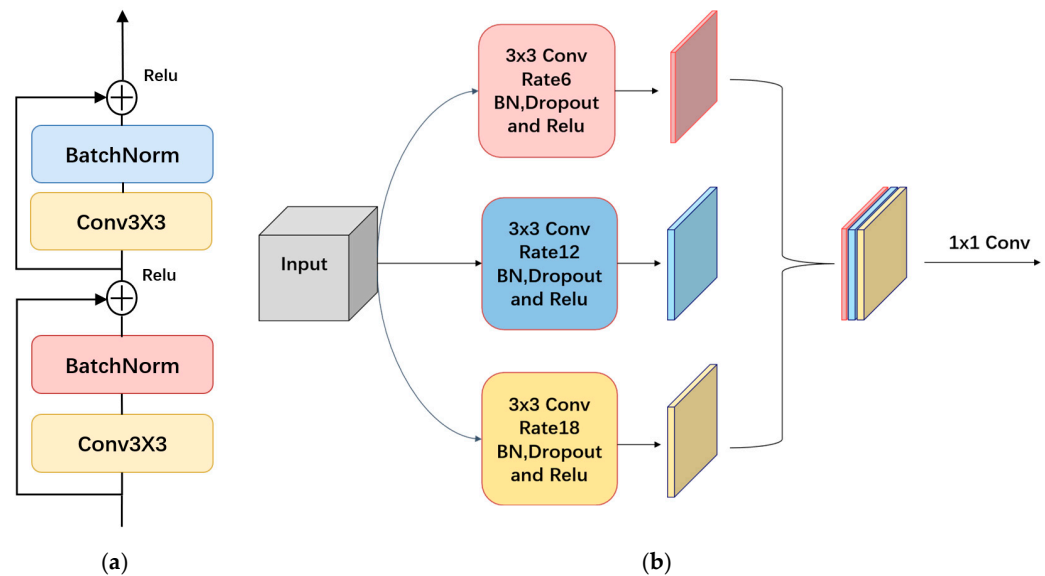


Figure 9. (a) ConvBlock network structure. (b) Multi-scale network structure.

The KiT module designs an efficient feature processing mechanism optimized for the fusion of global and local features in remote sensing image semantic segmentation tasks. This module first divides the input feature map into small blocks using convolution operations, then flattens each block (linearly maps it) to generate an embedding representation. To compensate for the Transformer's lack of sequence order perception, the KiT module adds positional encoding to each image block embedding, preserving spatial location information and ensuring the spatial contextual integrity of the features. The embedded sequence is then input to the Transformer Encoder, which integrates a multi-head self-attention mechanism and a KAN network. The multi-head self-attention mechanism effectively extracts global semantic features by modeling long-range dependencies between image blocks, enhancing the model's ability to perceive complex object distributions. The KAN network, based on the Kolmogorov-Arnold representation theorem, uses a learnable spline function instead of the fixed activation function (such as ReLU) in traditional MLPs, significantly improving the model's ability to fit complex nonlinear relationships. A specific

node of MLP and KAN in the neural network is shown in Figure 10, and the architecture of the KiT module is shown in Figure 11.

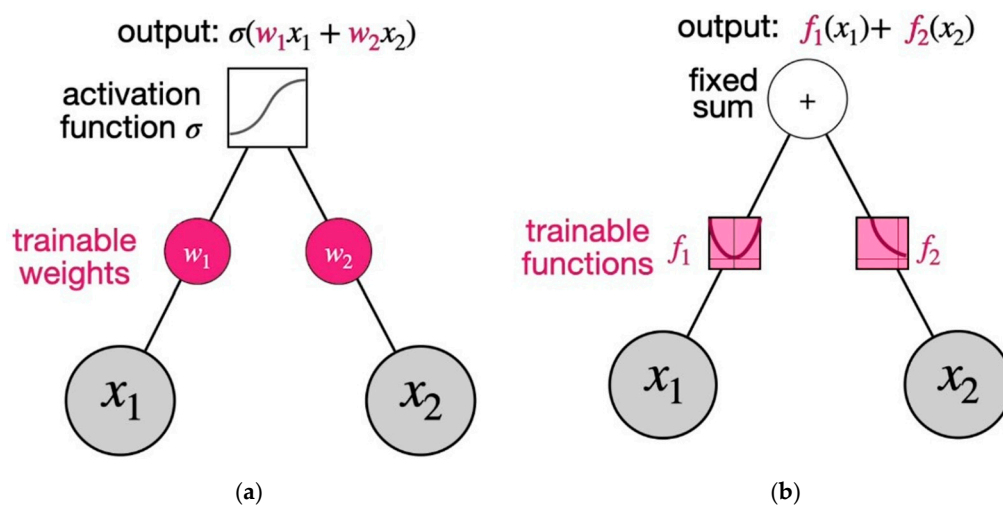


Figure 10. (a) MLP network nodes. (b) KAN network nodes.

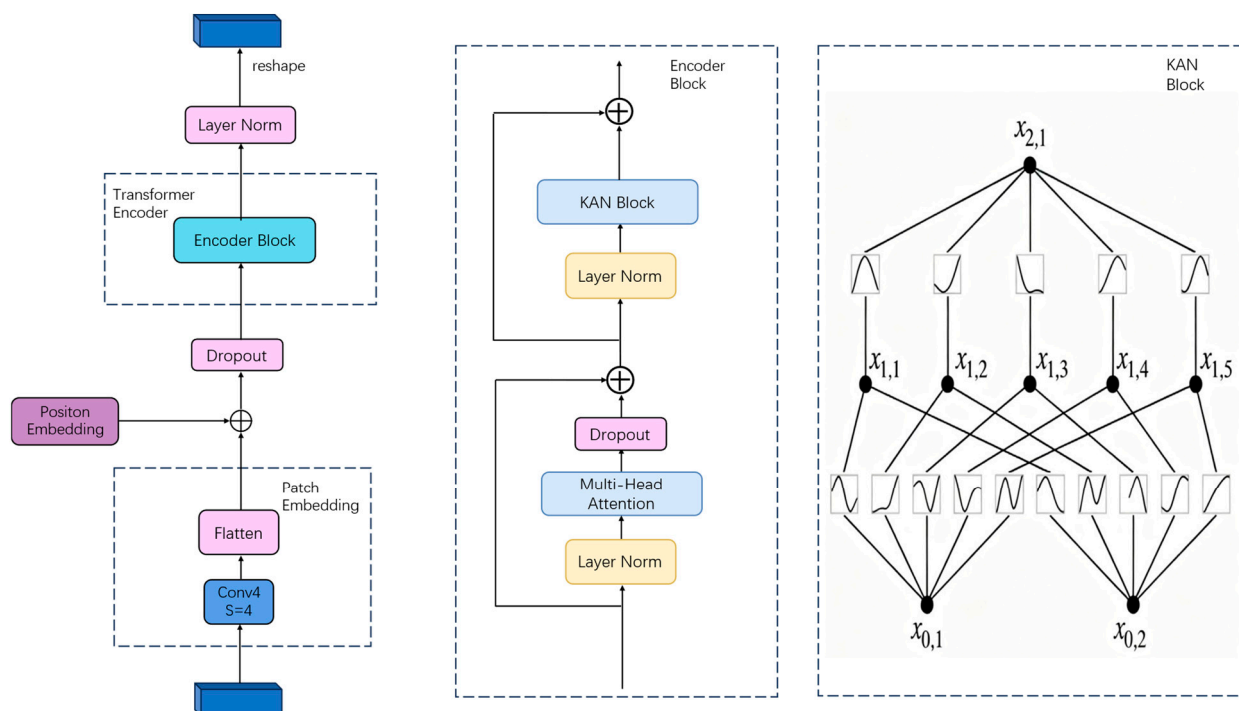
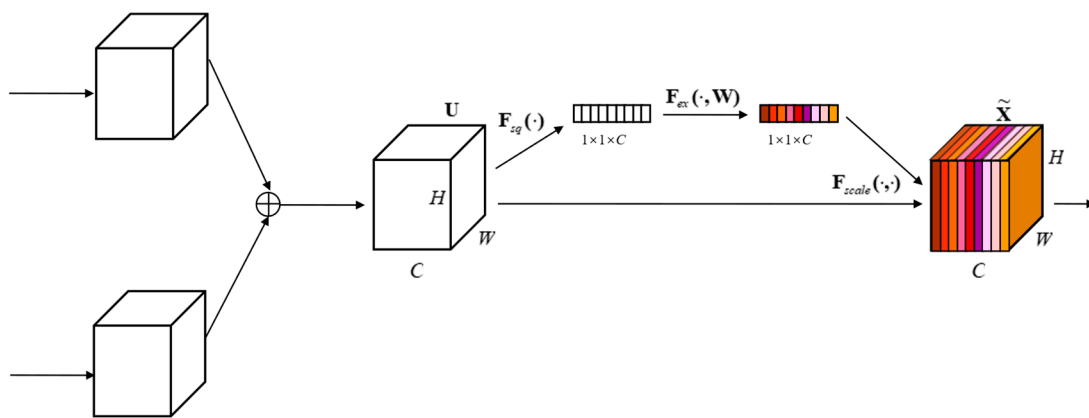


Figure 11. KiT module structure.

Figure 10a is a multi-layer perceptron with inputs  $x_1$  and  $x_2$ , weights  $w_1$  and  $w_2$ , which are trainable weights, and activation functions such as sigmoid, ReLU, or other functions. Suppose we choose sigmoid as the activation function, the output is  $\sigma(w_1x_1 + w_2x_2)$ . However, in the KAN of Figure 10b, the situation is the opposite; the input is still  $x_1$  and  $x_2$ , the trainable part is the function,  $f_1$  and  $f_2$  can be any function; the output will be  $f_1(x_1) + f_2(x_2)$ .

In the semantic segmentation of remote sensing images, the flexibility of KAN helps capture boundary details, texture features, and the complex morphology of small-scale targets, overcoming the expression limitations of traditional MLP in high-precision segmentation scenarios (such as blurred edges or small objects), thereby improving the segmentation accuracy and robustness of the model.

The SAFM module (Figure 12), also known as the Sum-Attention Feature Fusion Method, designs an efficient feature fusion mechanism to effectively integrate multi-scale features for semantic segmentation in remote sensing imagery. This module fuses shallow features (rich in spatial detail) with deep features (rich in high-level semantics) through a summation operation, achieving information complementarity while preserving their original feature representations and ensuring the integrity of detail and semantic information. To further optimize feature selection, the SAFM module introduces the Squeeze-and-Excitation (SE) attention mechanism. By dynamically weighting channel dimensions, it enhances the representation of key feature channels (such as the edges and textures of target objects) and suppresses the influence of redundant information, such as background noise. This adaptive feature weighting mechanism significantly improves the model's perception of complex object categories, particularly in remote sensing imagery with high background noise or blurred edges. Through precise feature fusion and attention modulation, the SAFM module enhances segmentation accuracy and model robustness.



**Figure 12.** SAFM module structure.

### 3.3. Evaluation Indicators

Semantic segmentation is a pixel-by-pixel classification task, and its performance is often quantified using various evaluation metrics. Key metrics include pixel classification accuracy, mean pixel accuracy (mPA), precision, recall, and mean intersection over union (mIoU). These metrics are calculated based on four predictions of the confusion matrix: true positives (TP), which are pixels correctly identified by the model as belonging to the target class; false positives (FP), which are pixels that are incorrectly labeled as belonging to the target class but actually belong to another class; false negatives (FN), which are pixels that are incorrectly classified as non-target classes but actually belong to the target class; and true negatives (TN), which are pixels correctly identified by the model as belonging to non-target classes. In the task of semantic segmentation of remote sensing images, these metrics can effectively evaluate the classification accuracy and stability of the model when dealing with complex target types (such as buildings, roads, water bodies, etc.) and multi-scale scenes. Among them, mIoU is a key metric for evaluating segmentation performance because it can comprehensively measure classification accuracy and completeness.

Accuracy reflects the pixel accuracy of semantic segmentation, which is the proportion of correct prediction results to the total prediction values.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (3)$$

Precision reflects the accuracy of the pixel categories in semantic segmentation, which is the probability that a certain category is predicted correctly in the prediction results.

High precision means that most of the pixels predicted by the model as a certain category do belong to that category.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

Recall reflects the probability that a certain category is predicted correctly in the true value; a high recall rate means that the model can find most of the actual positive pixels.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

Mean Pixel Accuracy (MPA) reflects the accuracy of the model in pixel-level classification of each category. Unlike accuracy, MPA calculates the accuracy of each category separately and then averages it to reduce the impact of category imbalance.

$$\text{MPA} = \frac{1}{n} \times \sum_{i=1}^n \frac{TP_i}{FN_i + TP_i} \quad (6)$$

The Intersection over Union (IoU) measures the degree of overlap between the predicted segmented area and the ground-truth segmented area. A higher IoU value indicates a better agreement between the model's predictions and the actual annotations. The mean Intersection over Union (mIoU) is the average of the IoUs across all categories and serves as a comprehensive metric for evaluating segmentation performance.

$$\text{mIoU} = \frac{1}{n} \times \sum_{i=1}^n \frac{TP_i}{FN_i + TP_i + FP_i} \quad (7)$$

The Dice loss function is a loss function specifically designed for image segmentation tasks. Its core concept is to measure the degree of overlap between the predicted segmentation results and the ground-truth annotations, and to improve model performance by optimizing the difference between the two. Compared to the cross-entropy loss, the Dice loss performs better in dealing with class imbalance because it comprehensively considers the contribution of each pixel rather than simply weighting the number of pixels [30].

$$\text{Dice Loss} = 1 - \frac{1}{C} \sum_{c=1}^C \frac{2 \sum_{i=1}^N p_{i,c} g_{i,c}}{\sum_{i=1}^N p_{i,c} + \sum_{i=1}^N g_{i,c} + \epsilon} \quad (8)$$

where  $p_{i,c}$  represents the predicted probability that pixel  $i$  belongs to category  $c$ ,  $g_{i,c}$  represents the true label of pixel  $i$ , and  $\sum_{i=1}^N p_{i,c} g_{i,c}$  represents the intersection (overlapping area) of the predicted probability and the true label.  $\sum_{i=1}^N p_{i,c}$  represents the total probability predicted as category  $c$ , and  $\sum_{i=1}^N g_{i,c}$  represents the total number of pixels whose true label is category  $c$ .  $\epsilon$  represents a smoothing term to prevent division by zero.

## 4. Results

All models in this experiment were tested on a computer equipped with an NVIDIA 4050 GPU and 14 GB of memory, running CUDA 11.8, PyTorch 2.0.0, and Python 3.8. The input image features are  $3 \times 512 \times 512$ , the batch size is 4, no pre-trained weights are used, and each model is trained for 200 rounds to obtain the best weight. Training uses the adaptive learning rate optimizer, Adam, with an initial learning rate of 0.0001. This is the recommended value for the Adam optimizer and is suitable for semantic segmentation tasks, avoiding instability caused by excessively large initial learning rates. The minimum learning rate is 0.000001, which is used for learning rate decay to ensure stability during later training. Adam's momentum parameter ( $\beta_1$ ) controls the first-order momentum update rate. The default value of 0.9 is suitable for most scenarios. A cosine annealing

learning rate schedule is used, gradually decaying from Init\_lr to Min\_lr. This smooths the learning rate, helps escape local optima, and is suitable for long-term training. A fixed random seed of 11 is used to ensure reproducibility.

#### 4.1. Comparison of Evaluation Parameters of Different Models

This paper trains five different semantic segmentation models and calculates the effects of precision, accuracy, average pixel accuracy, and average intersection-over-union ratio in each model (Table 1).

**Table 1.** Performance of different models.

Model	Accuracy	Precision	MPA	mIoU
U-net	76.36%	79.28%	70.45%	58.14%
Unet++	78.6%	76.31%	74.12%	60.57%
Transunet	74.64%	79.28%	63.66%	54.71%
Deeplabv3+	79.06%	78.48%	75.04%	62.65%
U-KAN	78.99%	76.20%	74.76%	61.27%
MKF-NET(Ours)	79.19%	78.53%	76.50%	64.31%

Experimental results on the LoveDA dataset show that MKF-NET achieves the best results compared to other models in terms of pixel accuracy (79.19%), average pixel accuracy (76.50%), and average intersection-over-union (mIoU) (64.31%). However, it lags slightly behind U-net and Transunet in precision. This is because the model is overly cautious when predicting the positive class, only predicting positive for pixels with high confidence. This results in high accuracy but also the risk of missed detections. Overall, MKF-NET outperforms other models in semantic segmentation evaluation metrics.

At the same time, the intersection-over-union (IoU) of each category is calculated (Table 2).

**Table 2.** Per-class IoU (%).

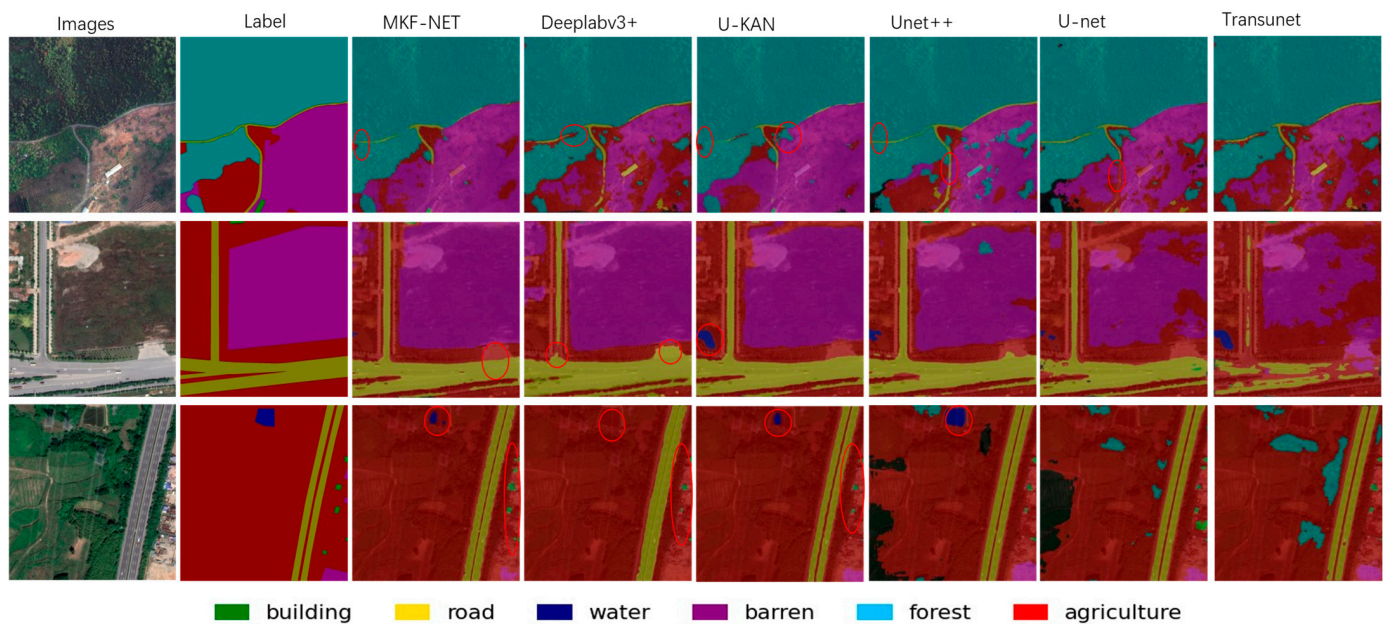
Model	Background	Water	Barren	Forest	Agriculture	Road	Building
U-net	66%	55%	67%	43%	66%	58%	52%
Unet++	73%	50%	69%	52%	76%	56%	48%
Transunet	62%	57%	65%	32%	63%	55%	49%
Deeplabv3+	73%	57%	71%	50%	79%	56%	52%
U-KAN	72%	54%	71%	49%	74%	58%	51%
MKF-NET(Ours)	74%	60%	73%	52%	75%	63%	53%

The data shows that MKF-NET achieves excellent results in multiple categories. It achieves 60% accuracy in the water category, 52% in the forest category, 73% in the barren category, 63% in the road category, and 53% in the building category, all of which outperform the other comparison models (U-net, Unet++, Transunet, Deeplabv3+, and U-KAN) in the corresponding categories. Although it performs inferior to Deeplabv3+ in the agriculture category, MKF-NET demonstrates strong segmentation capabilities in small object recognition.

#### 4.2. Segmentation Effects of Different Models

This study uses different models to segment remote sensing images. The effect after segmentation is as follows (Figure 13).





**Figure 13.** Segmentation results of different models.

From the first segmentation result, Unet++ performs best in segmenting the road boundary, but its performance in segmenting the barren boundary is not satisfactory. It also misses the small object on the left boundary of the image. Although MKF-NET is not as good as Unet++ in the boundary segmentation effect of the road category, it does not miss small targets and does better than other models in the details of the barren. In the second segmentation result, MKF-NET performs best. Deeplabv3+ also performs well, but has obvious flaws in handling the boundaries of the range class. The other models all exhibit some misclassification issues. Similarly, in the three segmentation result images, although Unet++ performs best in the water category, a large number of misclassifications occur. Compared to U-KAN, MKF-NET has a higher degree of small object detection. Deeplabv3+ missed the classification of water and was not as good as MKF-NET in the segmentation effect of the small object category building. Overall, MKF-NET has the best segmentation effect; U-net, the baseline model, achieved decent results.

Unet++ builds on U-net by adding dense connections, resulting in better segmentation of boundaries and small objects. Deeplabv3+ uses a large number of dilated convolutions to increase the receptive field, resulting in good overall segmentation results, but poorer control of details. Although Transunet can better obtain global context information through the Transformer module, in remote sensing image segmentation, relying solely on global information may not be enough, and accurate local feature positioning is also required, which has the worst effect. The U-KAN model's segmentation performance is impressive, but it is still slightly inferior to the MKF-NET model, which incorporates multi-scale fusion and dense connections. The segmentation results show a high number of misclassifications. This is likely due to the impact of similarly colored categories in the dataset on classification accuracy, making segmentation more challenging. This is particularly true in scenes with highly similar visual features (such as color and texture). This effect primarily affects the model's ability to distinguish between categories, potentially leading to misclassifications (for example, confusing forestry with agriculture).

## 5. Discussion

Convolution is the “eye” that obtains image features, and the size of the convolution kernel determines the range of image features that these “eyes” see each time; MKF-NET computes convolution kernels of different sizes on the original image, allowing the model to

see feature information of different sizes. Multi-scale feature fusion utilizes this to enhance the model's segmentation performance for objects of different sizes in the dataset (such as small roads and large areas of farmland). The problem of fuzzy target boundaries and the difficulty of detecting small targets are often due to the insufficient fusion of high-level semantic features and low-level detail features, and the features of small targets are diluted or submerged in the deep network. Dense connections enable the network to fuse features at different levels (from low-level details to high-level semantic information). Shallow features retain the location and detailed information of many small targets, while deep features contain the global semantics of large targets. Through dense connections, each layer can directly obtain the output of the previous layer, allowing small target features to participate in the decoding process earlier. This is equivalent to using multi-scale features for prediction at the same time, thereby maintaining high sensitivity to targets of different sizes, reducing information loss caused by downsampling, and allowing the restored feature map to retain more original spatial details. MKF-NET also uses the KIT module, which consists of ViT and KAN. ViT's self-attention mechanism directly models global pixel dependencies, eliminating the need for layer-by-layer stacking to expand the receptive field like a CNN. This makes it more efficient in capturing correlated features of large or multi-scale objects. For example, when detecting water and forests in remote sensing images, ViT's self-attention can directly correlate the context of the two (such as the spatial relationship between water and forest), avoiding the scale fragmentation caused by the local receptive field of CNN [31–34]. The KAN network is based on precise feature approximation using mathematical interpolation. Its core approach is to approximate arbitrarily complex functions using the superposition of single-variable functions. This allows it to accurately fit the local feature distribution of objects of varying sizes. While CNNs compress feature dimensions through downsampling, which inevitably loses detail, KANs preserve more detail by directly mapping input pixels to output features. Compared to CNNs, which use convolution kernels to “hard-extract” edges, KANs can more accurately restore the true contours of blurred boundaries. It is precisely because these designs complement each other that, in the experimental results, MKF-NET achieved the best results in small roads and buildings, with an intersection-over-union ratio of 63% and 53%, respectively, providing an effective approach to solving the problems of blurred target boundaries and difficult recognition of small-scale objects. In the future, the model can be extended to other remote sensing datasets to verify its generalization ability in data of different resolutions and modalities.

## 6. Conclusions

This study proposes a novel deep learning model, MKF-NET, that significantly improves semantic segmentation performance for remote sensing imagery by fusing KAN convolution with the Vision Transformer (ViT). MKF-NET leverages the nonlinear feature extraction of KAN convolution, the global context modeling of ViT, the multi-scale feature preservation of cross-layer connections, and the feature fusion optimization of the additive attention mechanism to achieve high-accuracy segmentation of remote sensing images. Experiments on the LoveDA dataset demonstrate that MKF-NET achieves a pixel precision of 78.53%, a pixel accuracy of 79.19%, an average class accuracy of 76.50%, and an average intersection-over-union (mIoU) of 64.31%, significantly outperforming traditional models such as U-net, Unet++, Deeplabv3+, Transunet, and U-KAN. This study proposes a reliable technical approach for the application of deep learning technology in remote sensing imagery. Future research directions include optimizing semantic segmentation models, improving detail perception, and exploring techniques such as model pruning and quantization to reduce computational load and memory usage.

**Author Contributions:** Conceptualization, N.Y.; Methodology, N.Y.; Software, N.Y.; Validation, N.Y.; Formal analysis, Y.-H.X., W.Z. and G.Y.; Investigation, Y.-H.X. and G.Y.; Resources, Y.-H.X., G.Y. and D.Z.; Data curation, G.Y. and D.Z.; Writing—original draft, N.Y.; Writing—review & editing, Y.-H.X.; Visualization, W.Z.; Project administration, Y.-H.X. and W.Z.; Funding acquisition, Y.-H.X. and W.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by: National Natural Science Foundation of China: 61601275; Jiangsu Graduate Research and Practice Innovation Program: SJCX24\_0385.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Navalgund, R.R.; Jayaraman, V.; Roy, P. Remote sensing applications: An overview. *Curr. Sci.* **2007**, *93*, 1747–1766.
2. Huan, H.; Liu, Y.; Xie, Y.; Wang, C.; Xu, D.; Zhang, Y. MAENet: Multiple attention encoder–decoder network for farmland segmentation of remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [\[CrossRef\]](#)
3. Zhang, R.; Chen, J.; Feng, L.; Li, S.; Yang, W.; Guo, D. A refined pyramid scene parsing network for polarimetric SAR image semantic segmentation in agricultural areas. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [\[CrossRef\]](#)
4. Jung, H.; Choi, H.-S.; Kang, M. Boundary enhancement semantic segmentation for building extraction from remote sensed image. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–12. [\[CrossRef\]](#)
5. Wang, L.; Fang, S.; Meng, X.; Li, R. Building extraction with vision transformer. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [\[CrossRef\]](#)
6. Xu, C.; Wang, J.; Sang, Y.; Li, K.; Liu, J.; Yang, G. An effective deep learning model for monitoring mangroves: A case study of the Indus Delta. *Remote Sens.* **2023**, *15*, 2220. [\[CrossRef\]](#)
7. Cui, L.; Jing, X.; Wang, Y.; Huan, Y.; Xu, Y.; Zhang, Q. Improved swin transformer-based semantic segmentation of postearthquake dense buildings in urban areas using remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *16*, 369–385. [\[CrossRef\]](#)
8. Pal, S.K.; Ghosh, A.; Shankar, B.U. Segmentation of remotely sensed images with fuzzy thresholding, and quantitative evaluation. *Int. J. Remote Sens.* **2000**, *21*, 2269–2300. [\[CrossRef\]](#)
9. Kanopoulos, N.; Vasanthavada, N.; Baker, R.L. Design of an image edge detection filter using the Sobel operator. *IEEE J. Solid-State Circuits* **1988**, *23*, 358–367. [\[CrossRef\]](#)
10. Canny, J. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *PAMI-8*, 679–698. [\[CrossRef\]](#)
11. Chandra, M.A.; Bedi, S. Survey on SVM and their application in image classification. *Int. J. Inf. Technol.* **2021**, *13*, 1–11. [\[CrossRef\]](#)
12. Rigatti, S.J. Random forest. *J. Insur. Med.* **2017**, *47*, 31–39. [\[CrossRef\]](#)
13. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
14. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
15. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In Proceedings of the International Workshop on Deep Learning in Medical Image Analysis, Granada, Spain, 20 September 2018; pp. 3–11.
16. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [\[CrossRef\]](#)
17. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587. [\[CrossRef\]](#)
18. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
19. Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T.Y.; Tegmark, M. Kan: Kolmogorov-arnold networks. *arXiv* **2024**, arXiv:2404.19756. [\[PubMed\]](#)
20. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.

21. Zhao, S.; Feng, Z.; Chen, L.; Li, G. DANet: A Semantic Segmentation Network for Remote Sensing of Roads Based on Dual-ASPP Structure. *Electronics* **2023**, *12*, 3243. [[CrossRef](#)]
22. Jia, J.; Song, J.; Kong, Q.; Yang, H.; Teng, Y.; Song, X. Multi-Attention-Based Semantic Segmentation Network for Land Cover Remote Sensing Images. *Electronics* **2023**, *12*, 1347. [[CrossRef](#)]
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
24. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306. [[CrossRef](#)]
25. Liu, B.; Li, B.; Sreeram, V.; Li, S. MBT-UNet: Multi-Branch Transform Combined with UNet for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2024**, *16*, 2776. [[CrossRef](#)]
26. Li, C.; Liu, X.; Li, W.; Wang, C.; Liu, H.; Liu, Y.; Chen, Z.; Yuan, Y. U-kan makes strong backbone for medical image segmentation and generation. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; pp. 4652–4660.
27. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv* **2021**, arXiv:2110.08733.
28. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60. [[CrossRef](#)]
29. Wang, S.; Korolija, I.; Rovas, D. Impact of traditional augmentation methods on window state detection. In Proceedings of the CLIMA 2022 The 14th REHVA HVAC World Congress, Rotterdam, The Netherlands, 22–25 May 2022.
30. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
31. Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; Dosovitskiy, A. Do vision transformers see like convolutional neural networks? *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12116–12128.
32. Liu, Y.; Zhang, Y.; Wang, Y.; Hou, F.; Yuan, J.; Tian, J.; Zhang, Y.; Shi, Z.; Fan, J.; He, Z. A survey of visual transformers. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 7478–7498. [[CrossRef](#)] [[PubMed](#)]
33. Khan, A.; Rauf, Z.; Sohail, A.; Khan, A.R.; Asif, H.; Asif, A.; Farooq, U. A survey of the vision transformers and their CNN-transformer based variants. *Artif. Intell. Rev.* **2023**, *56*, 2917–2970. [[CrossRef](#)]
34. Yao, W.; Bai, J.; Liao, W.; Chen, Y.; Liu, M.; Xie, Y. From cnn to transformer: A review of medical image segmentation models. *J. Imaging Inform. Med.* **2024**, *37*, 1529–1547. [[CrossRef](#)] [[PubMed](#)]

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