electronics

MDPI

Article

FURSformer: Semantic Segmentation Network for Remote Sensing Images with Fused Heterogeneous Features

Zehua Zhang [1], Bailin Liu [1,*] and Yani Li [2]

[1] School of Computer Science and Engineering, Xi'an Technological University, Xi'an 710021, China; zhangzehua@st.xatu.edu.cn
[2] School of Statistics, Lanzhou University of Finance and Economics, Lanzhou 730101, China; liyani@lzufe.edu.cn
* Correspondence: xatulbl@xatu.edu.cn

Abstract: Semantic segmentation of remote sensing images poses a formidable challenge within this domain. Our investigation commences with a pilot study aimed at scrutinizing the advantages and disadvantages of employing a Transformer architecture and a CNN architecture in remote sensing imagery (RSI). Our objective is to substantiate the indispensability of both local and global information for RSI analysis. In this research article, we harness the potential of the Transformer model to establish global contextual understanding while incorporating an additional convolution module for localized perception. Nonetheless, a direct fusion of these heterogeneous information sources often yields subpar outcomes. To address this limitation, we propose an innovative hierarchical fusion feature information module that this model can fuse Transformer and CNN features using an ensemble-to-set approach, thereby enhancing information compatibility. Our proposed model, named FURSformer, amalgamates the strengths of the Transformer architecture and CNN. The experimental results clearly demonstrate the effectiveness of this approach. Notably, our model achieved an outstanding accuracy of 90.78% mAccuracy on the DLRSD dataset.

Keywords: semantic segmentation; remote sensing images; Transformer; CNN

Citation: Zhang, Z.; Liu, B.; Li, Y. FURSformer: Semantic Segmentation Network for Remote Sensing Images with Fused Heterogeneous Features. *Electronics* **2023**, *12*, 3113. https://doi.org/10.3390/electronics12143113

Academic Editor: Silvia Liberata Ullo

Received: 4 July 2023
Revised: 12 July 2023
Accepted: 13 July 2023
Published: 18 July 2023

1. Introduction

Multi-scale and confusing geospatial objects appear in high-resolution remote sensing images (RSI), for example, images of rain, snow, or strong light can cause the objects to reflect light or weaken the edges. In terms of global representation, these images exhibit notable disparities stemming from various factors such as time of day (seasonal variations: spring, summer, fall, winter), weather conditions, sensor disparities, and other relevant influences. Conversely, from the perspective of local representation, objects such as buildings exhibit within-class dissimilarities while sharing commonalities across classes. Specifically, buildings exhibit substantial variations in terms of size, shape, height, and function. Moreover, factors like lighting conditions, viewing angles, occlusion, and shadows further contribute to the significant heterogeneity observed in high-resolution imagery. Furthermore, complex urban scenes, comprising spectrally similar entities like roads, bare ground, and parking lots, pose additional complexities when attempting precise building extraction. Consequently, the task of semantic segmentation in remote sensing images is intrinsically challenging and has engendered considerable interest among researchers.

Traditionally, conventional segmentation methods, including region-based techniques such as region growing (SRG [1]), graph-based approaches (e.g., histogram bimodal method [2]), and artificial neural networks (ANNs [3]), have been employed for semantic segmentation of remote sensing images. However, these methods necessitate manual interventions specifically for constructing extractions. Depending on the nature of manual processing, these approaches can be categorized as either initialization-based or segmentation process optimization-based methods. Nonetheless, conventional techniques heavily

rely on human a priori knowledge and entail extensive processing times, rendering them inefficient when confronted with intricate multiclassification problems.

Since Alexnet [4] won the 2012 championship at ImageNet [5], deep learning has demonstrated impressive representational learning capabilities and has been surprisingly successful in downstream tasks in computer vision. A convolutional neural network (CNN) uses massive amounts of training data to autonomously learn details and semantic information from images. Pioneering work such as FCN [6] has enabled the effective utilization of CNN models within end-to-end semantic segmentation frameworks. Similarly, U-net [7] has found widespread use in medical image processing by employing multiscale fusion techniques, while the DeepLab [8–11] series has employed feature pyramids and Atrous Convolution to enhance the perceptual field of the convolution receptive field. Researchers have improved the general semantic segmentation networks considering the characteristics of RSI tasks and have made CNNs more successful than ever in the RSI domain as well. It is true that the convolution operator has translation invariance and a good ability to obtain local information, and researchers strive to broaden the Receptive field of the convolution operator [12–14]. However, convolutional operators still cannot effectively model global information.

Initially designed for natural language processing (NLP), the Transformer [15] architecture has witnessed a notable expansion into the realm of computer vision following the introduction of the Vision Transformer (VIT) [16]. Researchers have discovered that the Transformer architecture exhibits remarkable efficacy in capturing global information from images. Consequently, numerous outstanding models rooted in the Transformer framework have emerged as potential replacements for CNN networks. Notably, a staged approach was employed by [17] to strike a balance between input computation and memory requirements in image analysis. Additionally, Segformer [18] employed a multiscale fusion mechanism, while the simple Transformer model demonstrated commendable performance in the domain of semantic segmentation. The Transformer architecture has demonstrated remarkable achievements across diverse computer vision tasks, showcasing its adeptness in modeling remote correlations by leveraging global information [19–21]. However, when confronted with the task of extracting local information devoid of spatial induction bias, the Transformer architecture struggles to deliver satisfactory performance [22]. Furthermore, due to the nature of its one-dimensional sequence input, the Transformer architecture inherently disregards crucial information pertaining to the channel dimension. In the context of semantic segmentation, partitioning an image into patches impedes comprehensive consideration of local details and hinders the refinement of object edges.

In order to mitigate the aforementioned challenges, we present a novel semantic segmentation model for remote sensing images termed FURSformer. By leveraging the inherent strengths of both Transformer and CNN architectures, FURSformer aims to enhance the overall performance of the model. The underlying principle of FURSformer lies in utilizing the Transformer branch to capture global information, while the CNN branch preserves and extracts local information. However, a direct concatenation of these branches would result in an inadequate aggregation of multi-level features during the decoding process. To address this, we introduce the Fusion of Local and Global Information mechanism (FLGM), which selectively combines the most salient aspects of global and local information. Specifically, we enhance information affinity modeling via the integration of information interaction and cross-attention modules, enabling effective feature fusion between global and local information.

The contribution of this work is in three main areas. The advantages and disadvantages of CNN and Transformer are analyzed, and the Transformer is designed as the extraction of global information, and an additional convolution structure is used to retain detailed local information. To RSI, the details of the object are very important.

We designed the FLGM module to increase the interaction of different levels of information and modeled the affinity between deep semantic information and low-level texture information to obtain a better feature map representation.

Our proposed method, FURSformer (shown in Figure 1), achieves competitive results on the remote sensing image dataset DLSRD.

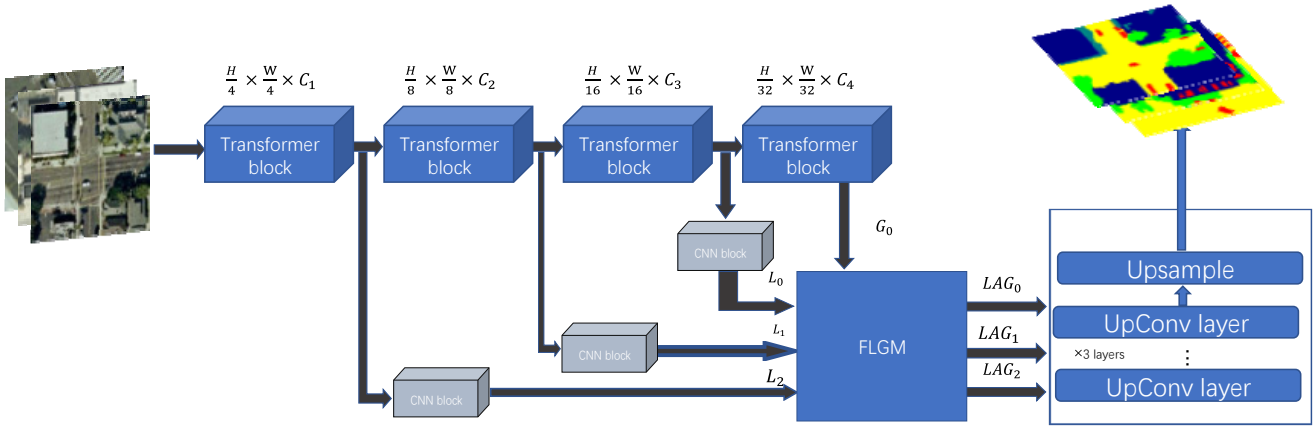

Figure 1. FURSformer network structure diagram: a semantic segmentation network for remote sensing images based on Transformer and CNN branches. The input image is obtained via the Transformer module and CNN module to obtain global and local information and then fused with the FLGM module we designed to obtain the feature map.

2. Related Work

Remote sensing images are widely used in urban planning, disaster monitoring, environmental protection, and agricultural management. Extracting and identifying information from images is the basis of these applications. Semantic segmentation, as a pixel-level image analysis technique, is one of the most important and challenging research directions in the field of image interpretation. Most of the traditional RSI semantic segmentation algorithms are based on manual feature-based machine-learning methods, such as support vector machines [23], random forests [24], and artificial neural networks [3]. These methods have poor efficiency and low generalization, which results in most image segmentation still relying mainly on manual labeling.

Semantic segmentation in remote sensing images [25,26]: 1. Remote sensing images exhibit higher resolutions, encompassing a wide range of scale variations among objects. Moreover, in addition to the objects of interest found in conventional domains (e.g., bridges, buildings, cars), remote sensing images also encompass semantically meaningful backgrounds, such as bodies of water, roads, and fields. 2. The foreground scale in remote sensing images tends to be considerably smaller compared to natural scene images, resulting in imbalanced foreground-background proportions. This poses challenges for networks when learning to accurately detect and classify smaller objects. Furthermore, remote sensing images encompass numerous complex categories (e.g., water bodies, tracks) that often exhibit fuzzy boundaries and significant spatial and spectral variations. 3. Remote sensing images entail multiple categories and diverse sources of noise, thereby introducing additional complexities to the task of semantic segmentation. The variability in multi-source data further compounds the challenges associated with accurately delineating and classifying objects within remote sensing imagery (as illustrated in Figure 2).

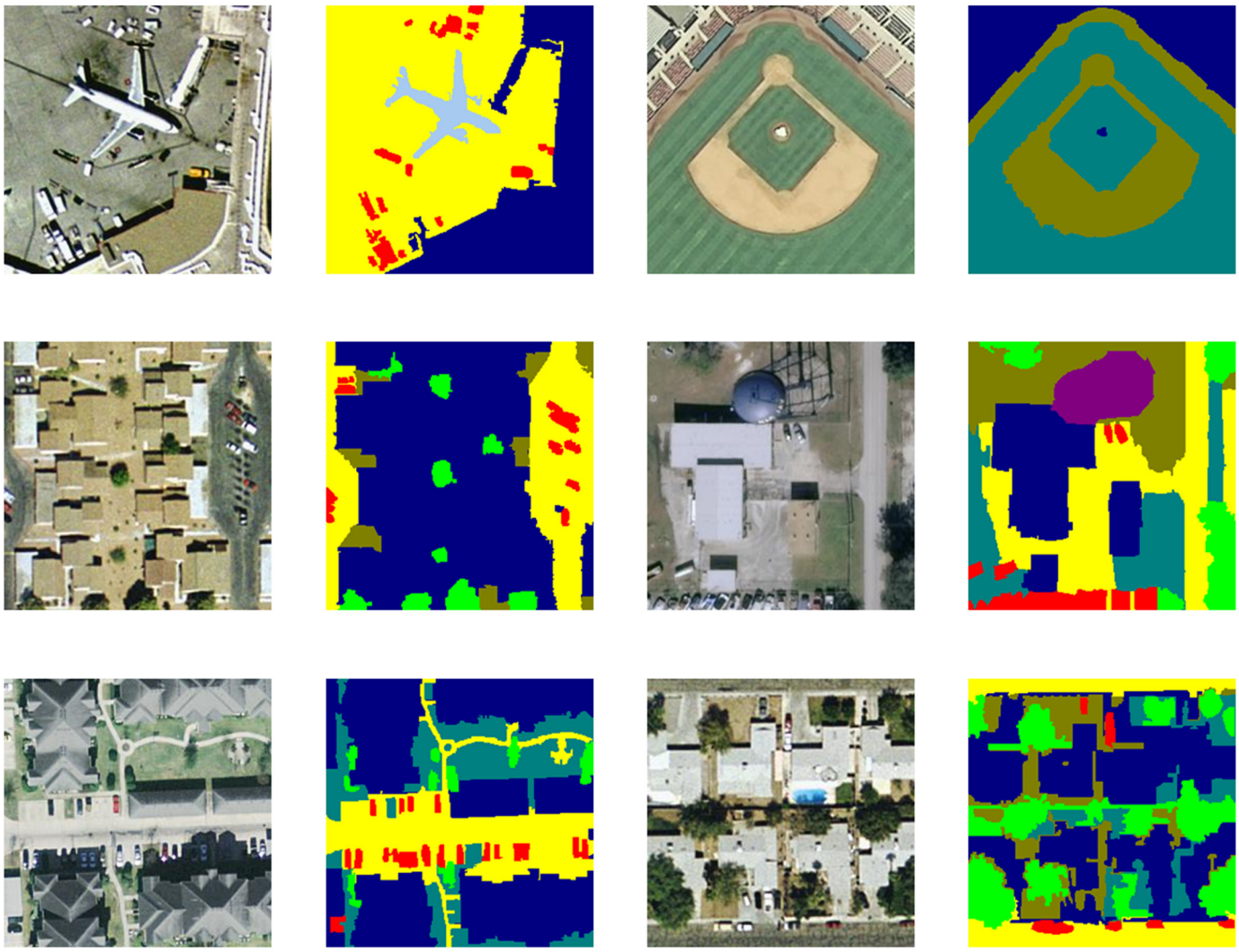

Figure 2. Partial images from the DLRSD dataset [27]. We took out several special semantic segmentation examples aiming to analyze the specificity of remote sensing images.

Remote sensing images based on CNN: Researchers have dedicated efforts to enhance the efficacy of semantic segmentation models specifically tailored for remote sensing images, resulting in the proposal of several remarkable approaches. Notably, a multi-task semantic segmentation learning model was introduced to address the challenges associated with large-resolution remote sensing images while effectively incorporating global contextual information [28]. Furthermore, researchers have put forth two-stage semantic segmentation networks [29], as well as the innovative AFNet approach that leverages multi-scale and multi-level fusion [30]. Previous studies have convincingly demonstrated that remote sensing images diverge from the general domain images, necessitating the careful design of modules to accommodate their distinctive attributes. In our work, we investigate ConvNext [31] and ResNext [32] methodologies to develop streamlined and efficient modules for the extraction of local information. This endeavor aims to overcome the limitations of Transformer architecture in capturing fine-grained details of small objects while maximizing the utilization of the CNN model's local modeling capabilities.

Remote sensing images based on Transformer: Given the remarkable achievements of the Transformer architecture across diverse computer vision applications, its feasibility has been thoroughly established. Notably, the integration of Transformer-based methodologies, such as the straightforward multi-scale fusion employed in Segformer [18], has yielded notable advancements in feature map representation. Exploiting the Transformer's

robust capacity for global modeling, previous studies have successfully applied the U-net framework in conjunction with Transformers within the domain of remote sensing, resulting in highly accurate segmentation outcomes [33]. In our research, we adopt the Transformer as an encoder to capture comprehensive global feature map representations, thereby addressing the limitations of CNN network architecture in acquiring relevant global information.

The encoder–decoder framework has been widely employed in various computer vision tasks. Notably, the DeepLab series [8–11] utilizes null convolutions to extend the perceptual field, while PsPnet [34] adopts an encoder–decoder structure with pyramidal pooling. Furthermore, Icnet [35] integrates the outputs of different feature maps to generate dense segmentation results. In our research, we leverage the standard encoder–decoder architecture of MixVisionTransformer [17] as a foundation. Given our objective of utilizing CNN for local information extraction and Transformer for global information acquisition, this framework aligns well with our approach. To aggregate information from multiple levels within the CNN and Transformer modules, we propose the Fusion of Local and Global Information mechanism (FLGM) module. This module enhances the interplay of features across all levels, facilitating the fusion of high-level semantic information and low-level local information. By doing so, our network benefits from improved feature aggregation and the extraction of heterogeneous information.

3. Methods and Motivation

In this section, we elucidate the underlying motivation that drives our research endeavors, accompanied by the presentation of graphical representations depicting experimental results. These results serve to substantiate the viability and efficacy of our study. Furthermore, we introduce the pivotal components of FURSformer as follows: (1) the Transformer branch, which facilitates the acquisition of global information, (2) the convolution branch, which enables the capture of local information, and (3) the hierarchical aggregation and Feature-Level Global–Local Fusion (FLGM) module.

3.1. Research Motivation

To illustrate the need for this work, we conducted a series of experiments to investigate the limitations of CNN or Transformer models as encoders in RSI.

Pilot research: We present a visual analysis in Figure 3 highlighting several instances of unsuccessful semantic segmentation achieved via CNN networks and Transformer networks. These instances serve as illustrative cases to examine and evaluate the merits and limitations of CNNs and Transformer models. Unique images, meticulously extracted from the DLRSD [27] dataset, including images featuring small objects (image c), large object images (image a), and images encompassing both small and large objects (Figure 3b,d), are utilized to systematically evaluate the robustness and generalization capabilities of the models. This comprehensive analysis aims to establish the expertise exhibited by CNNs and Transformer models in their respective domains. Moreover, considering the unique characteristics of remote sensing images, which necessitate the segmentation of images characterized by intricate boundaries and delicate objects, the ability of the model to comprehend and effectively utilize global information becomes imperative.

Analysis: The primary distinction between the general-purpose domain and the remote sensing domain lies in the challenges associated with extracting accurate feature information from objects within remote sensing images. Factors such as weather conditions, shadows, object resolution, noise, and more contribute to this difficulty. The introduction of the Vision Transformer [15] model has demonstrated its ability to reliably capture remote correlations within a global context, enabling categories to learn approximate features from one another and yield improved performance. In this context, we address three significant issues.

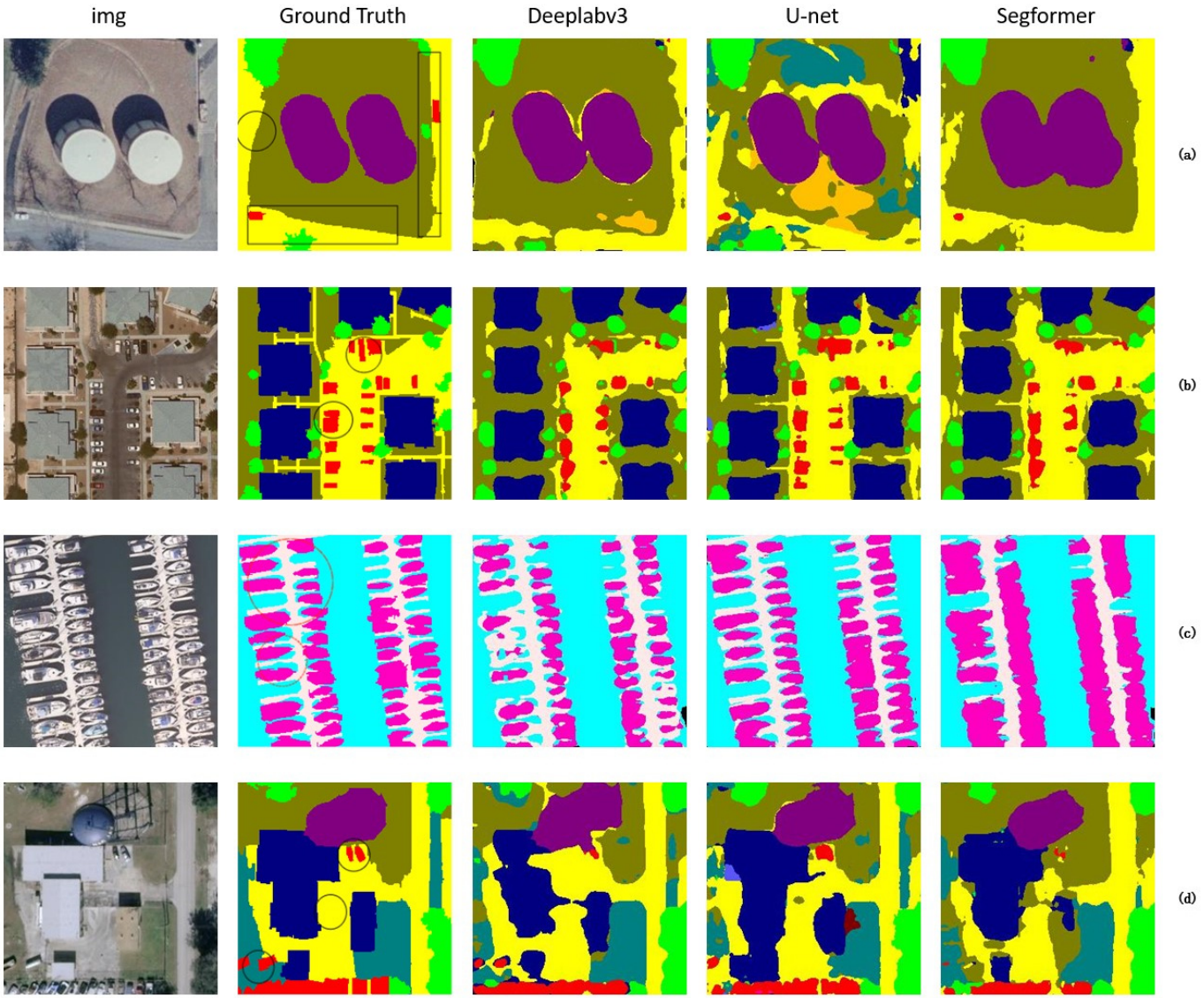

Figure 3. The semantic segmentation comparison chart between Transformer and CNN models (DLSRD [27] as an example). Each row represents the image, ground texture, and comparison model from left to right. (**a**–**d**) represents each image. The red and black circles drawn in the ground truth indicate the special areas that the CNN and Transformer models need to focus on.

Firstly, the Transformer model exhibits limitations in its ability to effectively learn local modeling capabilities, resulting in the loss of texture details during patch interactions. Secondly, due to the substantial variation in element scales within remote sensing images, the incorporation of multi-scale jump connections becomes crucial for robust performance. Additionally, the fixed nature of the Transformer's patch limits its ability to capture similar multi-scale information properties, thereby potentially hindering its performance.

Furthermore, despite the utilization of multi-scale feature fusion in U-net, the global information learned by the model remains coarse, primarily due to the limited perceptual receptive field provided by the CNN model. This limitation becomes evident in Figure 3a of the U-net model, where the confusion caused by direct concatenation during multi-scale fusion leads to simple downsampling, resulting in the model failing to effectively learn features in specific regions. Consequently, the output of the model exhibits confusion in representing these feature areas.

Lastly, when it comes to pixel-level grasping, such as in small pixel objects shown in Figure 3b,c, the regions obtained by the Transformer are not as refined as those obtained

by the CNN. This discrepancy leads to unclear boundaries and diminished detail. Further visual details can be found in Figure 3, we provide explicit visual distinctions between the concerns of the Transformer and CNN networks.

3.2. Transformer and CNN

To solve the above problem, we integrate the advantages of CNN and Transformer, and we borrow the MixVisionTransformer [18] module as the global information extraction branch of the encoder, use the designed CNN branch as the local information extraction branch of the encoder, and add the channel attention mechanism to complement the Transformer's focus on the channel dimension.

The structure diagram of Transformer is shown in Figure 4. The input is X, and the input is passed through the downsampling module, which is noted as Down_l, specifically our input $X \in B \times C \times H \times W$, after a convolutional layer to obtain $X'_{l-1} \in B \times C_{l-1} \times \frac{H}{2^{i+1}} \times \frac{w}{2^{i+1}}, i \in \{1, 2, 3, 4\}$. Then, it goes through a layer of normalization. Therefore, the process of downsampling module L-layer is expressed as follows:

$$X'_{l-1} = \mathrm{Down}_l(X) \tag{1}$$

$$X_{l-1} = \mathrm{LN}(X'_{l-1}) \tag{2}$$

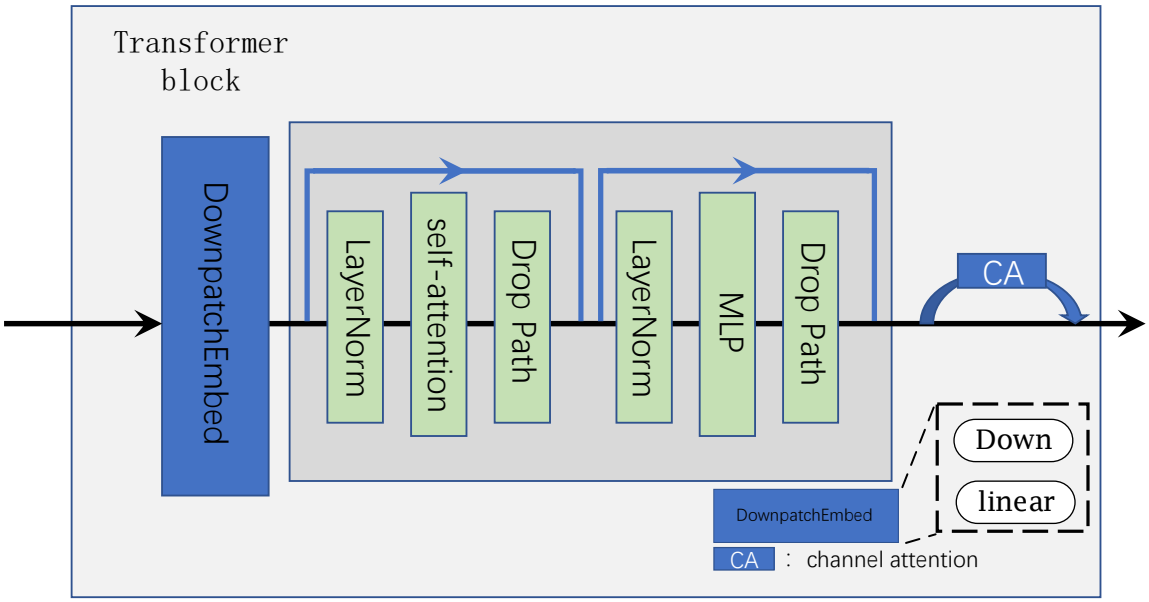

Figure 4. Transformer block architecture. The input is X, and after downsampling and patch embedding, the size of the feature map is reduced while increasing the dimension of the feature map. Layer normalization [36] and Drop path [37] improve network generalization performance. Finally, we add a channel attention [38] mechanism to the obtained output to capture the characteristics of the channel dimension.

By splitting the feature map X into non-overlapping patches with patch embedding and applying a linear layer to the input layer X projects the features to an arbitrary dimension. We will write the input of the first Transformer layer as Z_0. After that, the L-layer Transformer module is applied to extract the features. Specifically, each Transformer module consists of a multi-head self-attention (MSA) module and a multi-layer perceptron (MLP). The layer norm is used in front of each MSA and MLP, and a regularized drop path [37] is used to enhance the model generalization. Residual connections are used in

each module to enhance learning. Therefore, the process of the Transformer module L-layer is expressed as follows:

$$Z'_l = \mathrm{Drop}\left(\mathrm{MSA}\left(\mathrm{LN}\left(Z'_{l-1}\right)\right)\right) + Z'_{l-1} \tag{3}$$

$$Z_l = \mathrm{Drop}\left(\mathrm{MLP}\left(\mathrm{LN}\left(Z'_l\right)\right)\right) + Z'_l \tag{4}$$

And Z'_l and Z_l are the output features of MSA and MLP of L-layer.

For the CNN branch (as shown in Figure 5), we have built a local information extraction module to extract local information from the Transformer module near the input, which has the advantage of reducing parameters while obtaining useful local information. As in [39], they have too many parameters and use the standard CNN + Transformer architecture. Instead, we extracted valid local information from the CNN branch with a small number of parameters.

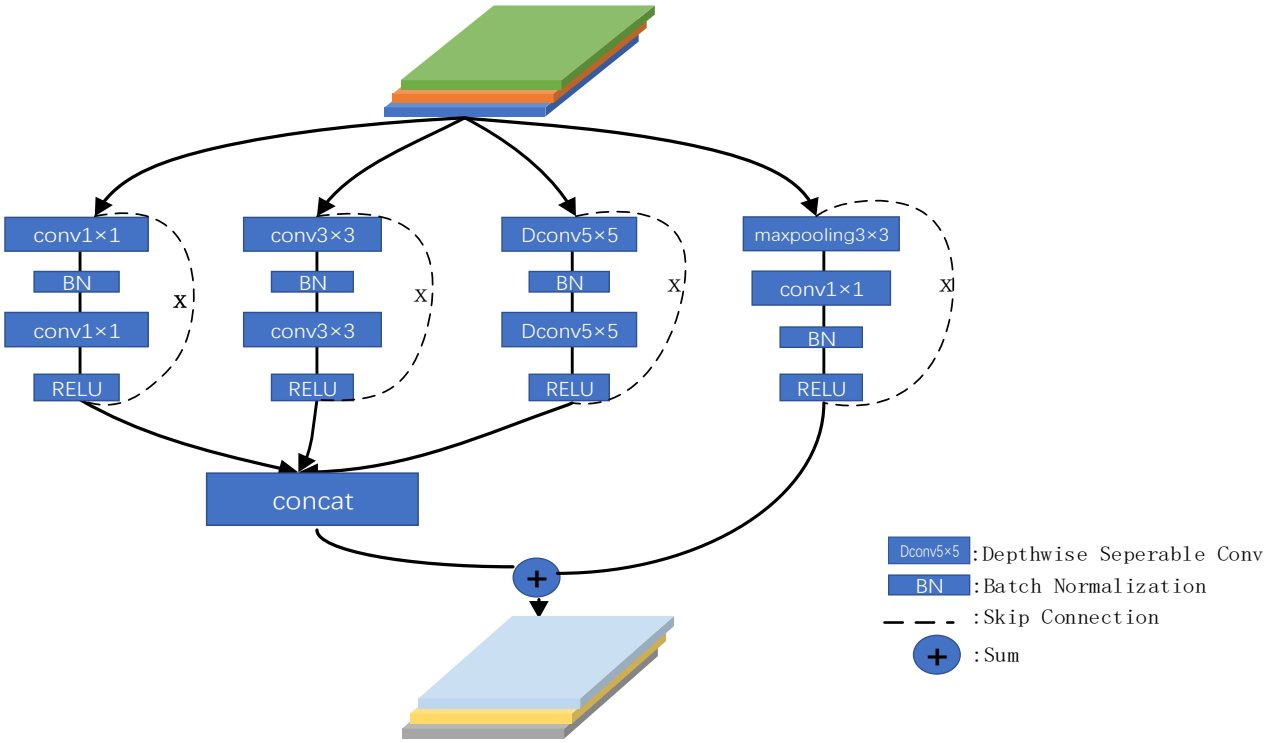

Figure 5. CNN block architecture. The input feature X first obtains local features via large kernel convolution and small kernel convolution, and then, the obtained features are concated. Finally, the concated features are combined with the enhanced features as outputs.

Specifically, our input is x. In the CNN block, a 1 × 1 convolution, a 3 × 3 convolution with padding = 1, stride = 1, and a 5 × 5 Depthwise Separable Convolution [40] were used to learn detailing information while reducing parameters, and finally, a maxpooling 3 × 3 with padding = 1 and stride = 1 was used to reinforce edge features and detailing features, and the residual connection is used to mitigate the vanishing gradient. We will loop the above feature extraction module L times so that the CNN module can learn the features of shallow information better.

3.3. FLGM

To alleviate the problem of insufficient fusion, we propose FLGM (in Figure 6), which enhances the global information of the Transformer and the local information of the CNN module to model the affinity of the two types of information in a set-to-set manner.

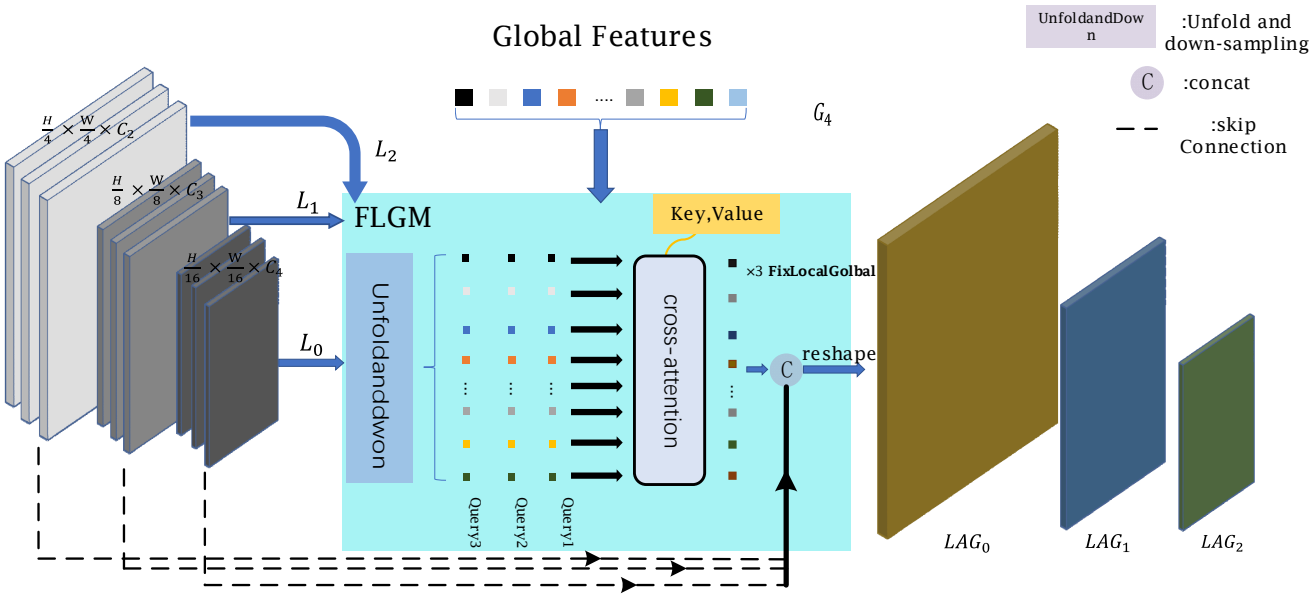

Figure 6. FLGM block architecture. The final mixed information feature was obtained via our FLGM module, which contains rich global and detailed local information.

We believe that simply concatenating global information with local information does not yield good results. Therefore, we construct the fusion of heterogeneous information using cross-attention. This approach has the advantage of fusing the information of the CNN module and the Transformer module as heterogeneous information, which can fully reconstruct the correlation between the two.

Specifically, we use the Transformer module to obtain the feature $G = \{G_4\}$ as a global feature, $G^R \in B \times C_4 \times (\frac{H}{2^5} \times \frac{w}{2^5})$, the feature has a wealth of global information. The local feature as $L = \{L_0, L_1, L_2\}$ is extracted using the CNN module to obtain a shallow feature map representation using $L^R \in B \times C \times \frac{H}{2^{i+1}} \times \frac{w}{2^{i+1}}$, $i \in \{2, 3, 4\}$. We first use MLP to project the local features to the same dimension as the global features, denoted as C_4, then reshape the local features into a two-dimensional matrix $X = \{X_0, X_1, X_2\}$, $X \in B \times C_4 \times (H_i \times W_i) I = \{0, 1, 2\}$. After that, cross-attention calculations are carried out using X and G, respectively, to construct hybrid feature maps. Here, we take the example of L_0, G_4, Q (query), K (key), and V (value) are calculated using linear layer projection with the following equations:

$$Q = P_Q(L_0) \tag{5}$$

$$K = P_K(G_4) \tag{6}$$

$$V = P_V(G_4) \tag{7}$$

where P_Q, P_K, and P_V are linear projections, respectively.

$$\text{Attention}(Q, K, V) = \text{softMax}\left(\frac{QK^T}{d_k}\right)V \tag{8}$$

$$\text{CrossAttention}(Q, K, V) = \text{Attention}\left(QK^T, K, V\right) \tag{9}$$

Query vector, Q, is first modeled with key vector, K, for attention score. After that, we divide by the dimension of multi-head attention and use multi-head attention in order to focus on different refinement features for each head again. Then, the attention weight $\text{softMax}(\frac{QK^T}{\sqrt{d_k}})$ is obtained using softMax. The final attention weights are calculated,

mapped to the corresponding dimensions, and connected by the residuals as the output. The FLGM module pseudo-code is shown below (Algorithm 1).

Algorithm 1 FLGM

```
#Description: cross: MutiHeadCrossAttention
#Input: local_c is the local information that passes through the cnn branch
#Input: G is the global information that passes through the transformer branch
#Output: FixF is the local and global information fixed

local_c = cnn(Loc_inf) #local information through cnn to get

local_c = self.linear_c(local_c).permute(0,2,1).reshape(n,1, local_c.shape[2], local_c.shape[3])

Fix = rearrange(self.cross(local_c, G),'b (h w) c->b c h w', h = local_c.size()[2]) #Repeat L time cross

FixF = torch.cat([Fix, local_c],dim = 2)
return FixF, G
```

The local_c are obtained using CNN local feature extraction module. After the operation of linear projection will be consistent with the G dimension, and then, it will be modeled with the global feature G for affinity, and finally, concat mix features and global features will be used as our multi-scale output features. In this way, we make the feature maps at each scale fuse global and local information via multi-scale cross-fusion, which facilitates us to learn the feature maps outputted by different branches. Finally, the final output is obtained by upsampling the output in a hierarchical manner during the decoder.

4. Experiment

4.1. The Dataset

DLRSD [27] is a densely labeled dataset that can be used for multi-labeling tasks, such as remote sensing image retrieval (RSIR) and classification, as well as other pixel-based tasks, such as semantic segmentation (also known as remote sensing classification). DLRSD has 21 large categories with 100 images each, the same as the UC Merced archive. We labeled the pixels of each image in the UC Merced archive with the following 17 class labels: aircraft, bare soil, buildings, cars, chaparral court, docks, fields, grass, mobile homes, sidewalks, sand, sea, boats, tanks, trees, and water. The 17 class labels were first constructed and defined in the multi-label RSIR archive, where UC Each image in the Merced archive is provided with a set of multiple labels. The dataset image size is $3 \times 256 \times 256$. The images in this dataset have many categories and scenes, giving us complex scene variations.

The LoveDA [41] dataset is constructed from 0.3 m images, collected in Nanjing, Wuhan, and Changzhou, China, with inconsistent urban-rural ratios due to each study area having its own planning strategy. The dataset was collected for both rural and urban areas referencing the urban-rural zoning codes published by the National Bureau of Statistics of China. Nine densely populated urban areas were selected from economically developed areas, and the other nine rural areas were selected from undeveloped prefectures. Each image is a 1024×1024 pixels image. The dataset has a total of seven categories: (1) background, (2) building, (3) road, (4) water, (5) barren, (6) forest, and (7) agriculture. Here, we have used the urban part of this dataset because there are label plots in the rural dataset that are not correctly classified.

4.2. Evaluation Indicators

Most of the evaluation metrics in semantic segmentation are based on accuracy assessment, usually based on a confusion matrix. In this paper, we have used mPA, mIoU, Recall, mAccuracy, and Precision for evaluating the performance of the proposed method. The confusion matrix is shown in Table 1.

Table 1. The confusion matrix consists of rows and columns for each of the values. The first one indicates whether it is correct or not, and the second one indicates the predicted result. It is represented by P_{ii}. In the matrix, true positives (TP) and true negatives (TN). The label marked as class i is incorrectly predicted as class j, denoted by P_{ij}, which is the false positive (FP) and false negative (FN) in the matrix.

	Prediction Results	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

4.2.1. mPA (Mean Pixel Accuracy)

mPA is a variation of PA (Pixel Accuracy). PA is the proportion of correctly labeled pixels to the total pixels. And mPA is the calculation of the proportion of the number of pixels per class that are correctly classified, after which all classes are averaged as follows:

$$\mathrm{PA} = \frac{\sum_{i=0}^{K} p_{ii}}{\sum_{i=0}^{K} \sum_{j=0}^{K} p_{ij}} \tag{10}$$

$$\mathrm{mPA} = \frac{1}{K+1} \sum_{i=0}^{K} \frac{p_{ii}}{\sum_{j=0}^{K} p_{ij} + \sum_{j=0}^{K} p_{ji} - p_{ii}} \tag{11}$$

4.2.2. mIoU (Mean Intersection over Union)

mIoU is a standard metric for semantic segmentation, which calculates the ratio of the intersection and the concatenation of two sets, which in the problem of semantic segmentation are the true value (ground truth) and the predicted value (predicted segmentation). This ratio can be morphed into the sum of the true positive, false negative, and false positive (concatenated sets) over the positive truth (intersection) ratio. The IoU is calculated on each class and averaged afterward.

$$\mathrm{mIoU} = \frac{1}{K+1} \sum_{i=0}^{K} \frac{p_{ii}}{\sum_{j=0}^{K} p_{ij} + \sum_{j=0}^{K} p_{ji} - p_{ii}} \tag{12}$$

4.2.3. Recall

Recall expresses how many of the actual positive samples the classifier was able to predict. It expresses the ratio of the number of samples correctly identified by the model as positive classes to the total number of positive samples. In general, the higher the Recall, the more positive samples are correctly predicted by the model, and the better the model is. It is for the original positive samples.

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{13}$$

4.2.4. Precision

Precision indicates the proportion of samples identified by the model as positive classes that are actually positive.

$$\mathrm{Precison} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{14}$$

4.2.5. mAccuracy

mAccuracy is the ratio of correctly predicted data to the total data.

$$\mathrm{mAccuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{15}$$

4.3. Implementation Detailed

In this section, we describe the model environment settings, hyperparameter settings, and the results of comparison tests.

Experimental Environment and Parameters

The comparison models in this paper (U-net, DeepLabV3, Segformer) all use the pre-training weights from the voc2012 dataset, and during the model training process, the total is 200 epochs, and the freeze epoch is 100. The meaning of the Freeze epoch is that in this epoch, we will not train the Transformer backbone network, nor will we perform backpropagation to update parameters. The AdamW [42] optimizer is used, and the initial learning rate is 1×10^{-4}. The value of the decal is set to 0.99. The specific initialization settings of the hyperparameters regarding the comparison model are shown in Table 2. For the data set, we set the ratio of the validation set to the training set to 1:9.

Table 2. The table is the initialized hyperparameter table for the comparison model used. We used the above initialization in all the next comparison experiments.

Model	Freeze Epoch	Epoch	Batch Size	Optimizer Tpye	Momentum	Weight Decay	Init lr/min lr	Lr Decay Type	FLOPs	Params
U-net	100	200	16	Sgd	0.9	1×10^{-4}	$7 \times 10^{-3}/7 \times 10^{-3} \times 0.01$	cos	56.52 G	24.89 M
DeepLabV3	100	200	16	Sgd	0.9	1×10^{-4}	$7 \times 10^{-3}/7 \times 10^{-3} \times 0.01$	cos	112.87 G	23.71 M
Segformer	100	200	16	AdamW	0.9	1×10^{-2}	$1 \times 10^{-5}/1 \times 10^{-5} \times 0.01$	cos	28.38 G	27.35 M
$FURSformer_{ours}$	100	200	16	AdamW	0.9	1×10^{-2}	$1 \times 10^{-5}/1 \times 10^{-5} \times 0.01$	cos	21.44 G	20.76 M

The hyperparameter initialization using Table 2 in the DLRSD dataset was compared with our model. The results of the evaluation are presented in Table 3, and the qualitative results are illustrated in. Our model is ahead of the baseline model we compared in the commonly used evaluation metrics. Specifically in the CNN model, the U-net model is higher than the DeepLabV3 model, probably because U-net uses multi-scale fusion and is designed for medical images. In terms of mIoU, the U-net model showed a 1.2% higher performance than the DeepLabV3 model, indicating that the predicted images differed less from the ground truth than the DeepLabV3's. Moreover, in terms of Mprecision, which represents the proportion of pixels that are actually correct in all predictions, the U-net model showed the second highest percentage, surpassing the DeepLabV3's. This could be attributed to the U-net's network architecture, which is more efficient in excluding noise and accurately identifying the small building, car, and ship boundaries in remote sensing images. It is worth noting that due to the unbalanced foreground and background of remote sensing images, mPA, mIoU, and mAccuracy should be analyzed together. In summary, there are no major fluctuations in the performance metrics of the CNN model.

Table 3. Experimental results obtained for different models on the DLRSD dataset.

Model	mIoU	mPA	mAccuracy	MRecall	mPrecision
U-net	71.59%	83.36%	86.16%	82.45%	83.80%
DeepLabV3	70.40%	83.36%	86.59%	83.36%	81.28%
Segformer	72.97%	85.63%	89.68%	85.63%	82.62%
$FURSformer_{ours}$	75.32%	86.04%	90.78%	86.04%	85.11%

As a representative semantic segmentation model of Transformer architecture, Segformer is frequently utilized as a baseline model for comparison in various studies. Our team has conducted an experimental evaluation where Segformer was employed as a baseline model for comparison with our proposed method. Segformer accuracy is far ahead; compared with U-net's mIoU, Segformer accuracy has increased by 1.3%, and mPA compared to DeepLabV3 likewise improved by 2.57%, benefiting from self-attention, which puts the images into the Transformer module as patches, weighting the different positions of the sequences, thus better enabling the ability to model sequences and providing correlation

between each patch throughout. For CNN, it is not possible to see the global correlation of the image from one convolution operator. Secondly, the Transformer architecture can better focus on the relationship between a region and another region, obtain arbitrary information from any sequence of the model, and give different weights at different times reasonably. Furthermore, the multi-head attention mechanism enables the model to autonomously focus on significant features while allowing each attention head to learn different features. CNN models are not as generalizable and flexible as Transformer modules.

Our experimental evaluations demonstrate that our proposed method outperforms Segformer using CNN branches augmented with a channel attention module, along with the integration of heterogeneous information from both branches to achieve an enhanced feature map representation. Specifically, our model achieves superior performance compared to Segformer as evidenced by a 2.35% improvement in terms of mIoU accuracy and a higher mPrecision than the U-net model, indicating its effective learning of local information. Furthermore, our approach demonstrates an exceptional ability to handle boundary refinement tasks involving cars, ships, and other objects, exploiting the strengths of both Transformer and CNN models to learn global and local information for precise pixel classification and boundary delineation. Our experimental results are presented in (Tables 4 and 5), while Figure 7 visualizes the outcomes of our approach.

Quantitative results presented in Tables 4 and 5 demonstrate that our proposed method achieves results closer to the CNN model in cases where CNN models perform well, especially in recognition of small object boundaries and detailed object awareness. This improvement can be attributed to our CNN branching and fusion module. In cases where the Segformer model exhibits high accuracy, and the CNN model shows low accuracy, our proposed method achieves similar or even superior accuracy compared to the Segformer model. This finding highlights the advantage of our approach, which not only leverages the strengths of the Transformer model for acquiring global information but also adds value via effective boundary feature extraction for the target class.

In Figure 7, we can see that the FURSformer is less influenced by noise than the Segformer, and our model refinement features are learned more abundantly. Compared to the CNN model, our model has a more expert understanding of the global picture and is better able to distinguish between category features. Figure 7c shows that the car category FURSformer model learns the features very well, while Segformer finds almost nothing for the smaller pixel categories. In terms of building construction, our model obtains a more significant effect closer to ground truth than Segformer. Figure 7f–h illustrate that our model learns details that the Segformer does not, i.e., rich local information. It can be seen that the prediction maps obtained by both the detailed object texture and the border FURSformer are better than the Segformer, and the complete prediction maps are also better than the CNN, which is due to the fact that we fuse the two kinds of information together.

Figure 7 provides a visual comparison of the performance of our proposed FURSformer model with that of Segformer. Our findings indicate that FURSformer is less susceptible to noise and demonstrates a more abundant feature refinement compared to Segformer. Furthermore, our model exhibits a superior understanding of the global picture and is better equipped to distinguish between category features compared to the CNN model. In particular, Figure 7c highlights that our FURSformer model effectively learns the features of the car category, while Segformer struggles to identify smaller pixel categories. In terms of building construction, our model achieves a more significant effect closer to ground truth than Segformer, as illustrated by Figure 7f–h). Our proposed method leverages the FLGM module to effectively integrate the strengths of both global and local information, resulting in a rich feature representation that enhances the accuracy of pixel classification and boundary delineation.

Table 4. Experimental results of IOU for each classification of DLRSD dataset. Bold represents the highest value of IOU for each classification in our experiment.

DLRSD IOU																	
Category	**Water**	**Trees**	**Tanks**	**Ships**	**Sea**	**Sand**	**Pavement**	**Mobile Home**	**Grass**	**Filed**	**Dock**	**Court**	**Chaparral**	**Cars**	**Buildings**	**Bare Soil**	**Airplane**
DeepLabV3	0.80	0.72	0.77	0.71	0.91	0.65	0.78	0.63	0.55	0.96	0.53	0.71	0.65	0.65	0.69	0.58	0.68
U-net	0.78	0.74	**0.82**	**0.78**	**0.94**	0.58	0.76	0.64	0.54	0.94	**0.60**	0.76	**0.71**	**0.70**	0.65	0.54	0.70
Segformer	0.81	0.77	0.76	0.72	0.92	0.69	0.81	0.65	0.62	0.97	0.51	0.82	0.65	0.65	0.76	**0.61**	0.67
$FURSformer_{ours}$	**0.83**	**0.79**	**0.84**	0.75	0.92	**0.73**	**0.83**	**0.70**	**0.64**	**0.98**	0.58	**0.83**	0.58	0.68	**0.79**	**0.61**	**0.73**

Table 5. Experimental results of PA for each classification of DLRSD dataset. Bold represents the highest value of PA for each classification in our experiment.

DLRSD PA																	
Category	**Water**	**Trees**	**Tanks**	**Ships**	**Sea**	**Sand**	**Pavement**	**Mobile Home**	**Grass**	**Filed**	**Dock**	**Court**	**Chaparral**	**Cars**	**Buildings**	**Bare Soil**	**Airplane**
DeepLabV3	0.88	0.84	0.94	0.79	**1.00**	0.76	0.86	0.78	0.73	0.96	0.71	0.96	**0.82**	0.79	0.77	0.75	0.84
U-net	0.86	**0.89**	0.91	0.84	0.99	0.68	0.89	0.83	0.63	0.95	**0.74**	0.93	0.78	**0.83**	0.76	0.72	0.77
Segformer	0.90	**0.89**	**0.95**	**0.90**	**1.00**	**0.77**	0.89	**0.87**	0.74	0.97	0.64	**0.97**	**0.82**	0.78	0.87	0.74	**0.87**
$FURSformer_{ours}$	**0.91**	**0.89**	**0.95**	0.88	0.99	**0.76**	**0.91**	0.86	**0.77**	**0.98**	0.73	0.96	0.72	0.80	**0.88**	**0.76**	**0.87**

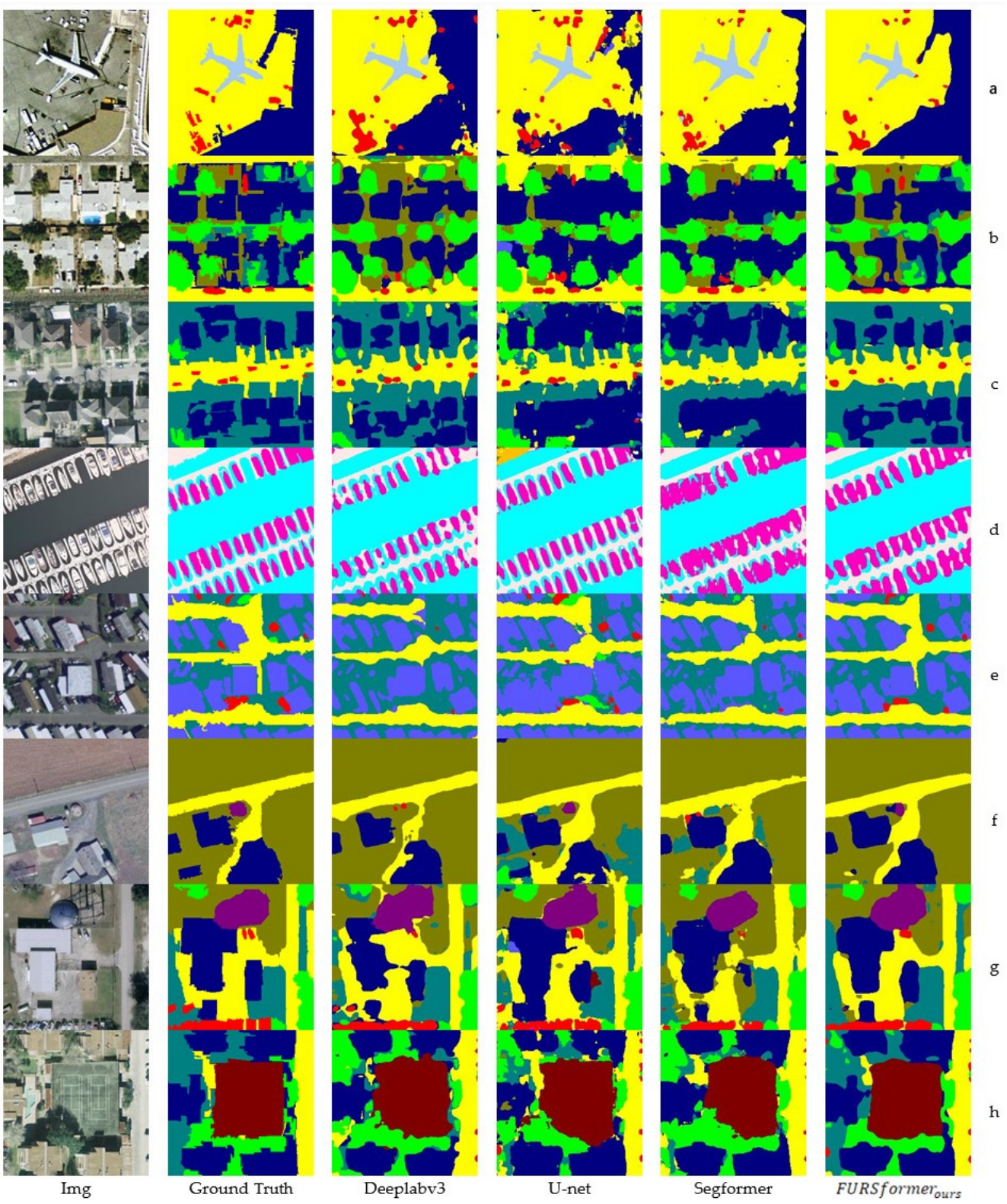

Figure 7. Comparison graph. We chose representative images to demonstrate that our model combines the advantages of CNN and Transformer models. (**a**–**h**) Each row, from left to right, represents the true image, ground truth, and the predicted results of the model.

In the dataset for training LoveDA-Urban, we only trained 40 epochs, and the freeze epoch was set to 20 epochs. We used the pre-trained model obtained from the DLSRD dataset to fine-tune the LoveDA dataset. In terms of lr, we have lowered it by an order of

magnitude on the basis of Table 2. The quantitative results are presented in Tables 6 and 7, and our qualitative results are shown in Figure 8. The table shows that CNN cannot be compared to Transformer. After comparing the differences between these two datasets, we believed that the LoveDA dataset has too few categories, such as confusing roads and cars. This resulted in CNN not achieving high performance in certain categories of the LoveDA dataset. As shown in Figure 8, with the respective advantages of the Transformer and CNN, our model has a better representation of classification features and a more accurate classification of edge features.

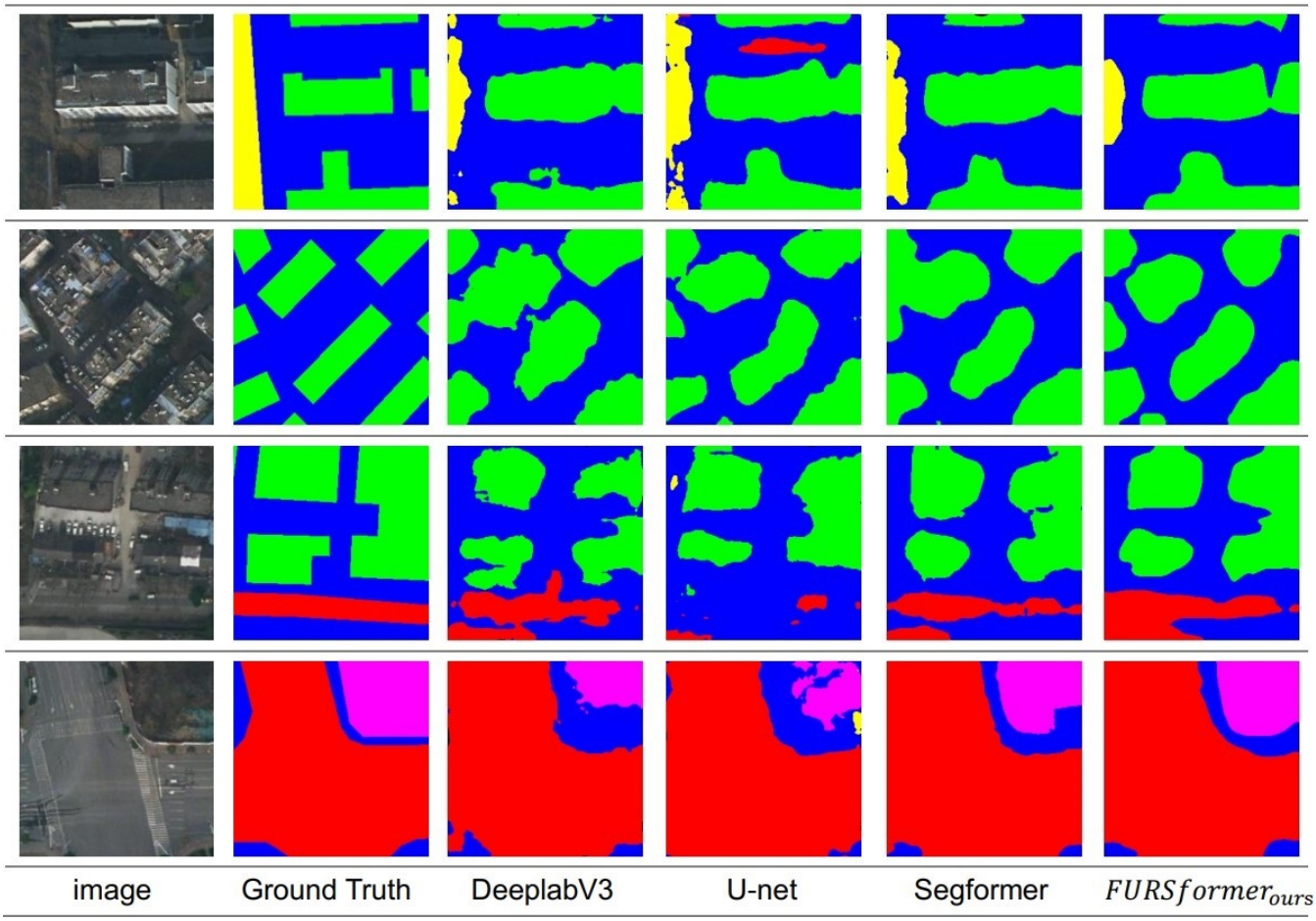

Figure 8. Comparison chart. This figure illustrates the Transformer and CNN models running on the LoveDA-Urban dataset.

Table 6. Experimental results obtained for different models on the LoveDA-Urban dataset.

	LoveDA-Urban				
	mIoU	**mPA**	**mAccuracy**	**MRecall**	**mPrecision**
Deeplabv3	66.20%	77.29%	77.13%	77.29%	80.24%
U-net	65.38%	78.48%	76.02%	78.48%	77.62%
Segformer	67.75%	80.06%	77.71%	80.06%	79.81%
$FURSformer_{ours}$	69.94%	80.55%	79.57%	80.55%	82.62%

Table 7. Experimental results of IOU for each classification of loveDA-Urban dataset. Bold represents the optimal result.

	Agriculture	Forest	Barren	Water	Road	Building	Background
Deeplabv3	0.55	0.52	0.53	0.84	0.62	0.60	0.64
U-net	0.56	0.51	0.48	0.84	0.62	0.61	0.61
Segformer	0.59	0.53	0.56	0.86	0.64	0.61	0.64
$FURSformer_{ours}$	**0.62**	**0.56**	**0.58**	**0.88**	**0.66**	**0.63**	**0.67**

5. Conclusions

We propose a simple and effective semantic segmentation method (FURSformer) based on deep-learning remote sensing images, which extracts global information via the Transformer branch and local information using a designed CNN branch. We used the normal Transformer module and downsampling to increase the channel information and reduce the feature map size. In the CNN module, we explored many SOTA models and simply built the CNN module to obtain local features. Since the later fusion branch leads to under-fusion of encoder feature aggregation, we propose a fusion heterogeneous information module (FLGM), which helps us to well aggregate local and global information and enhance the consistency of the network model with two branches. Finally, our proposed network achieves good results on several remote sensing datasets, and the FLGM module can be well used for the fusion of the Transformer model and CNN.

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

Funding: This research received no external funding.

Data Availability Statement: We used the publicly available dataset DLRSD provided by Shao, Z.; Yang, K.; Zhou, W.; The DLRSD dataset is freely available at https://sites.google.com/view/zhouwx/dataset#h.p_hQS2jYeaFpV0 (accessed on 15 April 2023).

Conflicts of Interest: The authors declare that there is no conflict of interest regarding the publication of this paper.

References

1. Liu, X.; Cui, B.; Yan, S. Seed region growing based on gradient vector flow for medical image segmentation. *Comput. Med. Imaging Graph.* **2008**, *32*, 124–131.
2. Oka, M.; Roy, R. Effective Image Segmentation using Fuzzy C Means Clustering with Morphological Processing. *Int. J. Image Graph. Signal Process.* **2012**, *4*, 49–56.
3. Basse, R.M.; Omrani, H.; Charif, O.; Gerber, P.; Bódis, K. Land use changes modelling using advanced methods: Cellular automata and artificial neural networks. The spatial and explicit representation of land cover dynamics at the cross-border region scale. *Appl. Geogr.* **2014**, *53*, 160–171. [CrossRef]
4. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
5. Deng, J.; Wei, D.; Socher, R.; Li, L.-J.; Li, K.; Li, F.-F. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 248–255.
6. Jonathan, L.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Santiago, Chile, 7–13 December 2015.
7. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015, Proceedings, Part III 18*; Springer International Publishing: Berlin/Heidelberg, Germany, 2015.
8. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
9. 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]

10. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
11. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
12. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
13. Liu, S.; Huang, D. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
14. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
15. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
16. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
17. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30, Proceedings of the 31th International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017*; Curran Associates Inc.: Red Hook, NY, USA, 2017.
19. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
20. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020, Proceedings, Part I 16*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
21. Guo, M.-H.; Lu, C.-Z.; Liu, Z.-N.; Cheng, M.-M.; Hu, S.-M. Visual attention network. *arXiv* **2022**, arXiv:2202.09741.
22. Xu, D.; Alameda-Pineda, X.; Ouyang, W.; Ricci, E.; Wang, X.; Sebe, N. Probabilistic graph attention network with conditional kernels for pixel-wise prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 2673–2688. [CrossRef]
23. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **2003**, *20*, 273–297. [CrossRef]
24. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
25. Volpi, M.; Tuia, D. Deep multi-task learning for a geographically-regularized semantic segmentation of aerial images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *144*, 48–60. [CrossRef]
26. Marmanis, D.; Wegner, J.D.; Galliani, S.; Schindler, K.; Datcu, M.; Stilla, U. Semantic segmentation of aerial images with an ensemble of CNSS. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *3*, 473–480. [CrossRef]
27. Shao, Z.; Yang, K.; Zhou, W. Performance Evaluation of Single-Label and Multi-Label Remote Sensing Image Retrieval Using a Dense Labeling Dataset. *Remote Sens.* **2018**, *10*, 964. [CrossRef]
28. Waldner, F.; Diakogiannis, F.I. Deep learning on edge: Extracting field boundaries from satellite images with a convolutional neural network. *Remote Sens. Environ.* **2020**, *245*, 111741. [CrossRef]
29. Ding, L.; Zhang, J.; Bruzzone, L. Semantic segmentation of large-size VHR remote sensing images using a two-stage multiscale training architecture. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 5367–5376. [CrossRef]
30. Liu, R.; Li, M.; Chen, Z. AFNet: Adaptive fusion network for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 7871–7886. [CrossRef]
31. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022. [CrossRef]
32. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
33. Heraa, S.; Gong, J. Transformer-Based Deep Learning Model for SAR Image Segmentation. In Proceedings of the 2021 International Conference on Intelligent Transportation, Big Data & Smart City, Xi'an, China, 27–28 March 2021; pp. 166–173.
34. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
35. Zhao, H.; Qi, X.; Shen, X.; Shi, J.; Jia, J. Icnet for real-time semantic segmentation on high-resolution images. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
36. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
37. Ghiasi, G.; Lin, T.-Y.; Le, Q.V. Dropblock: A regularization method for convolutional networks. In *Advances in Neural Information Processing Systems 31, Proceedings of the 32th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018*; Curran Associates Inc.: Red Hook, NY, USA, 2018.
38. Bastidas, A.A.; Tang, H. Channel attention networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
39. Li, Z.; Chen, Z.; Liu, X.; Jiang, J. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *arXiv* **2022**, arXiv:2203.14211.

40. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
41. 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.
42. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.