mathematics

MDPI

Article

HA-RoadFormer: Hybrid Attention Transformer with Multi-Branch for Large-Scale High-Resolution Dense Road Segmentation

Zheng Zhang [1], Chunle Miao [1], Changan Liu [1], Qing Tian [1,*] and Yongsheng Zhou [2]

[1] School of Information, North China University of Technology, Beijing 100144, China; zhangzheng@ncut.edu.cn (Z.Z.); chunle@mail.ncut.edu.cn (C.M.); furk0416@mail.ncut.edu.cn (C.L.)
[2] College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China; zhyosh@mail.buct.edu.cn
* Correspondence: tianqing@ncut.edu.cn

Abstract: Road segmentation is one of the essential tasks in remote sensing. Large-scale high-resolution remote sensing images originally have larger pixel sizes than natural images, while the existing models based on Transformer have the high computational cost of square complexity, leading to more extended model training and inference time. Inspired by the long text Transformer model, this paper proposes a novel hybrid attention mechanism to improve the inference speed of the model. By calculating several diagonals and random blocks of the attention matrix, hybrid attention achieves linear time complexity in the token sequence. Using the superposition of adjacent and random attention, hybrid attention introduces the inductive bias similar to convolutional neural networks (CNNs) and retains the ability to acquire long-distance dependence. In addition, the dense road segmentation result of remote sensing image still has the problem of insufficient continuity. However, multiscale feature representation is an effective means in the network based on CNNs. Inspired by this, we propose a multi-scale patch embedding module, which divides images by patches with different scales to obtain coarse-to-fine feature representations. Experiments on the Massachusetts dataset show that the proposed HA-RoadFormer could effectively preserve the integrity of the road segmentation results, achieving a higher Intersection over Union (IoU) 67.36% of road segmentation compared to other state-of-the-art (SOTA) methods. At the same time, the inference speed has also been greatly improved compared with other Transformer based models.

Keywords: dense road segmentation; transformer; multiscale patches; hybrid-attention

MSC: 68T01

Citation: Zhang, Z.; Miao, C.; Liu, C.; Tian, Q.; Zhou, Y. HA-RoadFormer: Hybrid Attention Transformer with Multi-Branch for Large-Scale High-Resolution Dense Road Segmentation. *Mathematics* **2022**, *10*, 1915. https://doi.org/10.3390/math10111915

Academic Editor: Radu Tudor Ionescu

Received: 6 May 2022
Accepted: 1 June 2022
Published: 2 June 2022

1. Introduction

Dense road segmentation based on high-resolution remote sensing images plays a vital role in urban planning, traffic management, vehicle navigation, and map updating [1–3]. With the development of deep learning, the task of road segmentation has made significant progress. It is essential to obtain multi-scale features for intensive detection tasks, such as object detection and semantic segmentation in computer vision. The road segmentation network based on CNNs achieves superior performance in intensive visual tasks through the multi-size of the convolution kernel and multiscale on the feature map. For example, Inception [4] utilizes the split-transform-merge strategy to fuse the features with different sizes of receptive fields generated by multiscale convolution kernels (such as 3×3, 5×5, and 7×7) and obtains a variety of receptive fields on the same level of the feature map, which is conducive to identifying objects with different scales. HRNet [5] uses multiple resolution branches to obtain fine and coarse features and constructs multiscale features through information interaction between different levels of features.

Although the model based on CNNs is widely used as a feature extractor in intensive prediction tasks [6,7], the performance of the visual Transformer [8,9] has surpassed CNNs [10] in many visual tasks. In the recent research on vision Transformer, ViT [11] and its variant models [12–14] focus more on solving the large-scale computing problem caused by the excessive length of the token sequence in the global attention mechanism. For example, Swin Transformer [15] divides the tokens sequence with length N into N/n groups through the sliding window and hierarchical design and uses the idea of divide and conquer to calculate local attention in groups, reducing the time complexity $O(N * n^2)$. CSWin Transformer [16] is also a local self-attention network, different from Swin Transformer. It adopts cross window self-attention and obtains better performance by gradually expanding the receptive field, and its time complexity is between ordinary window attention and global attention.

However, there are two main problems in high-resolution dense road segmentation based on Transformer. First, large-scale high-resolution remote sensing images have more pixels, which means a large amount of calculation. It is challenging to optimize the time complexity [17] of attention map calculation to ensure excellent segmentation performance. Second, remote sensing images containing dense roads have the characteristics of road density, complex and inconsistent scales. We need to build effective multiscale feature representations to solve the problem of inaccuracy and incompleteness in road segmentation [18]. However, for high-resolution dense prediction, less attention has been paid to how to establish compelling multiscale features. As shown in Figure 1, the above models have in common that they follow the general idea of CNNs and use a single-scale patch to build a simple multi-stage structure to obtain fine-to-coarse feature representations. Therefore, the multiscale feature representation based on multi-scale patch embedding still has room for improvement in the vision Transformer.

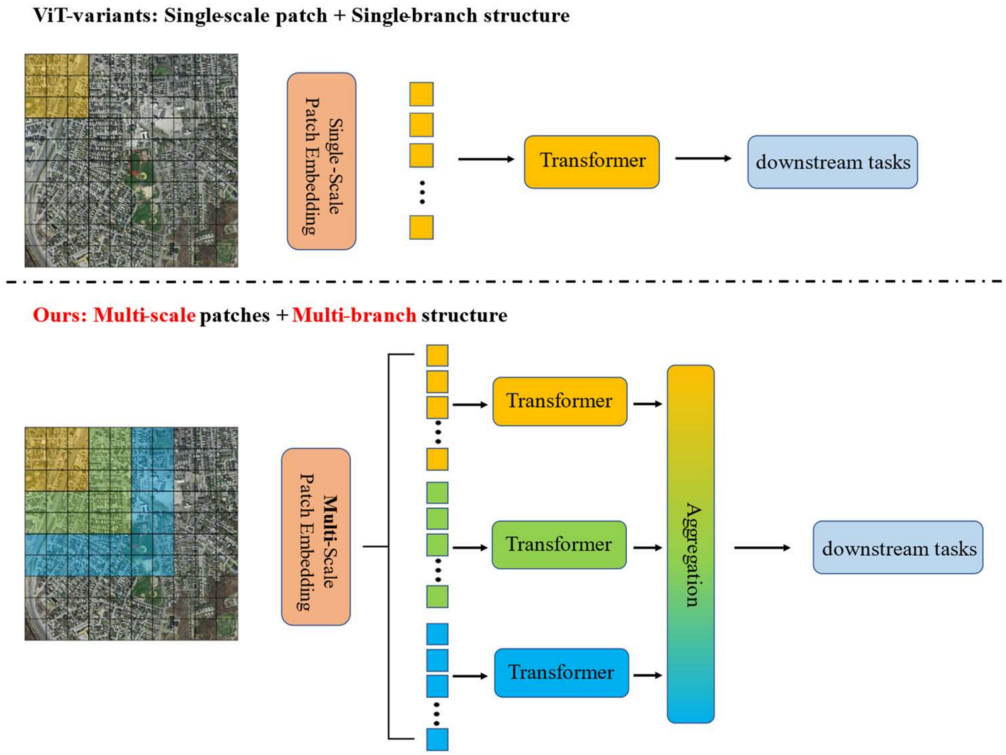

Figure 1. Top: single branch Transformer structure with single scale patch such as ViT-variants. Bottom: HA-RoadFormer, a multi-branch Transformer structure with multi-scale patches.

Our paper focuses on effectively constructing a hybrid-attention Transformer network with multi-scales embedding called HA-RoadFormer for dense road segmentation in large-scale high-resolution remote sensing images. We first exploit overlapping convolution

operation to obtain the visual patches with the same sequence length by appropriately changing the stride and padding. At the same time, the multiscale patch embedding tokenizes these visual patches with different sizes. Then, tokens with different scales are fed into respective Transformer branches independently. Due to the enormous computational cost for high-resolution remote sensing images, we design a hybrid-attention mechanism whose time complexity is linear with the number of tokens. Each Transformer branch encoder performs hybrid-attention on tokens of different scales. Finally, the output features from other Transformer encoder branches are aggregated to generate the fine-granularity and coarse-granularity feature representations. Considering convolution's powerful local feature extraction ability [17], we introduce convolution local features to augment the global features of Transformer in the feature aggregation model. Experiments on the Massachusetts dataset showed that the proposed HA-RoadFormer could improve the performance of dense road segmentation achieving a higher IoU (67.36%) compared to other state-of-the-art methods. The many performances also prove the power generation ability of our way.

The main contributions of the article are summarized as follows:

1. To reduce the high computational complexity of large-scale high-resolution remote sensing images, we propose a hybrid attention mechanism with linear complexity. Hybrid attention focuses on a few sampling points around the reference point. It assigns a small number of fixed keys to each query, which can alleviate the problem of limited input resolution. To retain the acquisition ability of long-distance dependence, we use random attention as a supplement.
2. We explore multiscale patch embedding and multi-branch structure from a unique perspective. Tokens of different scales are sent to the Transformer encoder independently through multiple paths. Then, the generated features are aggregated to realize coarse to fine feature representation at the same feature level.

2. Related Work

This section describes the related work from three aspects: road segmentation in remote sensing images, Transformer with local attention enhancement, and vision Transformer for semantic segmentation. Figure 2 clearly shows the context of this section.

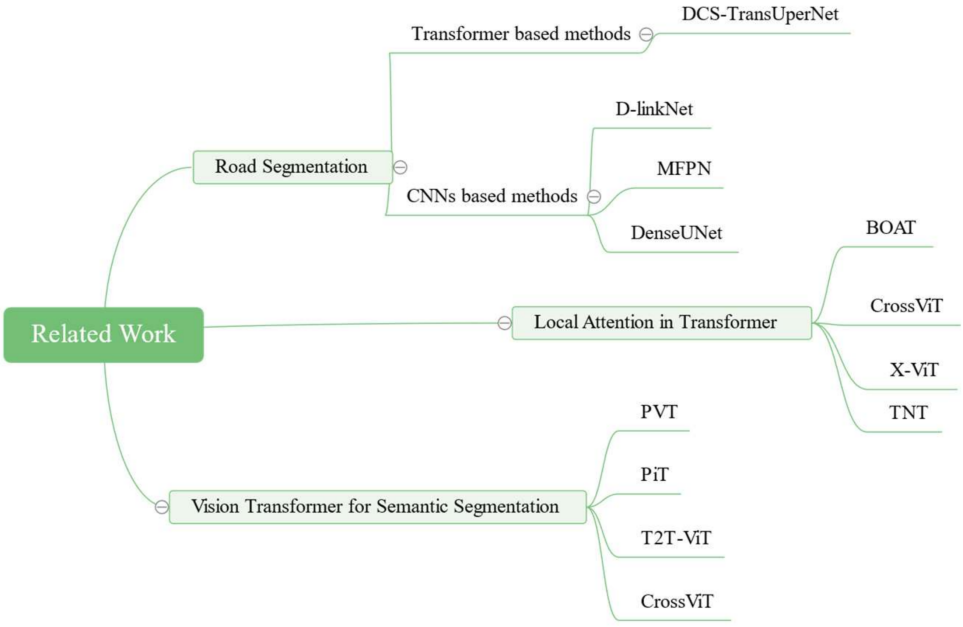

Figure 2. The overall context of related work.

2.1. Road Segmentation in Remote Sensing Image

Benefiting from deep learning, remote sensing image recognition has made significant progress [18–23]. Road segmentation based on deep learning is mainly the convolutional neural network. For example, D-linkNet [10] won the DeepGlobe 2018 Road Extraction Challenge championship by adding a cascaded dilated convolution structure between encoder and decoder. Gao et al. [24] design a multiscale feature pyramid structure to capture multiscale features achieving a better segmentation effect. Xin et al. [25] propose a deep learning network composed of dense residual block and jump connections, which fully uses the features at different levels.

Recently, the road segmentation framework based on Transformer has also been innovated. For example, Zhang et al. [26] propose a remote sensing road segmentation framework based on CSWin Transformer, improving IoU. However, the inference speed is relatively slow compared with the model based on CNNs.

2.2. Transformer with Local Attention Enhancement

ViT regards the input image as a patch sequence. Still, this rough patch embedding may destroy the local information of the image, which is also the difference between image and language. As a local feature extractor, convolution aggregates features through relatively fixed filters. This template matching process can effectively deal with most small datasets. Still, it faces the combination explosion of feature representation when dealing with large datasets such as massive remote sensing data. Compared with convolution, the local attention mechanism can dynamically generate attention weight according to the relationship between local elements.

To enhance the ability of local feature extraction and retain the non-convolution structure, many works [27–29] adapted to the patch structure through the local self-attention mechanism. For example, Swin Transformer limits the attention to one window, which introduces the locality of convolution operation and saves the amount of calculation. TNT [30] enhances local attention by more fine-grained patch division. We believe that there will be a stronger correlation between adjacent pixels. This paper proposes a hybrid attention mechanism with linear complexity composed of adjacent and random attention based on this prior knowledge. We will introduce it in detail in Section 3.

2.3. Vision Transformer for Semantic Segmentation

ViT inherits the columnar structure of the original Transformer and adopts the fixed resolution, so it ignores the representation of multiscale features. To better apply Transformer to intensive detection tasks, researchers introduced a pyramid structure to construct multiscale feature representation based on ViT. For example, PVT [31] adopts a multi-stage Transformer structure to obtain feature maps of different scales through different stages. This characteristic has achieved good results in semantic segmentation tasks. Similar to this structure, multi-stage Transformer networks include PiT [32], T2T-ViT [33]. At present, most segmentation networks adopt feature pyramid networks [34] structure, which can realize a seamless connection by replacing the CNNs backbone. However, this multiscale is only reflected in the feature level, ignoring the multiscale of patch embedding.

CrossViT [35] uses dual paths in a single-stage structure Transformer (such as ViT [11] and XCiT [36]) to embed patches of different scales. However, CrossViT only realizes the interaction between other branches through CLS (class) token [11], while HA-RoadFormer allows all patch interactions of different scales. In addition, in contrast with CrossViT (classification only), HA-RoadFormer more generally explores larger path dimensions (for example, more than two dimensions) and uses a multi-stage structure for semantic segmentation.

3. Method

3.1. Overall Structure

Figure 3 shows the overall structure of the hybrid-attention Transformer with a multi-branch (HA-RoadFormer). Since our goal is to explore dense road segmentation in high-

resolution remote sensing images, multiscale at feature levels is important for segmentation tasks. We propose a multi-stage structure rather than a single-stage one such as ViT. Specifically, a 5 stage Transformer encoder is constructed for multiple scales on feature level, which contains a conv-stem and four Transformer Blocks. Due to the large-scale high resolution of remote sensing images, it requires higher calculation. Therefore, our Transformer encoder adopts a local hybrid-attention mechanism and linear complexity of the token sequence. In addition, LeViT [37] absorbs the advantages of CNNs and adopts a convolutional stem block, which shows better low-level feature representation than non-overlapping patch embedding. Motivated by this, we also introduce a stem block composed of two 3×3 convolutions layers. The number of channels is $C_2/2$ and C_2, the stride is 2, and the size of the input feature map is $(W/4,\ H/4,\ C_2)$. Where C_2 represents the number of input channels in stage 2. Each convolution is followed by batch normalization and activation function.

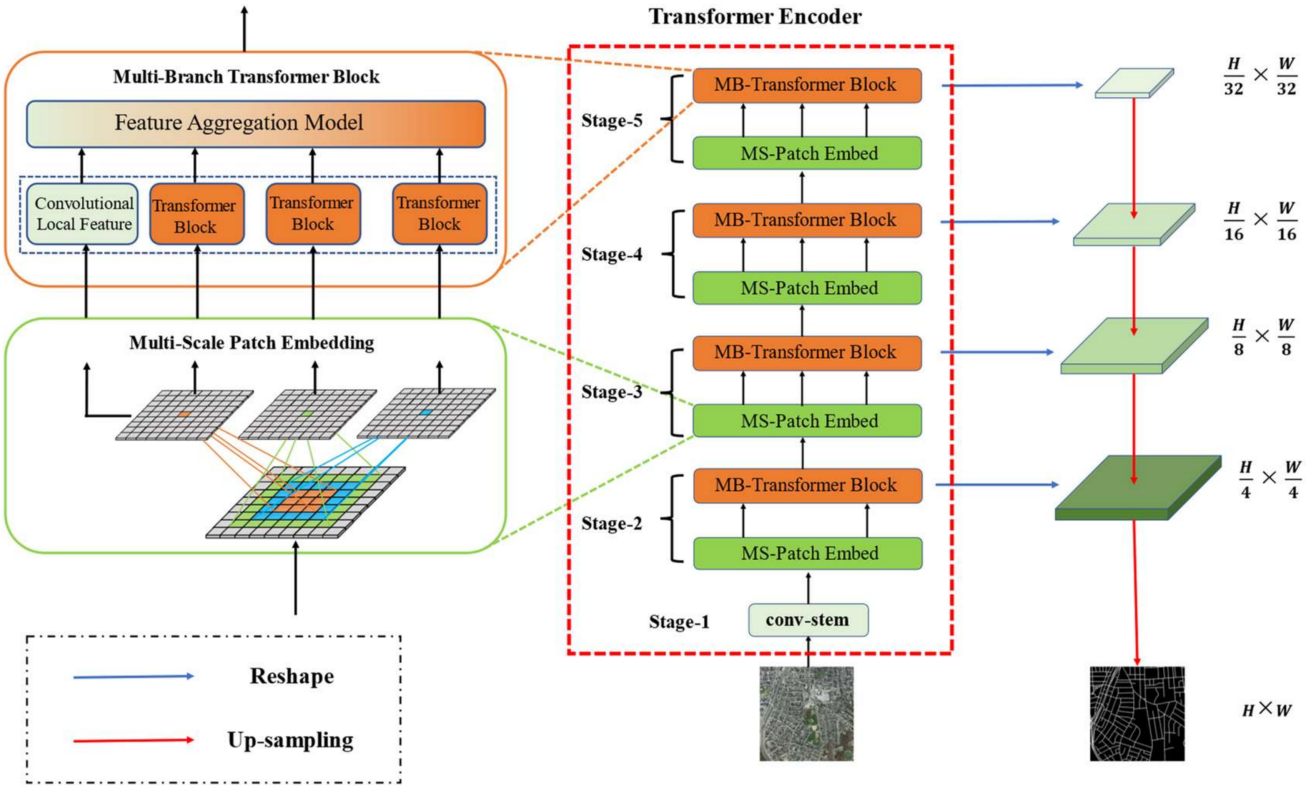

Figure 3. The overall structure of HA-RoadFormer.

From stages 2 to 5, each stage is composed of MS-Patch Embed and MB-Transformer Block. As shown in Figure 4, within the MB-Transformer Block, global average pooling is used to fuse the features from each branch. Finally, the feature maps of the 2 to 5 stages are gradually up-sampling and fused in the way of the U-net [38] to obtain the final road segmentation results.

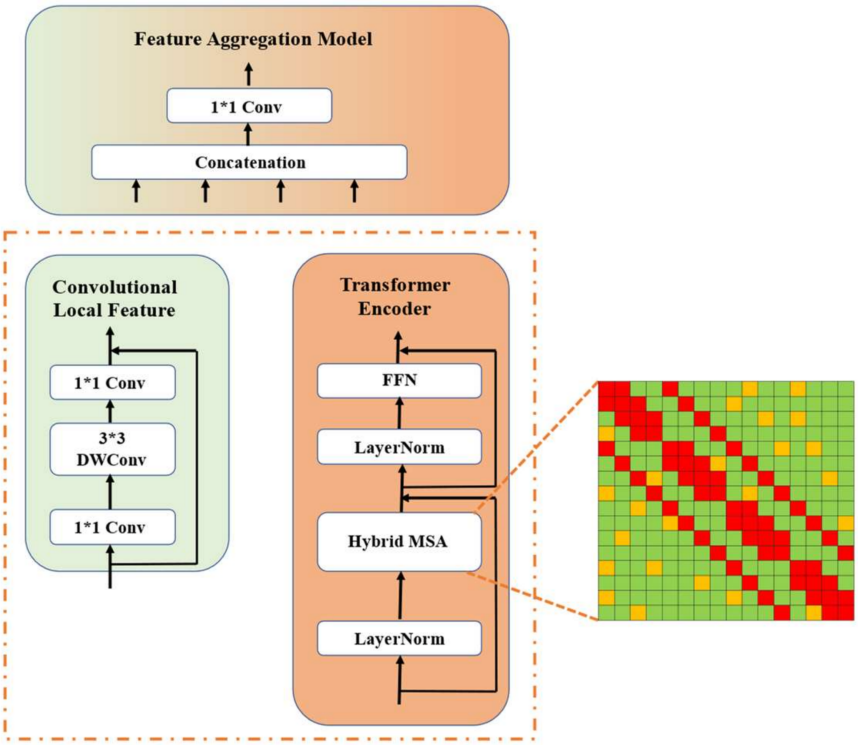

Figure 4. Illustration of the designed Muti-Branch Transformer Block.

3.2. Hybrid Attention with Linear Complexity

As the core operation of the Transformer, the attention mechanism [39] has been widely used in the field of natural language processing and computer vision. Its essence is to aggregate information according to the current query's weights given to the input information. The expression of attention mechanism is as follows:

$$Attention(Q, K, V) = Softmax(\frac{Q \times K^T}{\sqrt{d_k}}) \times V \tag{1}$$

where Q, K, and V are, respectively, from the mapping of the input $X \in \mathbb{R}^{N \times C}$ (N is the encoded patch length, C represents the encoding dimension) to different feature spaces. The expression of a linear transformation is as follows:

$$Q = X \times W_Q,\ K = X \times W_K,\ V = X \times W_V \tag{2}$$

For the global attention mechanism, there are the following two shortcomings. First, in terms of time complexity, because global attention needs to calculate an $N \times N$ attention matrix, its time complexity is $O(N^2)$, which is unacceptable for high-resolution images. Second, the global attention visual Transformer represented by ViT needs a larger amount of data and has a slower convergence speed than CNNs. This is because the global attention in Transformer equally treats the correlation between pixels with different distances and is driven by massive data to learn advanced semantic features. However, CNN's models have the advantage of inductive bias, which is the convolution kernel builds a stronger correlation between adjacent pixels. Based on such a priori knowledge, the convolution neural network can converge faster.

As shown in Figure 5, to solve the above problems, hybrid attention with linear complexity composed of adjacent and random attention is proposed in this paper. Inspired by inductive bias in CNNs, we focus on calculating the attention matrix between adjacent patches in adjacent attention. To ensure the Transformer's ability to capture long-distance

dependence, in random attention, each query randomly focuses on long-distance keys, and the super parameter determines the number of keys R. After the random superposition of multi-layer Transformers, excellent convergence can still be achieved in the end.

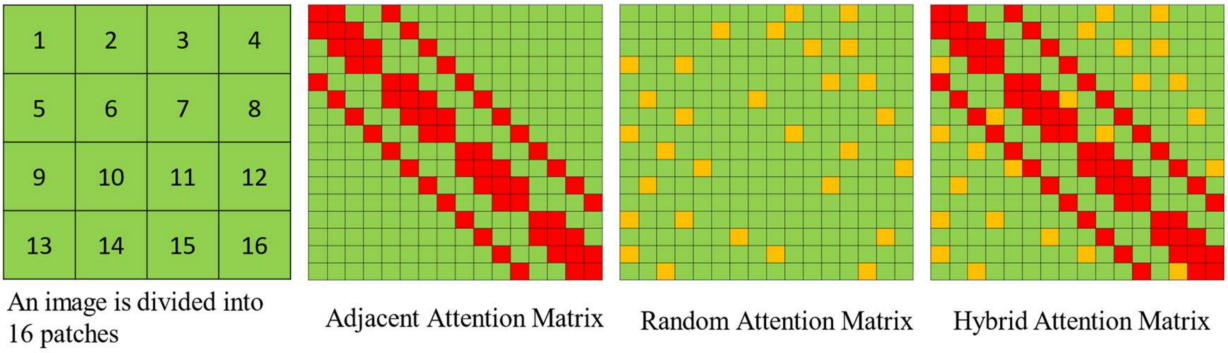

Figure 5. Illustration of the designed hybrid attention matrix composed of adjacent and random attention. In the attention matrix, green means not considered, red means considered by adjacent attention, and yellow means considered by random attention.

Hybrid attention has the following advantages: First, hybrid attention integrates the prior knowledge of inductive bias compared with global attention. From the perspective of optimization theory, the strong correlation between adjacent patches prioritizes backpropagation, while the weak correlation far away realizes the random gradient update randomly. Second, hybrid attention has the computational time cost of linear complexity, which is important for large-scale high-resolution remote sensing images. Through the analysis, the adjacent attention only needs to calculate the five diagonals in the attention matrix selected according to the length of the patch, so its time complexity is $O(N)$. The $O(N)$ time complexity can also be achieved by controlling the parameter R for random attention.

Specifically, given the input sequence $X = (x_1, x_2, \ldots, x_N) \in \mathbb{R}^{N \times C}$, the sequence is processed by the hybrid attention mechanism proposed in this paper. The attention mechanism is described by a directed graph D, the node-set $[N] = \{1, 2, \ldots, N\}$, and the directed edge weight of the graph is the self-attention score between two nodes. For a more straightforward introduction, this paper defines an adjacency matrix $A \in [0, 1]^{N \times N}$ belonging to graph D. If q_i and k_i participate in attention calculation, then $a_{ij} = 1$, otherwise $a_{ij} = 0$. For example, a matrix whose adjacency matrix is all 1, that is, attention is calculated between all patches, which is global attention.

Therefore, in the random attention mechanism proposed in this paper, the vector Q_i randomly selects vector K_j for attention calculation. Random attention is expressed as the Formula (3):

$$A[i, random(j)] = 1 \tag{3}$$

You only need to calculate the attention between two adjacent patches for the adjacent attention matrix. For example, in patch 6 in Figure 5, you only need to set $A[6:2]$, $A[6:5]$, $A[6:6]$, $A[6:7]$, $A[6:10]$ to 1.

3.3. Multiscale Patch Embedding

In this paper, a multiscale patch embedding layer is designed to obtain coarse and fine visual tokens at the same feature level. We use convolution operation to acquire overlapping patches by controlling a certain stride and padding. Specifically, the input of the stage i is obtained by the output $X_i \in \mathbb{R}^{H_{i-1} \times W_{i-1} \times C_{i-1}}$ of the stage $i-1$ through 2D reshape. We map X_i to a new token with the number of channels C_i through function $F_{k \times k}(\cdot)$, where $F_{k \times k}(X_i)$ is a 2D convolution operation, the convolution kernel size is $k \times k$,

the stride is s, and padding is p. The formula expression of the length and width of the output 2D feature $F_{k\times k}(X_i) \in \mathbb{R}^{H_i \times W_i \times C_i}$ is as follows:

$$H_i = floor(\frac{H_{i-1} - k + 2 \times p}{s}) + 1,\ W_i = floor(\frac{W_{i-1} - k + 2 \times p}{s}) + 1 \quad (4)$$

Through (4), it is found that we can control the length of the tokens sequence by adjusting stride and padding so as to output the same resolution feature map with different patch sizes. Therefore, we use convolution operations with different convolution kernel sizes in parallel in the patch embedding layer. For example, as shown in Figure 3, we can obtain multi-sized visual tokens of the same sequence length with patch sizes.

To reduce the amount of calculation, we will make two detailed explanations in our work. First, by superimposing the convolution operation with a small kernel and gradually expanding the receptive field, the same effect of convolution operation with a large kernel is achieved. It reduces the number of parameters [40]. Therefore, we use the convolution of two 3×3 to replace a 5×5 convolution and three 3×3 convolutions instead of a 7×7 convolution. Second, we use depthwise separable convolution, composed of depthwise and pointwise convolution [41]. All convolution layers are followed by batch normalization and activation functions. Finally, visual tokens of different patch sizes are independently sent to the Transformer encoder.

4. Experiment

4.1. Dataset and Preprocessing

Massachusetts roads [42] is the largest public road dataset globally, consisting of 1171 images, including 1108 training images, 14 verification images, 49 test images, and corresponding labels. The size of each image is $1500 * 1500$ pixels, and the resolution is 1.2 m/pixel. This dataset contains various features, such as roads, grasslands, forests, and buildings.

According to the structure of HA-RoadFormer, the remote sensing image needs to be down-sampling 32 times, and 512 is exactly a multiple of 32. Considering the memory of hardware devices and segmentation results, the remote sensing images and the corresponding label images in the dataset are randomly cut into 512×512 [43] pixel image samples. Then, all the image samples are randomly divided into a new training set and validation set in the proportion of 4:1 because 20 percent of the verification set is widely used [44]. Finally, the training set contains 14,366 image samples, and the verification set contains 3592 image samples.

4.2. Implementation Details

In this experiment, we implemented HA-RoadFormer using the open-source framework PyTorch. Before training, we use data enhancement, such as random rotation, to expand the data again. The sum of cross entropy loss and dice loss function is used in the training process as the final loss function [26] and optimized with Adam [45] algorithm. We used 4 NVIDIA RTX 2080ti GPUs (11 G) by a workstation with four graphic card slots. The data flow of multi-GPU training is shown in Figure 6. According to [46], we set the batch size to 2, and the initial learning rate was 0.001. We set the training epoch to 100 and use early stopping by observing the decline curve of the loss function. According to [47], when the loss of the training set no more prolonged decreases and tends to be stable, while the loss of the verification set begins to increase, we stop the iteration.

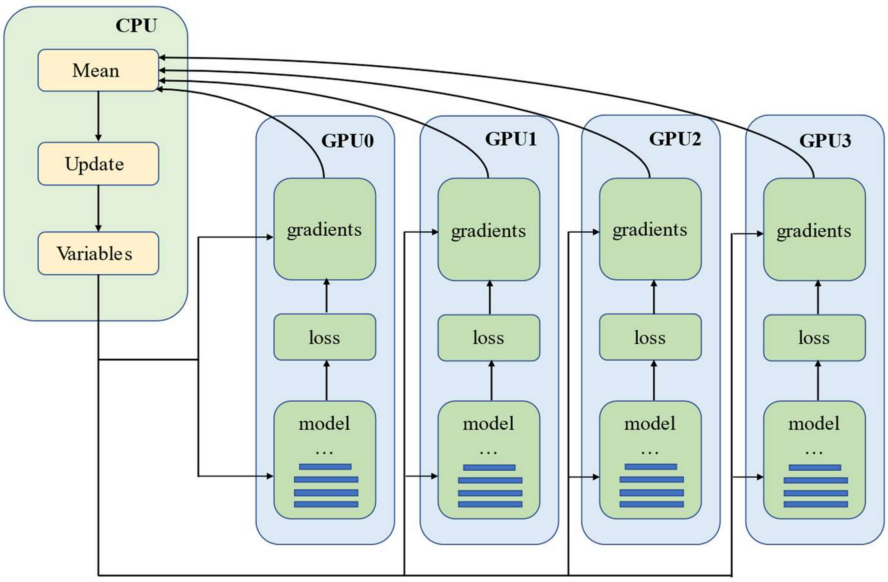

Figure 6. The data flow of multi-GPU training.

4.3. Evaluation Metrics

To quantify the results of road segmentation, the most common evaluation metrics [48,49] in the field of semantic segmentation are used: Precision, Recall, F1-score, and IoU. Road segmentation is regarded as a binary classification problem in our work, and the prediction results are divided into two categories: road and non-road. For this problem, the sample data can be classified into four cases: true positive (*TP*), false positive (*FP*), true negative (*TN*), and false-negative (*FN*). The evaluation metric formulas deduced by them are as follows:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1\text{-}score = \frac{2TP}{2TP + FN + FP} \tag{7}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{8}$$

4.4. Experimental Results on Massachusetts Road Dataset

To verify the performance of HA-RoadFormer in road segmentation task, we select representative road segmentation networks, such as DeepLabV3+ [50], D-LinkNet [10], D-ResUnet [51], and Transformer based networks DCS-TransUperNet [26] and ViT [11], making comparative experiments on Massachusetts road dataset.

Table 1 shows the quantitative analysis under different models. As shown in Table 1, HA-RoadFormer achieved the best performance in Precision (82.94%), Recall (79.43%), F1-Score (81.14%), and IoU (67.36%). Experiments show that our HA-RoadFormer has better road segmentation ability. In addition, the attention mechanism of linear complexity reduces the amount of calculation. Compared with other Transformer based models, the inference speed of HA-RoadFormer has been greatly improved, which further narrowed the inference speed gap between HA-RoadFormer and CNNs based models.

Table 1. Quantitative analysis results with different methods on the Massachusetts roads dataset.

Method	Precision (%)	Recall (%)	F1-Score (%)	IoU (%)	Inference (ms)
DeepLabV3+	75.47	72.23	73.81	58.86	1.65
D-LinkNet	77.91	69.85	73.66	58.54	**1.22**
D-ResUnet	75.95	77.58	76.75	61.69	2.32
ViT	80.55	77.64	79.06	63.85	2.45
DCS-TransUperNet	82.44	78.43	80.38	65.56	2.84
HA-RoadFormer	**82.94**	**79.43**	**81.14**	**67.36**	1.86

Figure 7 shows a very representative visualization result. We selected two scenes with relatively dense roads (the first and third row) to compare the results. To compare the segmentation results more clearly, we use the red rectangular box to circle the critical comparison places, and enlarge the area (the second and fourth lines). In Figure 7, (a) is a remote sensing image, (b) is the corresponding label, and then the third to seventh columns are the road segmentation results of DeepLabV3+, D-LinkNet, D-ResUnet, DCS-TransUperNet and HA-RoadFormer (the last column). Through the experimental comparison in the second row, it is found that HA-RoadFormer with multiscale patch embedding has apparent advantages in the segmentation of minor roads. In image 2, we highlight and compare the scenes of multiple adjacent main roads with great difficulty in detection. The difficulty is that the pixels at the edge of the road are very easy to confused, resulting in the segmentation result being an adhesive block. From the locally enlarged view (the fourth row) in image 2, compared with other networks, the segmentation result of HA-RoadFormer (the last column) has better continuity, and the edge of the road is clearer. In addition, Figure 8 shows more examples of segmentation results.

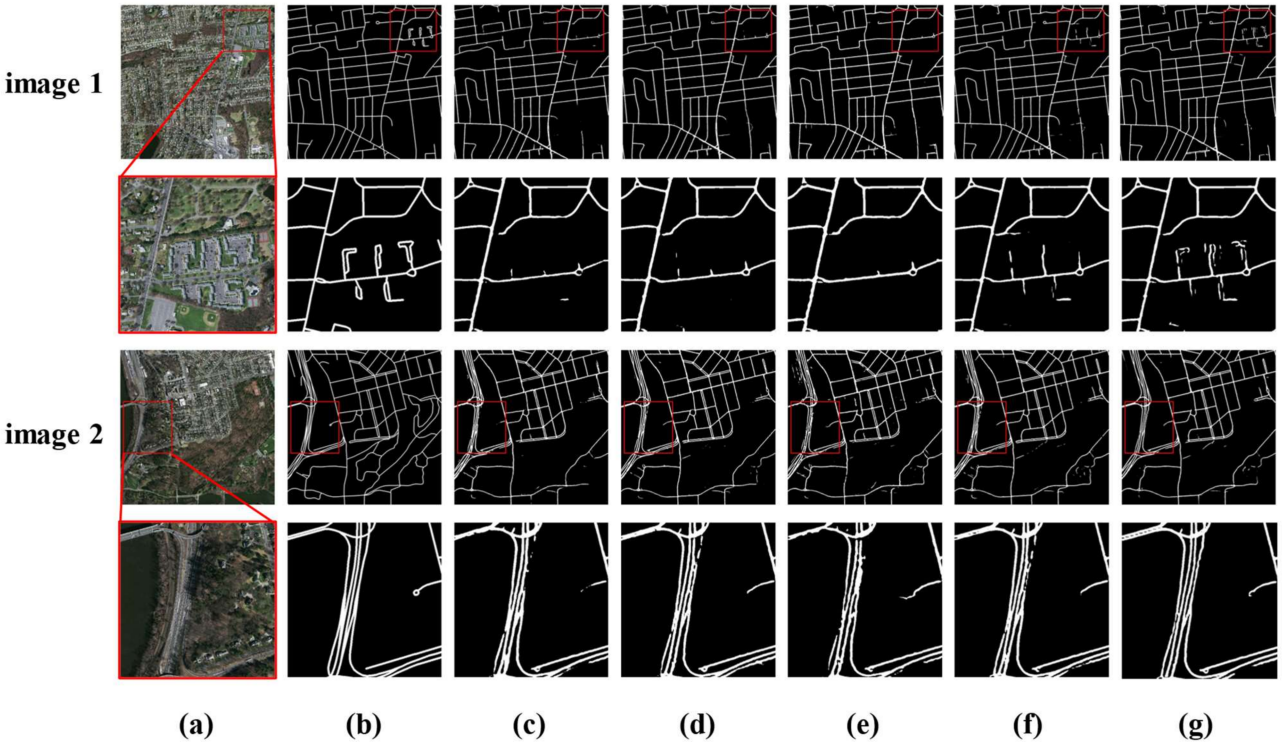

Figure 7. Visualization result of road segmentation from different methods. (**a**) The test image. (**b**) The ground truth. (**c**) Results with DeepLabV3+. (**d**) Results with D-LinkNet. (**e**) Results with D-ResUnet. (**f**) Results with DCS-TransUperNet. (**g**) Results with HA-RoadFormer.

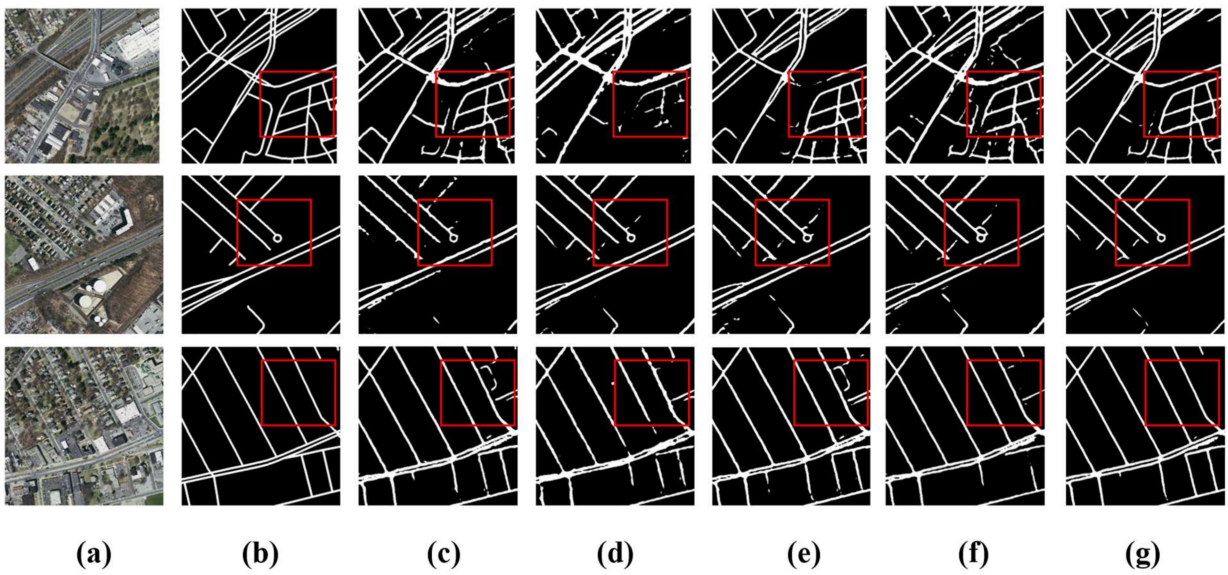

Figure 8. More examples of segmentation results. (**a**) The test image. (**b**) The ground truth. (**c**) Results with DeepLabV3+. (**d**) Results with D-LinkNet. (**e**) Results with D-ResUnet. (**f**) Results with DCS-TransUperNet. (**g**) Results with HA-RoadFormer.

4.5. Ablation Experiment

To further explore the role of multiscale patch embedding in road segmentation, we visualized the attention feature map. As shown in Figure 9, (a) is the input remote sensing image, (b) is the ground truth, and (c) represents the result of road segmentation. (e–g) are from different scale patch embedded branches in the fifth stage of HA-RoadFormer. (e) is the heat map from small patches. It pays more attention to the detailed information of the road and captures the road pixels more accurately. Still, it lacks the overall continuity, which is reflected in the intermittent state of the highlighted points in the heat map. (f, g) are the heat maps from medium-scale and large-scale patches. It is a higher-level semantic feature which pays more attention to the broad range of the image and enhances the continuity of road segmentation. The heat map (g) shows that large patches filter out the noise and small road targets. At the same time, it will also misjudge large areas such as roads, such as gullies and dark slender regions. Heat map (h) is the result of the feature fusion of three different branches (e–g). It combines the overall features of large patches and improves the continuity of road segmentation and retains the fine features of small patches, and improves the segmentation ability of minor roads. Finally, the road segmentation result (c) is obtained. These results show that the multi-branch structure embedding based on a multiscale patch can acquire fine and coarse features and capture segmented objects of different scales more accurately.

In addition, (d) represents the heat map of the attention feature matrix in the second stage of HA-RoadFormer. We find that the model gradually began to pay attention to the pixels near the road. Although the heat map does not show great attention in yellow or red, it can still show that the hybrid attention with linear complexity proposed has an effective feature extraction ability in the shallow layer of the network. It has been proved that focusing on adjacent local information and randomly focusing on long-distance dependence is effective.

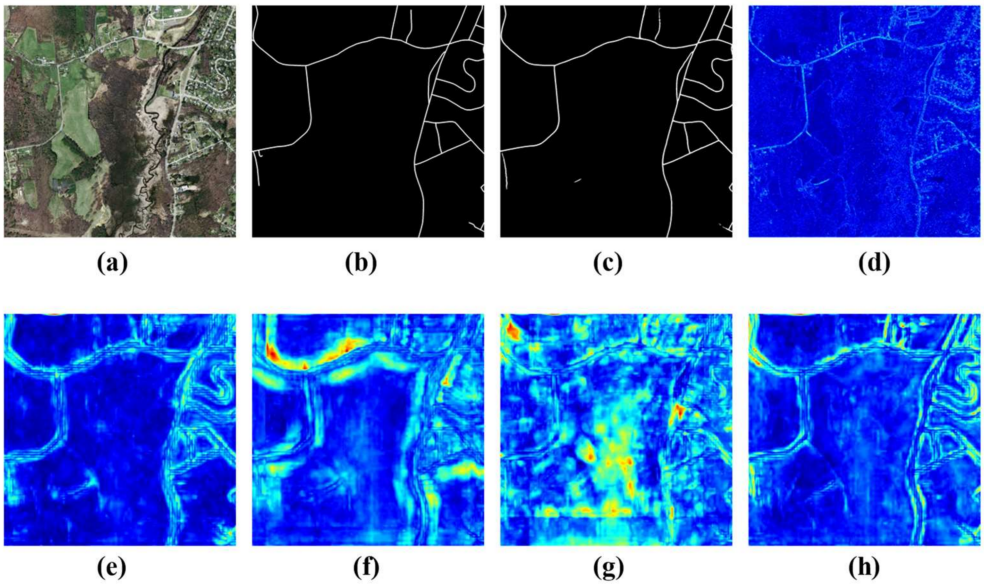

Figure 9. Analysis of attention heat map of the patch from different scales. (**a**) The test image. (**b**) The ground truth. (**c**) Results with HA-RoadFormer. (**d**) the heat map of the attention feature matrix in the second stage of HA-RoadFormer. (**e**) the heat map from the small-scale branch. (**f**) the heat map from the medium-scale branch. (**g**) the heat map from the large-scale branch. (**h**) the heat map of the feature fusion of three different branches.

Table 2 quantitatively analyzes different ablation experiments. Compared with HA-RoadFormer (IoU: 67.36%), we use the models with single branch embedding and double branch embedding, achieving the IoU as follows: 63.94%, 64.27%, respectively. Experiments show that patch embedding with different scales plays significant roles in feature extraction, proving the effectiveness of multiscale embedding in HA-RoadFormer. In addition, we design a contrast model called RoadFormer, which replaces the hybrid attention with global attention and the other parts are consistent with HA-RoadFormer. The experimental results show that the IoU (67.36%) of HA-RoadFormer is slightly higher than that of RoadFormer (66.50%). However, the inference time is greatly improved, reduced from 2.94 ms to 1.86 ms.

Table 2. Quantitative analysis results in different patch embedding combinations and attention mechanisms.

Method	IoU (%)	Inference (ms)
HA-RoadFormer with two patches	64.27	1.66
HA-RoadFormer with one patch	63.94	**1.40**
RoadFormer	66.50	2.94
HA-RoadFormer with three patches	**67.36**	1.86

5. Conclusions

This paper aims to solve the problem of dense road segmentation in high-resolution remote sensing images. Since the time complexity of the global attention in the original Transformer is directly proportional to the square of the length of the token sequence, this amount of calculation results in lower inference speed for large-scale high-resolution remote sensing images. Therefore, this paper proposes local hybrid attention with linear complexity, which retains the acquisition ability of Transformer's long-distance dependence and introduces the idea of inductive bias such as CNNs. This prior knowledge makes the model converge faster. We propose a Transformer Road segmentation network based on multiscale patch embedding, which obtains the multiscale feature representation under the same level of features through multiscale patch embedding to solve the problem of dense

segmentation. We have got the road segmentation results of SOTA on the Massachusetts roads dataset. The segmentation results are more significantly improved in the dense road remote sensing images. In addition, we conducted ablation experiments. Through the comparative analysis of heap maps, it is found that there is an interpretable effect of multiscale patch embedding in attention.

Although HA-RoadFormer has achieved excellent road segmentation results, there is still room for optimization. Roads are continuous in space, and this unique topology is the most prominent feature of roads. It is a very promising job to study how to fuse this topology information into neural networks to improve the quality of road segmentation.

Author Contributions: Conceptualization, Z.Z. and C.M.; methodology, Z.Z. and C.M.; software, C.M.; validation, Z.Z.; formal analysis, C.L.; investigation, Q.T.; resources, Q.T.; data curation, C.M.; writing, Z.Z., Y.Z. and C.M.; original draft preparation, C.M.; visualization, Z.Z.; supervision, C.L., Y.Z. and Q.T.; project administration, C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by the Fundamental Research Fund of Beijing Municipal Education Commission and North China University of Technology Research Start-up Funds.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Hinz, S.; Baumgartner, A.; Ebner, H. Modeling contextual knowledge for controlling road extraction in urban areas. In Proceedings of the IEEE/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas (Cat. No. 01EX482), Rome, Italy, 8–9 November 2001; IEEE: Piscataway, NJ, USA, 2001; pp. 40–44.
2. Wang, J.; Qin, Q.; Gao, Z.; Zhao, J.; Ye, X. A New Approach to Urban Road Extraction Using High-Resolution Aerial Image. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 114. [CrossRef]
3. Shi, W.; Miao, Z.; Debayle, J. An Integrated Method for Urban Main-Road Centerline Extraction from Optical Remotely Sensed Imagery. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 3359–3372. [CrossRef]
4. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
5. Yu, C.; Xiao, B.; Gao, C.; Yuan, L.; Zhang, L.; Sang, N.; Wang, J. Lite-Hrnet: A lightweight high-resolution network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10440–10450.
6. Oliveira, G.L.; Burgard, W.; Brox, T. Efficient deep models for monocular road segmentation. In Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Korea, 9–14 October 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 4885–4891.
7. Levi, D.; Garnett, N.; Fetaya, E.; Herzlyia, I. StixelNet: A deep convolutional network for obstacle detection and road segmentation. In Proceedings of the British Machine Vision Conference, Swansea, UK, 7–10 September 2015; Volume 1, p. 4.
8. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y. A Survey on Visual Transformer. *arXiv* **2020**, arXiv:2012.12556.
9. Zhang, Z.; Xu, Z.; Liu, C.; Tian, Q.; Wang, Y. Cloudformer: Supplementary Aggregation Feature and Mask-Classification Network for Cloud Detection. *Appl. Sci.* **2022**, *12*, 3221. [CrossRef]
10. Zhou, L.; Zhang, C.; Wu, M. D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 182–186.
11. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
12. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. Vivit: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 6836–6846.
13. Chen, Z.; Xie, L.; Niu, J.; Liu, X.; Wei, L.; Tian, Q. Visformer: The vision-friendly transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 589–598.
14. Zhou, D.; Kang, B.; Jin, X.; Yang, L.; Lian, X.; Jiang, Z.; Hou, Q.; Feng, J. Deepvit: Towards Deeper Vision Transformer. *arXiv* **2021**, arXiv:2103.11886.

15. 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.
16. Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; Guo, B. Cswin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. *arXiv* **2021**, arXiv:2107.00652.
17. Broggi, A. Parallel and Local Feature Extraction: A Real-Time Approach to Road Boundary Detection. *IEEE Trans. Image Processing* **1995**, *4*, 217–223. [CrossRef]
18. Li, H.-C.; Hu, W.-S.; Li, W.; Li, J.; Du, Q.; Plaza, A. A^3CLNN: Spatial, Spectral and Multiscale Attention ConvLSTM Neural Network for Multisource Remote Sensing Data Classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *33*, 747–761. [CrossRef]
19. Ma, F.; Zhang, F.; Xiang, D.; Yin, Q.; Zhou, Y. Fast Task-Specific Region Merging for SAR Image Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5222316. [CrossRef]
20. Ma, F.; Zhang, F.; Yin, Q.; Xiang, D.; Zhou, Y. Fast SAR Image Segmentation With Deep Task-Specific Superpixel Sampling and Soft Graph Convolution. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5214116. [CrossRef]
21. Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T. FAIR1M: A Benchmark Dataset for Fine-Grained Object Recognition in High-Resolution Remote Sensing Imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 116–130. [CrossRef]
22. Yang, J.-Y.; Li, H.-C.; Hu, W.-S.; Pan, L.; Du, Q. Adaptive Cross-Attention-Driven Spatial-Spectral Graph Convolutional Network for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 6004705. [CrossRef]
23. Yue, Z.; Gao, F.; Xiong, Q.; Wang, J.; Huang, T.; Yang, E.; Zhou, H. A Novel Semi-Supervised Convolutional Neural Network Method for Synthetic Aperture Radar Image Recognition. *Cogn. Comput.* **2021**, *13*, 795–806. [CrossRef]
24. Gao, X.; Sun, X.; Zhang, Y.; Yan, M.; Xu, G.; Sun, H.; Jiao, J.; Fu, K. An End-to-End Neural Network for Road Extraction from Remote Sensing Imagery by Multiple Feature Pyramid Network. *IEEE Access* **2018**, *6*, 39401–39414. [CrossRef]
25. Xin, J.; Zhang, X.; Zhang, Z.; Fang, W. Road Extraction of High-Resolution Remote Sensing Images Derived from DenseUNet. *Remote Sens.* **2019**, *11*, 2499. [CrossRef]
26. Zhang, Z.; Miao, C.; Liu, C.; Tian, Q. DCS-TransUperNet: Road Segmentation Network Based on CSwin Transformer with Dual Resolution. *Appl. Sci.* **2022**, *12*, 3511. [CrossRef]
27. Yu, T.; Zhao, G.; Li, P.; Yu, Y. BOAT: Bilateral Local Attention Vision Transformer. *arXiv* **2022**, arXiv:2201.13027.
28. Lin, H.; Cheng, X.; Wu, X.; Yang, F.; Shen, D.; Wang, Z.; Song, Q.; Yuan, W. Cat: Cross Attention in Vision Transformer. *arXiv* **2021**, arXiv:2106.05786.
29. Bulat, A.; Perez Rua, J.M.; Sudhakaran, S.; Martinez, B.; Tzimiropoulos, G. Space-Time Mixing Attention for Video Transformer. *Adv. Neural Inf. Processing Syst.* **2021**, *34*, 5223512.
30. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in Transformer. *Adv. Neural Inf. Processing Syst.* **2021**, *34*, 5216488.
31. Zhang, C.; Wan, H.; Liu, S.; Shen, X.; Wu, Z. Pvt: Point-Voxel Transformer for 3d Deep Learning. *arXiv* **2021**, arXiv:2108.06076.
32. Heo, B.; Yun, S.; Han, D.; Chun, S.; Choe, J.; Oh, S.J. Rethinking Spatial Dimensions of Vision Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 11936–11945.
33. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.-H.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-Token Vit: Training Vision Transformers from Scratch on Imagenet. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 558–567.
34. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
35. Chen, C.-F.R.; Fan, Q.; Panda, R. Crossvit: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 357–366.
36. Ali, A.; Touvron, H.; Caron, M.; Bojanowski, P.; Douze, M.; Joulin, A.; Laptev, I.; Neverova, N.; Synnaeve, G.; Verbeek, J. Xcit: Cross-Covariance Image Transformers. *Adv. Neural Inf. Processing Syst.* **2021**, *34*, 5241254.
37. Graham, B.; El-Nouby, A.; Touvron, H.; Stock, P.; Joulin, A.; Jégou, H.; Douze, M. LeViT: A Vision Transformer in ConvNet's Clothing for Faster Inference. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 12259–12269.
38. 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; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
39. Fukui, H.; Hirakawa, T.; Yamashita, T.; Fujiyoshi, H. Attention branch network: Learning of attention mechanism for visual explanation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 10705–10714.
40. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
41. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
42. Levy, J.I.; Houseman, E.A.; Spengler, J.D.; Loh, P.; Ryan, L. Fine Particulate Matter and Polycyclic Aromatic Hydrocarbon Concentration Patterns in Roxbury, Massachusetts: A Community-Based GIS Analysis. *Environ. Health Perspect.* **2001**, *109*, 341–347. [CrossRef]

43. Ding, C.; Weng, L.; Xia, M.; Lin, H. Non-Local Feature Search Network for Building and Road Segmentation of Remote Sensing Image. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 245. [CrossRef]
44. Bella, G.; Massacci, F.; Paulson, L.C. An Overview of the Verification of SET. *Int. J. Inf. Secur.* **2005**, *4*, 17–28. [CrossRef]
45. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
46. Wan, J.; Xie, Z.; Xu, Y.; Chen, S.; Qiu, Q. DA-RoadNet: A Dual-Attention Network for Road Extraction from High Resolution Satellite Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 6302–6315. [CrossRef]
47. Prechelt, L. Early Stopping-but When? In *Neural Networks: Tricks of the Trade*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 55–69.
48. Panboonyuen, T.; Jitkajornwanich, K.; Lawawirojwong, S.; Srestasathiern, P.; Vateekul, P. Road Segmentation of Remotely-Sensed Images Using Deep Convolutional Neural Networks with Landscape Metrics and Conditional Random Fields. *Remote Sens.* **2017**, *9*, 680. [CrossRef]
49. Sun, Z.; Geng, H.; Lu, Z.; Scherer, R.; Woźniak, M. Review of Road Segmentation for SAR Images. *Remote Sens.* **2021**, *13*, 1011. [CrossRef]
50. 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; pp. 801–818.
51. Liu, Z.; Feng, R.; Wang, L.; Zhong, Y.; Cao, L. D-Resunet: Resunet and dilated convolution for high resolution satellite imagery road extraction. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 3927–3930.