symmetry

MDPI

Article

Road Extraction Method of Remote Sensing Image Based on Deformable Attention Transformer

Ling Zhao [1], Jianing Zhang [1], Xiujun Meng [2,3], Wenming Zhou [4], Zhenshi Zhang [5,*] and Chengli Peng [1,*]

[1] School of Geosciences and Info-Physics, Central South University, Changsha 410083, China; zhaoling@csu.edu.cn (L.Z.); 215011046@csu.edu.cn (J.Z.)
[2] Tianjin Zhongwei Aerospace Data System Technology Company Limited, Tianjin 300301, China; mengxiujun1000@126.com
[3] Tianjin Enterprise Key Laboratory of Intelligent Remote Sensing and Information Processing Technology, Tianjin 301899, China
[4] China Railway Design Corporation, Tianjin 300308, China; wenmingchow@126.com
[5] Undergraduate School, National University of Defense Technology, Changsha 410080, China
* Correspondence: zhangzhenshi@nudt.edu.cn (Z.Z.); pengcl@csu.edu.cn (C.P.)

Abstract: Road extraction is a typical task in the semantic segmentation of remote sensing images, and one of the most efficient techniques for solving this task in recent years is the vision transformer technique. However, roads typically exhibit features such as uneven scales and low signal-to-noise ratios, which can be understood as the asymmetry between the road and the background category and the asymmetry in the transverse and longitudinal shape of the road. Existing vision transformer models, due to their fixed sliding window mechanism, cannot adapt to the uneven scale issue of roads. Additionally, self-attention, based on fully connected mechanisms for long sequences, may suffer from attention deviation due to excessive noise, making it unsuitable for low signal-to-noise ratio scenarios in road segmentation, resulting in incomplete and fragmented road segmentation results. In this paper, we propose a road extraction based on deformable self-attention computation, termed DOCswin-Trans (Deformable and Overlapped Cross-Window Transformer), to solve these problems. On the one hand, we develop a DOC-Transformer block to address the scale imbalance issue, which can utilize the overlapped window strategy to preserve the overall contextual semantic information of roads as much as possible. On the other hand, we propose a deformable window strategy to adaptively resample input vectors, which can direct attention automatically to the foreground areas relevant to roads and thereby address the low signal-to-noise ratio problem. We evaluate the proposed method on two popular road extraction datasets (i.e., DeepGlobe and Massachusetts datasets). The experimental results demonstrate that the proposed method outperforms baseline methods. On the DeepGlobe dataset, the proposed method achieves an IoU improvement ranging from 0.63% to 5.01% compared to baseline methods. On the Massachusetts dataset, our method achieves an IoU improvement ranging from 0.50% to 6.24% compared to baseline methods.

Keywords: road segmentation; remote sensing image; CSwin transformer; deformable attention transformer

check for updates

Citation: Zhao, L.; Zhang, J.; Meng, X.; Zhou, W.; Zhang, Z.; Peng, C. Road Extraction Method of Remote Sensing Image Based on Deformable Attention Transformer. *Symmetry* **2024**, *16*, 468. https://doi.org/10.3390/sym16040468

Academic Editor: Changxin Gao

Received: 2 March 2024
Revised: 1 April 2024
Accepted: 6 April 2024
Published: 12 April 2024

1. Introduction

Road extraction based on high-resolution remote sensing images is an important task in the remote sensing image processing community, and ensuring the connectivity and integrity of the extracted roads is of great importance for many applications, such as urban construction, transportation planning, road network updating, and route navigation [1–3]. However, the connectivity and integrity of the extracted roads can hardly be ensured due to the following asymmetry questions regarding roads in high-resolution remote sensing images. The first is the uneven problem of length and width scale caused by the asymmetrical shape of the road in transverse and longitudinal planes. In particular, roads are densely distributed in narrow and elongated patterns, with lengths extending to several kilometers and widths rarely exceeding several tens of meters, contrasting sharply with other land cover categories that often appear as dense blocks. This extreme discrepancy in the length-to-width ratio leads to a highly imbalanced scale of road features. The second challenge stems from the low signal-to-noise ratio (SNR) of roads, resulting in an asymmetry in the number of samples representing two categories. Roads, as foreground elements, occupy a small proportion of the entire image. For instance, in the DeepGlobe dataset, roads account for only 5%. Meanwhile, the background contains diverse and complex land cover categories, serving as noise that can interfere with model recognition and decision-making. Factors such as roadside vegetation and shadows, vehicles on the road surface, and occlusions like tunnels all pose challenges to the continuity of road features. The aforementioned issues lead to sub-optimal performance in connectivity and integrity for existing methods, both of which are crucial visual metrics for evaluating road extraction results.

The developmental history of road extraction tasks based on high-resolution remote sensing imagery can be roughly divided into two stages based on their principles:

(1) Model-driven handcrafted feature methods. The core of these methods lies in designing road feature representations based on manual priors, including threshold segmentation methods [4–7] for pixel selection and clustering methods [8–11]. In general, traditional algorithms directly consider the underlying features of roads, such as shape, color, edges, texture, and grayscale. However, heavy reliance on these priors may lead to limited precision and robustness, leading to their inferior performance in complex scenarios.

(2) Data-driven learning feature methods. The essence of these methods lies in constructing extensive annotated datasets and using deep learning to automatically learn features from the labeled data. Representative works mainly build upon classical CNN backbone networks, incorporating strategies to integrate multi-scale semantic information and enlarge receptive fields to address issues such as fitting road scales or improving the connectivity of segmentation results. For instance, Deep ResUnet [12] combines the advantages of the U-shaped feature fusion pattern from the Unet with residual connections from the Resnet to form a deep convolution neural network architecture. The D-Linknet [13], based on the original Linknet [14], adds extra dilated convolution layers to increase the receptive field, facilitating the extraction and fusion of semantic information at different levels and obtaining detailed road information. However, stacking layers in CNN models to enlarge receptive fields may lead to issues such as model complexity and over-fitting, hindering models from capturing global context, which is indispensable for tasks like road extraction. To address these challenges, recent transformer [15] architectures have enabled long-range modeling to capture global contextual information, making them more suitable for tasks with significant horizontal and vertical scale variations like road extraction. Classic works employing a. vision transformer as the backbone network for feature extraction include ViT [16], PVT [17], and Swin Transformer [18], among others. Subsequently, transformer-based models tailored for semantic segmentation tasks have emerged, including SETR [19], TransUNet [20], SegFormer [21], and CAS-Net [22], gradually confirming the feasibility of using transformer structures instead of CNNs to construct feature extraction modules for segmentation tasks. These versatile models demonstrate that transformer structures perform well in tasks such as image classification, segmentation, and detection. However, the development of transformer network designs more tailored to the characteristics of road features is still ongoing.

To address the obstacles encountered in road extraction tasks when using a vision transformer, this paper utilizes the shape and distribution characteristics of road features in remote sensing images as prior knowledge to guide the construction of the model. The motivation can be mainly divided into two parts:

(1) Due to the uneven scale issue of road features, this paper adopts the CSwin transformer [23] as the basis for the model architecture because the cross-shaped window partitioning of the CSwin transformer intuitively matches the shape of road features. How-

ever, this window partitioning method mechanically hinders the interaction of information between adjacent regions. Therefore, this paper introduces an overlapped window strategy to increase the shared information between neighboring windows, enhancing the model's perception of contextual information. This approach can retain the overall and continuous semantic information of roads to some extent, thereby improving the completeness of road segmentation results.

(2) Due to the low SNR of road features, the mechanism of self-attention involves fully connected operations on long sequences, allocating attention to all positions, thus making the model more sensitive to noise. Excessive background noise can interfere with the model's attention, preventing it from capturing the specific locations of target features. This paper adopts the deformable receptive field strategy, wherein the model's attention to patches can adaptively change with each input image, concentrating limited computational resources on foreground feature-relevant areas while reducing the intrusion of noise information.

Figure 1 provides a schematic illustration of how three classical transformer models and our method extract information from images. Taking a long vertical road as an example, in Figure 1a, no window partitioning is performed because, due to the complexity of the computations, the entire image cannot be input directly, and only very small patches can be cropped. In Figure 1b, square windows with symmetrical length and width clearly do not match the road morphology, with patches around the same coordinates on the road being separated into different windows under the square partitioning scheme. In Figure 1c, although the cross-shaped windows generally conform to the road morphology, they still cannot perform flexible and effective interaction calculations due to the fixed window range, and the windows cover a large amount of background area. In Figure 1d, which represents our method, the three patches are encompassed under the same deformable strip-shaped window, and the selected attention areas are concentrated on road information.

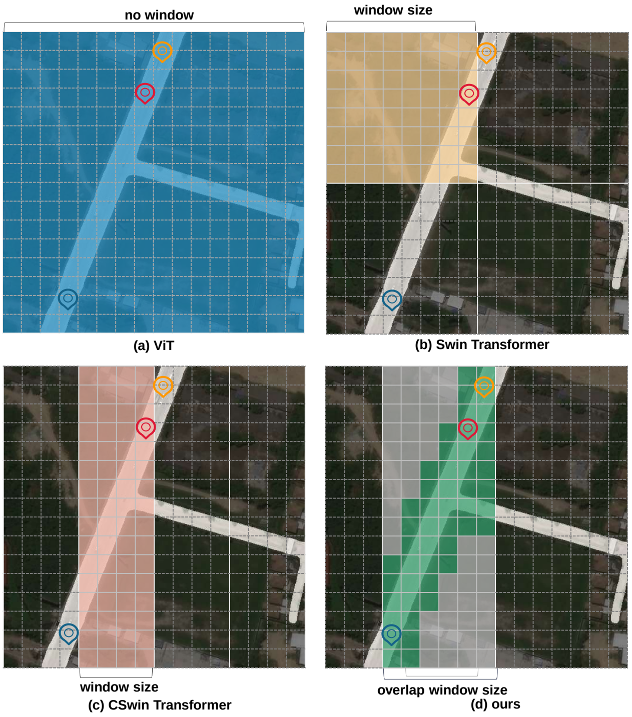

Figure 1. Comparison of the window schematics of our method with other classical transformer models. (**a**) ViT directly computes attention across the entire image. (**b**) The Swin transformer uses square-shaped window partitioning attention. (**c**) The CSwin transformer adopts cross-shaped window attention in both horizontal and vertical directions. (**d**) Ours incorporates overlap and deformable strategies into the cross-shaped windows.

The innovations of the proposed module can be summarized as follows:

- We propose an overlapped window, which uses key and value patches larger than the query patch during attention computation. It can facilitate information exchange between adjacent windows in both horizontal and vertical directions, completing road details effectively.
- The proposed deformable window introduces adaptive offset to flexibly resample key and value elements in attention computation, which can reduce attention allocation biases caused by excessive noise in the background. The purpose is to overcome the problem of the asymmetrical number of samples of road and background categories as much as possible.
- Our proposed network can achieve significant performance enhancement compared with state-of-the-art deep-learning-based road extraction methods on two popular datasets (i.e., DeepGlobe and Massachusetts datasets).

2. Related Works

2.1. Road Extraction Method

In addition to the Deep ResUnet and D-Linknet methods mentioned earlier, many other CNN-based road extraction methods have also addressed multi-scale issues and dense prediction tasks through improvements such as feature fusion and refinement. For instance, CADUNet [24] builds upon DenseUNet [25] by incorporating global attention modules and core attention modules, thereby improving road connectivity. Mosinskal et al. [26], aiming to ensure road topological connectivity, propose a topological loss as a replacement for the cross-entropy loss function. They continue to input road segmentation prediction results into the training model for iterative optimization to extract detailed features. The ATP-QDCNNRE [27] designed by Khan et al. utilizes quantum mechanical concepts and dilated convolutions to enhance the model's ability to capture long-range dependencies and employs automatic hyperparameter tuning to achieve road extraction. Ref. [28] proposed a fine-tuning network based on U-Net to preserve precise road geometry features, thereby improving edge detection effectiveness. Additionally, they applied the BRISQUE preprocessing technique to the dataset to enhance performance. Meanwhile, Tao et al. [29] analyzed these networks and found that blindly fusing multi-scale and multi-level features and expanding receptive fields could introduce irrelevant contextual information. Therefore, SIIS-Net is designed with a Spatial Information Inference Structure (SIIS) to better model context information along road-specific directions. Currently, DCS-TransUperNet [30] attempts to apply transformers to road extraction tasks. DCS-TransUperNet designs a dual-resolution branch encoder to extract coarse-grained and fine-grained features. Then, it uses Feature Fusion Modules (FFM) to merge the feature maps' output with the dual branches, thereby enhancing the feature representation with global dependencies. As it focuses solely on the improvement of vision transformer performance through multi-scale feature information, this paper aims to incorporate effective road context extraction methods into the vision transformer framework to enhance model performance.

2.2. Vision Transformer

The initial application of transformer technology to the visual domain was the ViT, which converted the entire image into a token sequence for self-attention computation. However, ViT only outputs fixed-resolution feature maps throughout the process, which is not conducive to fine-grained boundary segmentation tasks and incurs high computational complexity. Subsequent works, such as PVT [17], attempted to leverage various forms of feature pyramids to obtain multi-scale features and long-range information, which proved beneficial for tasks such as detection and segmentation. To address the issue of high computational costs associated with global attention computation in ViT, several local vision transformers emerged. These models, akin to convolution layers' local inductive bias in processing image information, partition patches into different local windows for subsequent self-attention computation. For instance, SWin Transformer utilizes block-based non-overlapping windows and introduces shift windows to enhance interaction between local windows. CSwin employs horizontal and vertical cross-shaped windows for self-attention and proposes a locally enhanced position encoding module. Moreover, some works aimed to judiciously integrate the strengths of both the convolution layer and transformer. For example, CvT [31] utilizes strided convolutions to reduce the token size and simplify self-attention computation. CvT inserts convolution layers between layers of the multi-level transformer network to supplement local image features. Drawing from these experiences, our approach tailors a sparse window attention method suitable for road extraction tasks and properly employs convolution modules in deformable self-attention to predict offsets, enabling more accurate acquisition of image information.

2.3. Deformable CNN and Attention

Due to previous methods like ViT, which directly partition the image into a token sequence based on fixed sizes and positions, this approach may disrupt the overall semantic information of the image and treat the foreground and background with equal attention, which is not conducive to dense recognition tasks. The main concept of Deformable CNN [32] is to concentrate receptive fields on the regions or targets of interest by resampling the spatial positions of input pixels. Some works have already incorporated the idea of deformability into different modules of transformer networks in various ways to enhance the attention paid to key areas. For instance, Deformable-DETR [33], aimed at object detection, employs a sparse spatial sampling attention module, which selects only a small group of important sampling points around reference points for calculation. PS-ViT [34] introduces an iterative progressive sampling module before the ViT backbone, which iteratively updates the attended foreground regions and then performs adaptive sampling. DePatch [35] is a plug-and-play module that flexibly segments samples in a deformable manner during the patch segmentation stage, reducing the disruption of fixed segmentation on semantic information. DAT [36] can be regarded as a spatial adaptive method, where the deformable self-attention module designed is added to the backbone network, forming a powerful and effective pyramid backbone network.

3. Method

This section introduces a new multi-stage transformer segmentation network tailored for road features called DOCswin-Trans (Deformable and Overlapped Cross-Window Transformer), which follows an encoder-decoder structure, as illustrated in Figure 2. In the following, the encoder and decoder will be introduced.

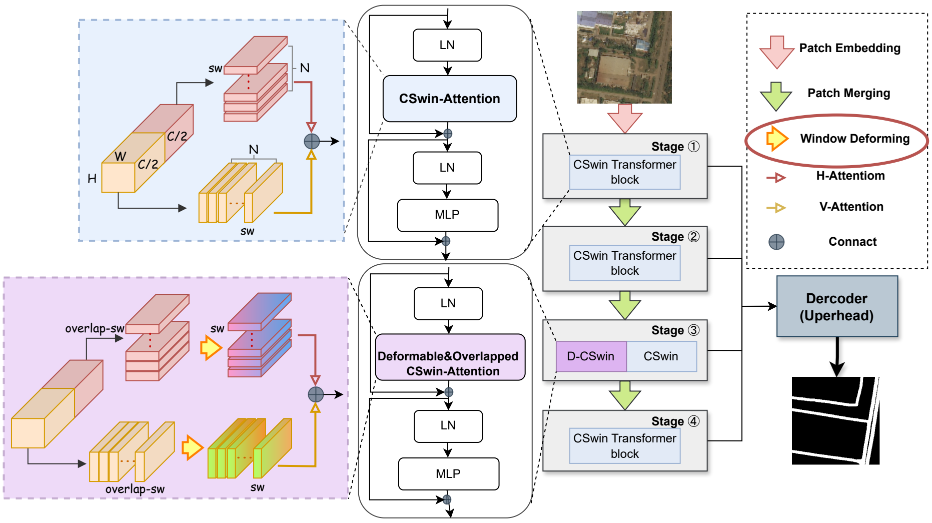

Figure 2. The overall structure of the developed method.

3.1. Encoder

The encoder consists of four stages and is illustrated in Figure 2. In order to maintain the stability of the model in the early stage and to avoid the new module disrupting the extraction of background information from the model, the CSWin-Transformer Block and DOC-Transformer Block are used interchangeably in the third stage only, and the basic CSWin-Attention is still used in the rest of the stages for feature extraction, which can balance and supplement the complete road and background features in time. This section will detail these two blocks, respectively.

3.1.1. Cswin-Transformer Block

As shown in Figure 2, the CSWin-Transformer Block consists of multi-layer perceptron (MLP) and cross-shaped window self-attention (CSWin-Attention). Additionally, layer normalization (LN) and residual connections are incorporated to ensure training stability. Its mathematical representation is as follows:

$$\hat{X}_l = CSWin - Attention(LN(X_{l-1})) \tag{1}$$

$$X_l = MLP(LN(\hat{X}_l)) + \hat{X}_l \tag{2}$$

The core part leveraged to extract features in the CSWin-Transformer block is the CSWin-Attention, which can leverage sparse attention to reduce the computational complexity of self-attention for long sequence inputs. More importantly, for the segmentation of the roads, this window partitioning method ensures that each patch participates in the calculation range extending in both horizontal and vertical directions. This aligns with the strip-like distribution characteristics of road features, allowing for more reasonable and effective feature extraction within a suitable range.

The self-attention layer operates by transforming the input X into three new matrices (Q, K, V) through linear transformations (W^Q, W^K, W^V) applied to query, key, and value. Subsequently, it calculates scores using scaled dot-product attention, indicating the relevance.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^o \tag{4}$$

The multi-head attention layer enhances the self-attention layer by adding the ability to focus on different positions, thereby providing a more comprehensive mapping representation of the input sequence. This helps the model capture richer information. In multi-head attention, multiple sets (denoted as) of Q, K, and V matrices are defined to focus on different positions and perform calculations. Output results are obtained for each set of Q, K, and V, and after concatenation and linear transformation of the output results from each set, the final result is obtained.

CSWin-Attention divides the input feature into two parts along the channel dimension C. One part performs the horizontal strip self-attention computation, and the other part performs the vertical strip self-attention computation, which ensures the completeness of the feature extraction in the horizontal and vertical scales. Finally, the results of the two groups are concatenated along the channel dimension C to obtain the feature vector.

$$CSWin - Attention(X) = Concat(H - Attention, V - attention)W^o \tag{5}$$

$$where\begin{cases} H - Attention(X_H), X_H = X[1:C/2] \\ V - Attention(X_V), X_V = X[C/2:C] \end{cases} \tag{6}$$

The following formulas all use horizontal (H) stripe self-attention computation as an example, with vertical (V) stripe self-attention being similar. In the H direction, the input vector is divided into several equally wide stripe window segments. Then, multi-head self-attention computation is performed only within the window range for the patches contained inside. Different widths of window segmentation are set in each stage, with the stripe width increasing with the stage. This helps the model capture road features of different scales, adapting to road entities of different thicknesses. Here, $N = H/sw_2$ represents the number of stripe windows, and F_H^i denotes the feature vectors generated after self-attention computation for the i_{th} horizontal stripe window. The results obtained from each stripe window are concatenated sequentially to obtain the complete feature map.

$$X_H \longrightarrow [X_H^1, X_H^2, ..., X_H^N], \tag{7}$$

$$F_H^i = Attention(X_H^i W^Q, X_H^i W^K, X_H^i W^V) \tag{8}$$

$$H - Attention(X_H) = Concat[F_H^1, F_H^2, ...F_H^N] \tag{9}$$

Compared to the square windows utilized in other vision transformer methods, the horizontal and vertical stripe windows designed in CSwin are aligned with the distribution of road shapes. However, there are opportunities for improvement in the detailed and accurate extraction of road features. Therefore, in the next section, the DOC-Transformer block with overlapped and deformable strategies will be proposed to improve the CSWin-Transformer block in these aspects.

3.1.2. DOC-Transformer Block

The proposed DOC-Transformer block structure is similar to the CSWin-Transformer block. Compared with the CSWin-Attention, our developed DOC-Transformer block has two efficient modifications. The first is the overlapped windows, which use windows with overlapping edges to generate K and V. The second is the deformable windows, which are realized by resampling the feature vectors using an offset prediction network. These two windows are detailed in the following.

Overlapped Window: Overlapped attention is used in both MOA [37] and HaloNet [38], verifying that local attention using K and V slightly larger than Q improves the performance at the same time while saving computation costs. The implementation of overlapping windows involves partitioning the input features into two types of strip windows with different sizes, denoted as X_{H_1} and X_{H_2}. Here, X_{H_2} strip width is slightly wider than X_{H_1}. As demonstrated in the horizontal window of Figure 3, the center of X_{H_1} and X_{H_2} are aligned, and the effect of the overlapping region is formed in the upper and lower sides of the window, where X_{H_2} is overlapped with respect to X_{H_1}. The overlap operation can obtain the boundary information of the neighboring windows and better preserve the local continuity of the input features, it enhances self-attention to better focus on the positional information of roads within the inputs and then addresses the challenge of the road's low SNR. In order to ensure that the overlap degree is reasonable and does not confuse the extraction of the features, it is stipulated that Q is mapped from X_{H_1} without overlap, while K and V are mapped from $\tilde{X_{H_2}}$.

$$X_H \longrightarrow X_{H_1} = [X_{H_1}^1, X_{H_1}^2, ..., X_{H_1}^{N_1}], X_{H_1} = [X_{H_2}^1, X_{H_2}^2, ..., X_{H_2}^{N_2}] \tag{10}$$

$$Q = X_{H_1} W^Q, K = \tilde{X_{H_2}} W^K, V = \tilde{X_{H_2}} W^V \tag{11}$$

where $N_1 = H/sw_1, N_2 = H/sw_2$ are the number of windows, $\tilde{X_{H_2}}$ is the result of X_{H_2} after the resampling operation, and the specific process of offset acquisition and resampling will be described in detail in the next section.

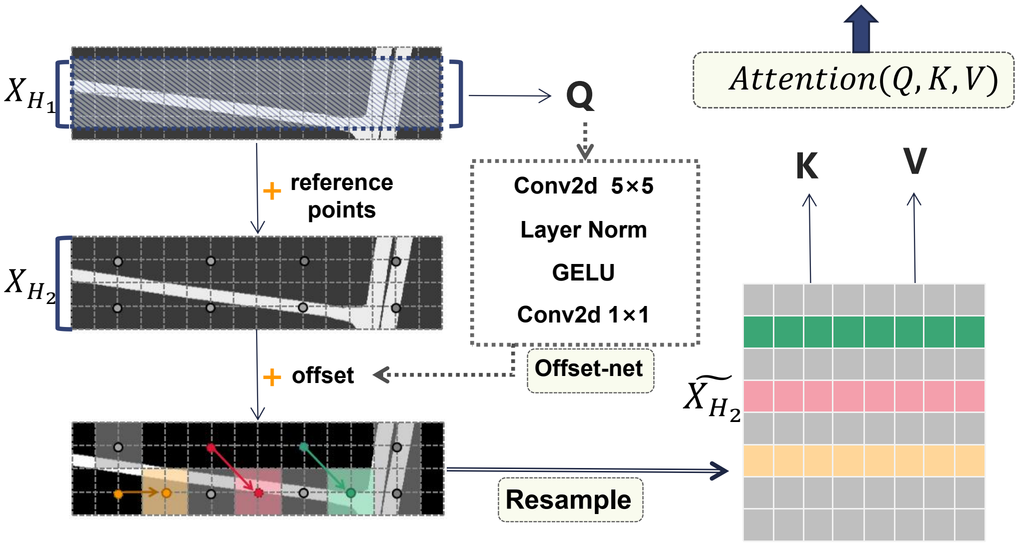

Figure 3. Detail structure of the deformable window.

Deformable Window: The module structure in Figure 3 shows the details of the deformable window in the red circle in the right legend in Figure 2. Its core components consist of offset generation and vector resampling. In this paper, it is assumed that relying on the self-learning ability of the offset prediction network, the feature information contained in Q can be integrated through this network to find the most suitable and optimal sampling region, and the whole process can be described as: the grid reference point predicts the offset based on Q, and then the focus area, which is the road-related region, is screened out by resampling in the window with a larger background. Firstly, the standard reference grid coordinates $Ref \in R^{H_G \times W_G \times 2}$, with equal intervals, are generated in the strip window of size $sw \times H$ (the reference grid points are sparsely plotted in the figure for the convenience of demonstrating the process), which can be expressed as a series of $2D$ coordinates, and then all the $2D$ coordinates are normalized to the range of $[-1,+1]$, where $(-1,-1)$ represents the top-left corner of the grid, and $(+1,+1)$ represents the bottom-right corner, facilitating subsequent resampling operations. Subsequently, Q is input into the offset prediction network composed of two convolutional layers and the activation function to obtain the offset $Off \in R^{H_G \times W_G \times 2}$, and then added to the standard reference grid coordinates to obtain the new offset coordinates $Pos \in R^{H_G \times W_G \times 2}$. Finally, using Pos as the coordinate index to resample X_{H_2} by bilinear interpolation, which can be interpreted as the use of each point of the 4 nearest cell values of the neighborhood to be computed, and finally, obtain the reconstructed vector $\tilde{X_{H_2}}$.

$$Off = offset - network(Q) \tag{12}$$

$$Pos = Ref + Off \tag{13}$$

$$\tilde{X_{H_2}} = Re - sample(X_{H_2}, Pos) \tag{14}$$

- **Offset network:**The middle part of Figure 3 shows the offset prediction network. Its specific implementation includes a nonlinear activation function GELU and two convolution structures, in which 5×5 convolution is used to extract global and local features, and the role of the 1×1 convolution kernel is to adjust the output dimensionality to 2 dimensions as the offset in the two directions of (x, y).
- **Re-sample:** The purpose of re-sample is to reconstruct the feature vector based on the offset coordinates, and since the probability will be out of the integer coordinate points, we employ bilinear interpolation. This involves linear interpolation separately in the horizontal and vertical directions, based on the values of the four-neighboring pixels. This operation involves the distance from each point as a weight in the calculation,

which ensures a reasonable correlation of the predicted vector values in terms of spatial location.

3.2. Decoder

The decoder adopts the Uperhead architecture shown in Figure 4, which is composed of a pyramid pooling module and a multi-scale feature pyramid. Feature maps of different dimensions generated from the four stages are simultaneously inputted into the decoder. The multi-level feature maps are upsampled and fused in the decoder to generate the final prediction.

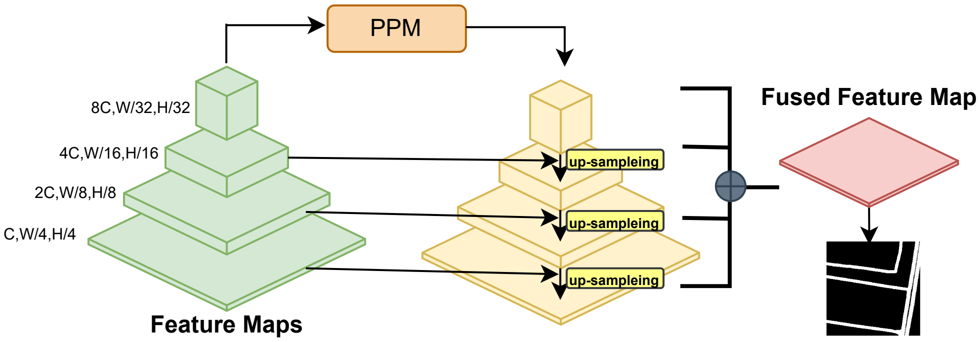

Figure 4. Detail structure of the decoder architecture.

4. Results

4.1. Datasets

The high-resolution remote sensing road extraction datasets used in this study are two classic open-source datasets: the DeepGlobe dataset [39] and the Massachusetts road dataset [40].

The DeepGlobe dataset consists of remote sensing images with a resolution of 0.5 m, containing various types of roads in urban and rural scenes. It comprises a total of 8570 images of size 1024 × 1024 pixels, with only 6226 images annotated for road segmentation. Following the common practice, we randomly split the dataset into training and validation sets in an 8:2 ratio and then cropped them in a non-overlapping way to a size of 512 × 512 pixels. In total, 19,924 images were used for training, while 4980 images were used for validation.

The Massachusetts road dataset consists of 1-meter resolution remote sensing images from Massachusetts, USA. Unlike the full road coverage in the DeepGlobe dataset, the labeling approach here involves expanding road centerlines downloaded from OpenStreetMap to approximate real road masks with a width of 7 pixels. The dataset comprises 1171 images of size 1500 × 1500. To obtain more data samples and enhance sample diversity from the limited original images, we cropped them into 512 × 512 image blocks using a sliding window approach with a stride of 256. Due to the presence of irregular areas of white space with no features in some original images, we filtered out ineligible blank images, resulting in 14,420 training images and 1008 testing images.

4.2. Evaluation Metrics

Evaluation metrics are used to judge the performance of a model, and we utilized classic pixel-level evaluation metrics, including IoU, F1-score, precision, and recall. IoU is the ratio of the intersection and union of the road class pixels in the predicted and GT, recall can reflect the completeness of the road extraction by the model, and precision is used to characterize the correctness of the road extraction result, usually recall and precision will be constrained by each other, and the F1-score can be used as a comprehensive consideration of the above two.

For the strong class imbalance issue in this task, the calculation details are as follows: aggregate and accumulate TP, FP, and FN from all images in the test set, as shown in Equation (15), where there are a total of m images and i denotes the i_{th} image. These values are then used to compute precision and recall to avoid significant biases caused by too few positive samples. Finally, the F1-score is calculated using Equation (18).

$$TP = \sum_{1}^{m} TP^{(i)}, FP = \sum_{1}^{m} FP^{(i)}, FN = \sum_{1}^{m} FN^{(i)} \tag{15}$$

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{18}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{19}$$

4.3. Experimental Details

The experiments were conducted on NVIDIA RTX A6000 (NVIDIA, Santa Clara, CA, USA), and the model was trained using the AdamW optimizer with an initial learning rate of 0.0001 and momentum parameters of 0.9 and 0.999. The learning rate strategy is "poly", and the warm-up mechanism was used to warm up the model and decay the model according to the polynomials in later iterations. The batch size was 8, and the total epochs was 100.

4.4. Result on Deepglobe Dataset

To evaluate the performance of our road extraction model on the DeepGlobe dataset, we selected several high-performing segmentation networks as references, including the classic model of road extraction D-Linknet and the CNN-based semantic segmentation network Deeplabv3+, as well as the transformer-based backbone networks Swin and Cswin (using Uperhead as the decoder). Comparison with the multiple types of networks above ensures the comprehensiveness of the experimental control. Table 1 demonstrates a series of evaluation metrics to determine the accuracy of different models on the DeepGlobe dataset. Our method (referred to as "Ours") has a 0.63% to 5.01% improvement on IoU compared to other methods, and 0.47% to 3.68% on F1. The improvement of recall is more obvious when compared with Cswin, which can indicate that our model improves the capture rate of positive samples of roads in the case of an extreme sample imbalance between foreground and background.

Table 1. Quantitative analysis results of different models on the DeepGlobe dataset.

Method	Precision	Recall	F1	IoU	mPrecision	mRecall	mF1	mIoU
D-Linknet	77.53	73.30	75.36	60.46	88.21	86.21	87.18	79.24
Deeplabv3+	79.84	71.17	75.26	60.33	89.32	85.21	87.14	79.20
Swin + Uperhead	82.42	74.30	78.15	64.14	90.68	86.82	88.64	81.21
Cswin + Uperhead	82.06	75.37	78.57	64.71	90.52	87.34	88.86	81.50
Ours + Uperhead	82.24	76.08	79.04	65.34	90.62	87.69	89.10	81.83

Scenario Analysis To validate the performance ability of the model in different scenes, we divide the validation set images into multiple scenes after observing the overall situation of the DeepGlobe dataset: rural areas, town areas, and urban concentrated areas. In Figure 5, these 5 sets of images come from three scenarios that are both representative and diverse.

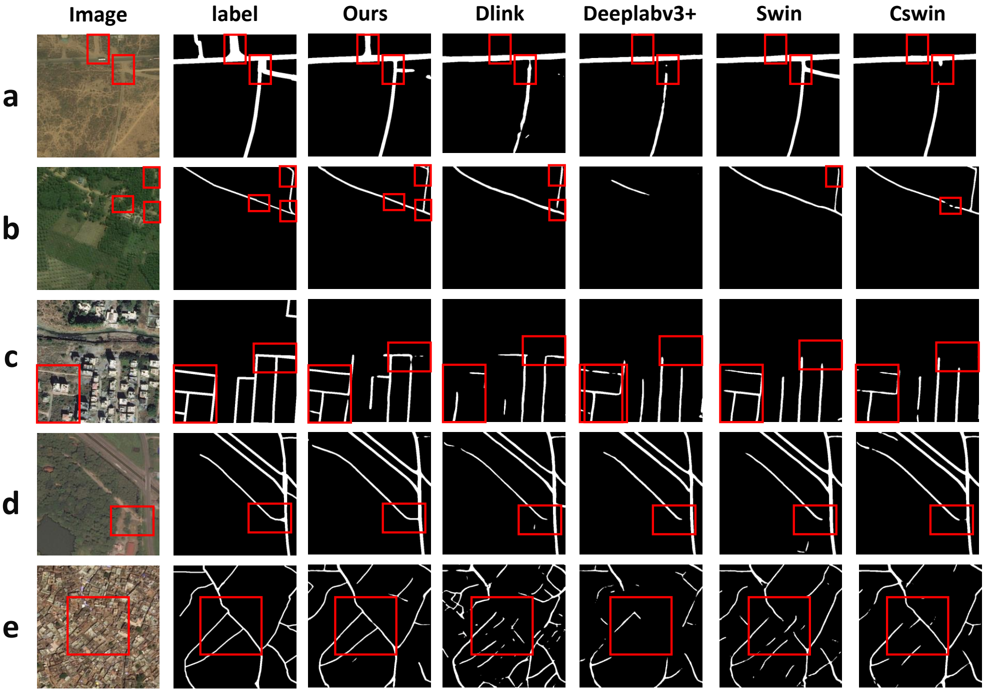

Figure 5. Visual analysis of road segmentation results of different models on the DeepGlobe dataset. (**a**,**b**) are the roads in rural areas and the background of (**a**) is a large area of wilderness, and the background of (**b**) is a forested area. (**c**,**d**) are the roads in town areas, and the main body of (**c**) is a small path next to the building, the main body of (**d**) is a main traffic artery, and (**e**) is the roads in concentrated urban areas. The area highlighted in red boxes shows regions with noticeable road discontinuities.

In rural areas (Figure 5a,b), the road structure is relatively single, and the background mainly consists of wasteland, cultivated land, and woodland. The difficulty of recognition in (a) arises from the similarity in texture between the road and the wasteland, which leads to the omission of detection. A short road in the upper part of the image in (a) is ignored by the other models, and there is a break in the extraction of road intersections by Dlink, Deeplabv3+, and Cswin. (b) reveals occlusion issues, and the extraction results of DOCswin-Trans(ours) are the most complete. Comparing DOCswin-Trans with Cswin, it is evident that DOCswin-Trans supplements the obscured parts caused by vegetation, ensuring road connectivity.

In town areas (Figure 5c,d), road structures are more complex, leading to issues like "the same thing different spectrum" and "the same spectrum foreign matter", which may lead to challenges in identification due to spectral similarities between roads and objects, like buildings or parking lots. Roads in building gaps in (c) were more difficult to extract, while DOCswin-Trans also failed to identify all roads, but there was some improvement over the other methods.

In concentrated urban areas (Figure 5e), roads are intertwined and complex, and the interference is exacerbated due to the high density of buildings, making it difficult even for human eyes to judge. This increases the difficulty of continuous road extraction. The results of other models are highly fragmented, while DOCswin-Trans extracts the main roads and ensures overall connectivity, but there are still some false detections present.

Table 2 shows the evaluation measures mIoU and mF1 in each scenario, and the line chart is drawn as shown in Figure 6. Firstly, it can be visualized that the recognition difficulty in town areas is higher than that in rural areas, while the accuracy in concentrated urban areas is the lowest, which is consistent with the above analysis of the recognition difficulty. Additionally, it is notable that DOCswin-Trans shows a larger improvement in densely urban areas. The comparison images in (e) also indicate that the enhancements made by our method may be more applicable to concentrated urban areas. The proposed modules manage to capture road details even in highly complex backgrounds.

Table 2. Scenario analysis results of different models on the DeepGlobe dataset.

Method	Rural Area		Town Area		Urban Concentrated Area	
	mIoU	mF1	mIoU	mF1	mIoU	mF1
D-Linknet	80.01	87.62	78.78	86.91	77.23	86.18
Deeplabv3+	79.61	87.31	78.99	87.06	76.95	85.94
Swin+Uperhead	81.87	89.01	80.75	88.36	79.68	87.93
CSWin+Uperhead	82.06	89.15	81.08	88.61	79.98	88.14
Ours+Uperhead	82.24	89.28	81.34	88.8	80.69	88.64

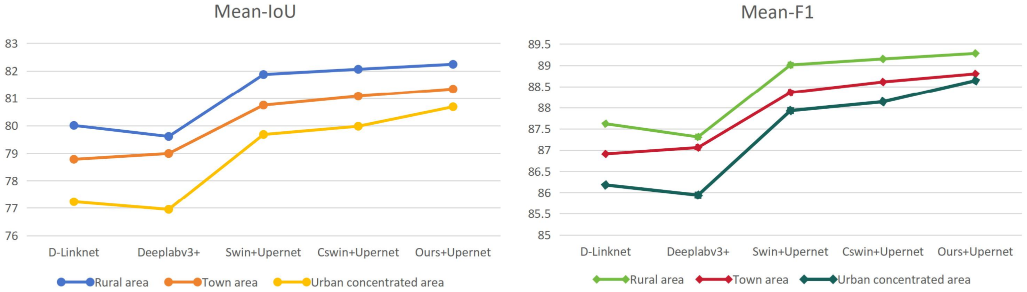

Figure 6. Line chart of scene analysis results on the DeepGlobe dataset.

4.5. Result on Massachusetts Road Dataset

The image scenes in the Massachusetts road dataset also include rural and urban areas, but the overall differences are smaller compared to the DeepGlobe dataset. Additionally, the ground truth (GT) annotations are derived from OpenStreetMap and rasterized to a uniform road width of 7 pixels, which may not align with the actual road widths in the original images. Consequently, improvements in accuracy metrics may be subject to some disturbance, but ensuring consistency across models still allows for experimental comparisons. The evaluation metric results for road extraction methods on the Massachusetts road dataset are presented in Table 3. Our method shows improvements of 0.50% to 6.24% in IoU compared to other methods, and enhancements of 0.39% to 5.05% in F1-score.

Table 3. Quantitative analysis results of different models on the Massachusetts roads dataset.

Method	Precision	Recall	F1	IoU	mPrecision	mRecall	mF1	mIoU
D-Linknet	79.86	62.52	70.14	54.01	88.93	80.83	84.35	75.59
Deeplabv3+	79.53	64.96	71.51	55.65	88.82	82.03	85.06	76.45
Swin + Uperhead	81.3	67.21	73.59	58.21	89.77	83.19	86.14	77.82
CSWin + Uperhead	79.15	70.91	74.80	59.75	88.79	84.95	86.76	78.60
Ours + Uperhead	80.37	70.64	75.19	60.25	89.40	84.85	86.97	78.88

The comparison of prediction results for each model is shown in Figure 7. In Figure 7a, a segment of the road in the top-left corner is entirely obscured by vegetation, yet the road morphology is preserved, and its continuation can be inferred along the extension of surrounding roads.

DOCswin-Trans supplements the occluded portions by leveraging its ability to capture contextual information along both horizontal and vertical scales. It even correctly identifies road areas not labeled in the ground truth. The vertical road on the right side in Figure 7b and the circular road in Figure 7c are only partially detected by D-Linknet, Deeplabv3+, and Swin, while they are completely ignored by CSwin.

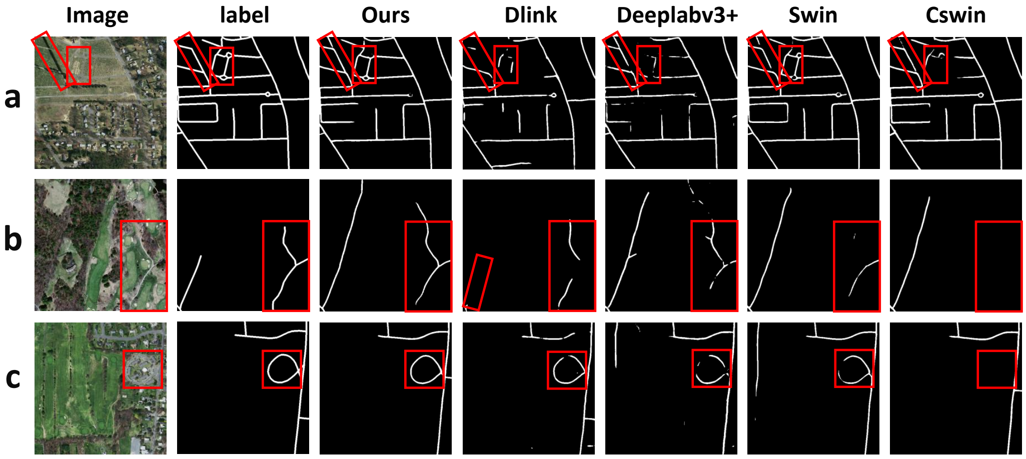

Figure 7. Visual analysis of road segmentation results of different models on the Massachusetts roads dataset. (**a**–**c**) represent road segmentation results of subareas with different networks, respectively. The area highlighted in red boxes shows regions with noticeable road discontinuities.

4.6. Parameter Experiment

Model performance is closely related to parameter settings. To evaluate the impact of the number of modules on the model, we designed the following two sets of experiments conducted on the DeepGlobe Dataset. The training configuration remains consistent with the experimental details in Section 4.3 of the experiment.

(1) Model variant: The CSwin transformer builds some variants with different parameter settings such as the number of blocks, the attention head numbers, and the channel dimension, etc. The detailed parameters for each variant are provided in Table 4. In this experiment, different variants of the CSwin transformer were applied as the base encoder to test their performance on the road extraction task.From Table 5, it can be seen that the performance of the model increases with the number of parameters and the depth of the architecture. However, blindly pursuing larger hidden layers, attention head numbers, and other parameters does not guarantee a significant improvement in accuracy. This also depends on factors such as the size of the training dataset and hyperparameter settings. In practical applications, different variants of the model can be chosen based on the task complexity, requirements for model accuracy, and deployment considerations.

Table 4. Detailed configurations of different variants of the CSwin transformer.

Method	Hidden Size	Block Number	Head Number	Window Size	Param
CSWin-tiny	64	1, 2, 21, 1	2, 4, 8, 16	1, 2, 7, 7	23 M
CSWin-small	64	2, 4, 32, 2	2, 4, 8, 16	1, 2, 7, 7	35 M
CSWin-base	96	2, 4, 32, 2	4, 8, 16, 32	1, 2, 7, 7	78 M

Table 5. Quantitative analysis results of different model variants on the DeepGlobe dataset.

Method	Precision	Recall	F1	IoU	mPrecision	mRecall	mF1	mIoU
CSWin-tiny	80.61	75.07	77.74	63.59	89.79	87.16	88.42	80.91
CSWin-small	80.97	76.43	78.63	64.79	89.99	87.84	88.88	81.54
CSWin-base	82.06	75.37	78.57	64.71	90.52	87.34	88.86	81.50
Ours-tiny	80.43	75.71	78.00	63.93	89.71	87.47	88.55	81.08
Ours-small	81.37	76.82	79.03	65.33	90.20	88.04	89.09	81.82
Ours-base	82.24	76.08	79.04	65.34	90.62	87.69	89.10	81.83

(2) DOC-Number: To balance the model's attention between foreground and background, we replaced some of the Cswin-Transformer blocks with DOC-Transformer blocks only in the third stage of the model. To assess the impact of the replaced module count (DOC-Number) on the model's performance, we designed the following 5 sets of comparative experiments. There are a total of 32 blocks in the third stage, and DOC-Transformer Blocks are inserted at regular intervals among them. The results are shown in Table 6. It can be observed that as the DOC-Number increases, the accuracy exhibits a trend of initially increasing and then decreasing. Additionally, the results are optimal when the parameter setting is 8, and they are worst when the parameter setting is 32. This indicates that when fewer modules are replaced (DOC-Number = 0 or 4), there is still room for improvement in the model's attention to road features, suggesting that the model's performance has not been maximally enhanced. However, when too many modules are replaced (DOC-Number = 16 or 32), the accuracy tends to slightly decrease. This could be due to the model relying too heavily on deformable attention mechanisms, leading to insufficient attention to background features and resulting in misjudgments. Moreover, the increased number of replaced modules introduces redundancy in parameters, affecting the model's inference speed. Table 6 also provides statistics on the inference time of each model. Considering both the improvement in accuracy and the impact of parameter quantity on model performance, we have selected parameter setting 8 as the final design for the model architecture.

Table 6. Experiment results of DOC-Number setting on the DeepGlobe dataset.

Method	DOC-Number	Precision	Recall	F1	IoU	Inference
Ours-base	0	82.06	75.37	78.57	64.71	0.255
	4	81.61	76.45	78.94	65.21	0.328
	8	82.24	76.08	79.04	65.34	0.373
	16	81.57	75.71	78.53	64.65	0.515
	32	81.46	70.0	75.29	60.38	0.736

4.7. Ablation Study

In order to better understand the performance of the two modified modules in the road extraction task, we designed two ablation experiments to evaluate the effectiveness of each component. The experiments were conducted on the DeepGlobe dataset, and the training configurations remained consistent with the experiments described earlier.

As shown in Table 7, retaining only the deformable window resulted in a 0.34% improvement in IoU compared to CSWin. This experiment illustrates that after the feature vectors undergo offset resampling in the deformable attention module, certain complex and redundant background feature information can be ignored. This process retains more representative information related to road features, aiding the self-attention mechanism in recognizing the heterogeneity of road information and addressing the issue of low signal-to-noise ratio in road data.

Table 7. Quantitative analysis of the ablation experiments on the deformable window and overlapped window modules.

Method	Deformable Window	Overlapped Window	Precision	Recall	F1	IoU
CSWin-base	×	×	82.06	75.37	78.57	64.71
Ours-base	✓	×	81.58	76.25	78.82	65.05
Ours-base	×	✓	80.80	76.88	78.79	65.00
Ours-base	✓	✓	82.24	76.08	79.04	65.34

Retaining only the overlapped window setting resulted in a 0.29% improvement in IoU compared to CSWin, and there was a significant increase in recall. This suggests that this strategy enhances the model's recall rate for road-positive samples, thereby improving its ability to capture positive sample features. This ablation experiment demonstrates that overlapping windows enhance the model's ability to supplement information between adjacent windows. By providing closely related contextual information, it improves the completeness of the identification of road-class pixels and alleviates the issue of uneven horizontal and vertical scales in road data.

4.8. Heat Map Visualization

To intuitively observe and understand how the model attends to different parts of the inputs to enhance interpretability and accuracy, we utilized Grad-CAM to generate activation maps. This allowed us to observe how different models concentrate their attention on the road foreground and other background elements. Grad-CAM computes the gradients of the feature maps with respect to the classification output and multiplies them by the weight matrices calculated by the model. The resulting activation maps can demonstrate the contribution of each pixel position to the output probability. Then, we designate warmer-colored regions as areas where the model pays higher attention.

As shown in Figure 8, it can be observed that Swin and Cswin models exhibit high sensitivity not only to road areas but also to building regions. This might be due to the interference from the problem of different objects sharing similar spectra. Additionally, the highlighted areas in Swin appear more scattered, possibly because the block-wise windowing approach does not adapt well to the distribution characteristics of roads, resulting in a lack of continuous activation areas around roads. The performance of Cswin is slightly better than Swin, as it forms highlighted areas around roads, but the activation level in background areas remains relatively high, without focusing on foreground parts. The proposed model, on the other hand, to some extent shifts attention to road-related areas. The darkening of colors around roads indicates increased attention, while the tendency towards blue in background areas suggests that attention is suppressed there.

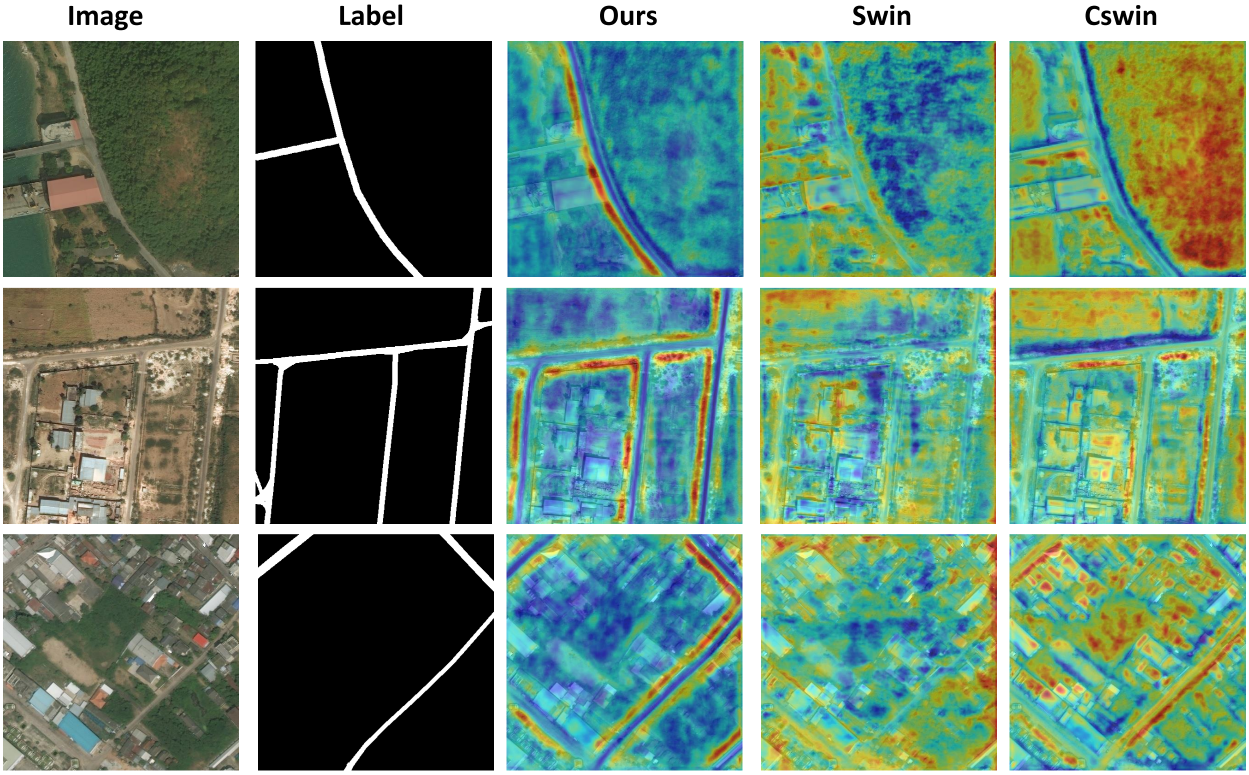

Figure 8. Heat map visualization of different models on DeepGlobe Dataset.

5. Discussion

5.1. Conclusions

This paper proposes a transformer-based semantic segmentation network called DOCswin-Trans for road extraction tasks in high-resolution remote sensing images, which mainly focuses on the prior knowledge of roads. Our work aims to design rational modules to overcome the asymmetry in scale and sample quantity of roads as much as possible. We use horizontal and vertical bar windows in the transformer framework to more reasonably partition the regions for multi-head self-attention computation and use the overlapped window strategy to obtain continuous contextual information about the road, solving the problem of road scale imbalance. We add the deformable self-attention module to let the model focus its attention more on the patches containing road information, which can be used to balance the gap between the foreground and the background in terms of the amount of information of the samples, i.e., to solve the problem of the road's low SNR.

We conducted experiments to validate the proposed model on two road datasets and compared it with several baseline methods. Both qualitative and quantitative analyses were performed using accuracy evaluation metrics and prediction maps. We designed ablation experiments specifically targeting the overlapped window strategy and deformable window strategy integrated into the model to validate their effectiveness. Additionally, visualization through heatmaps confirmed that our method improved the model's perception of road-related pixels. Since road image data may vary due to geographical factors, we also designed scene analysis experiments to assess each model's performance in different scenarios. Our findings indicate that DOCswin-Trans demonstrated the most significant optimization effect in dense urban areas, with performance improvements also observed in rural and town areas. Through these experiments, we demonstrated that the modules designed in this method enhance the model's ability to extract road features.

5.2. Future Directions

The approach presented in this paper does not address the optimization of model computational complexity. However, the computational complexity of the model is crucial for its deployability in real-world applications, especially in scenarios such as traffic safety and emergency response, where real-time road extraction is essential. Future work aims to develop lighter-weight transformer-based models to accelerate inference speed. Furthermore, the current model heavily relies on annotated samples, leading to challenges in generalizing to unknown domains. In future work, strategies such as generative learning and self-training will be explored to tackle domain adaptation tasks, enabling the trained model to be applied to a wider range of road data.

Author Contributions: Conceptualization, L.Z. and J.Z.; methodology, L.Z. and J.Z.; software, L.Z. and W.Z.; validation, J.Z., W.Z., Z.Z. and C.P.; formal analysis, W.Z. and C.P.; resources, L.Z. and J.Z.; data curation, C.P. and Z.Z.; writing—original draft preparation, L.Z.; writing—review and editing, J.Z., W.Z., Z.Z., X.M. and C.P.; visualization, C.P.; supervision, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (Grant No.: 42171458), and Tianjin Key Laboratory of Rail Transit Navigation Positioning and Spatio-temporal Big Data Technology.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

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

References

1. Cheng, G.; Zhu, F.; Xiang, S.; Pan, C. Road centerline extraction via semisupervised segmentation and multidirection nonmaximum suppression. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 545–549. [CrossRef]
2. Song, Y.; Ju, Y.; Du, K.; Liu, W.; Song, J. Online road detection under a shadowy traffic image using a learning-based illumination-independent image. *Symmetry* **2018**, *10*, 707. [CrossRef]

3. Abdollahi, A.; Pradhan, B.; Alamri, A. VNet: An end-to-end fully convolutional neural network for road extraction from high-resolution remote sensing data. *IEEE Access* **2020**, *8*, 179424–179436. [CrossRef]
4. Singh, P.P.; Garg, R.D. Automatic road extraction from high resolution satellite image using adaptive global thresholding and morphological operations. *J. Indian Soc. Remote Sens.* **2013**, *41*, 631–640. [CrossRef]
5. 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]
6. Shanmugam, L.; Kaliaperumal, V. Junction-aware water flow approach for urban road network extraction. *Iet Image Process.* **2016**, *10*, 227–234. [CrossRef]
7. Mu, H.; Zhang, Y.; Li, H.; Guo, Y.; Zhuang, Y. Road extraction base on Zernike algorithm on SAR image. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1274–1277.
8. Singh, P.P.; Garg, R.D. A two-stage framework for road extraction from high-resolution satellite images by using prominent features of impervious surfaces. *Int. J. Remote Sens.* **2014**, *35*, 8074–8107. [CrossRef]
9. Xu, G.; Zhang, D.; Liu, X. Road extraction in high resolution images from Google Earth. In Proceedings of the 2009 7th International Conference on Information, Communications and Signal Processing (ICICS), Macau, China, 8–10 December 2009; pp. 1–5.
10. Ali, I.; Rehman, A.U.; Khan, D.M.; Khan, Z.; Shafiq, M.; Choi, J.G. Model selection using K-means clustering algorithm for the symmetrical segmentation of remote sensing datasets. *Symmetry* **2022**, *14*, 1149. [CrossRef]
11. Miao, Z.; Wang, B.; Shi, W.; Zhang, H. A semi-automatic method for road centerline extraction from VHR images. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 1856–1860. [CrossRef]
12. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual u-net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
13. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 182–186.
14. Chaurasia, A.; Culurciello, E. Linknet: Exploiting encoder representations for efficient semantic segmentation. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017; pp. 1–4.
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
16. 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 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
17. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 568–578.
18. 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, BC, Canada, 11–17 October 2021; pp. 10012–10022.
19. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, DC, USA, 14–19 June 2020; pp. 6881–6890.
20. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
21. 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.
22. Yang, Z.; Wu, Q.; Zhang, F.; Zhang, X.; Chen, X.; Gao, Y. A New Semantic Segmentation Method for Remote Sensing Images Integrating Coordinate Attention and SPD-Conv. *Symmetry* **2023**, *15*, 1037. [CrossRef]
23. 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. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12124–12134.
24. Li, J.; Liu, Y.; Zhang, Y.; Zhang, Y. Cascaded attention DenseUNet (CADUNet) for road extraction from very-high-resolution images. *Isprs Int. J.-Geo-Inf.* **2021**, *10*, 329. [CrossRef]
25. Cao, Y.; Liu, S.; Peng, Y.; Li, J. DenseUNet: Densely connected UNet for electron microscopy image segmentation. *Iet. Image Process.* **2020**, *14*, 2682–2689. [CrossRef]
26. Mosinska, A.; Marquez-Neila, P.; Koziński, M.; Fua, P. Beyond the pixel-wise loss for topology-aware delineation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3136–3145.
27. Khan, M.J.; Singh, P.P.; Pradhan, B.; Alamri, A.; Lee, C.W. Extraction of Roads Using the Archimedes Tuning Process with the Quantum Dilated Convolutional Neural Network. *Sensors* **2023**, *23*, 8783. [CrossRef]
28. Khan, M.J.; Singh, P.P. Advanced road extraction using CNN-based U-Net model and satellite imagery. *Prime-Adv. Electr. Eng. Electron. Energy* **2023**, *5*, 100244. [CrossRef]
29. Tao, C.; Qi, J.; Li, Y.; Wang, H.; Li, H. Spatial information inference net: Road extraction using road-specific contextual information. *ISPRS J. Photogramm. Remote Sens.* **2019**, *158*, 155–166. [CrossRef]

30. 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]
31. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 22–31.
32. 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; pp. 764–773.
33. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
34. Yue, X.; Sun, S.; Kuang, Z.; Wei, M.; Torr, P.H.; Zhang, W.; Lin, D. Vision transformer with progressive sampling. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 21–27 October 2021; pp. 387–396.
35. Chen, Z.; Zhu, Y.; Zhao, C.; Hu, G.; Zeng, W.; Wang, J.; Tang, M. Dpt: Deformable patch-based transformer for visual recognition. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 2899–2907.
36. Xia, Z.; Pan, X.; Song, S.; Li, L.E.; Huang, G. Vision transformer with deformable attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4794–4803.
37. Patel, K.; Bur, A.M.; Li, F.; Wang, G. Aggregating global features into local vision transformer. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 1141–1147.
38. Vaswani, A.; Ramachandran, P.; Srinivas, A.; Parmar, N.; Hechtman, B.; Shlens, J.R. Scaling local self-attention for parameter efficient visual backbones. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12894–12904.
39. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. Deepglobe 2018: A challenge to parse the earth through satellite images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 172–181.
40. Friedland, M.L. *The University of Toronto: A History*; University of Toronto Press: Toronto, ON, Canada, 2013.

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