applied sciences

MDPI

Article

DCS-TransUperNet: Road Segmentation Network Based on CSwin Transformer with Dual Resolution

Zheng Zhang, Chunle Miao, Chang'an Liu and Qing Tian *

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.)
* Correspondence: tianqing@ncut.edu.cn

Abstract: Recent advances in deep learning have shown remarkable performance in road segmentation from remotely sensed images. However, these methods based on convolutional neural networks (CNNs) cannot obtain long-range dependency and global contextual information because of the intrinsic inductive biases. Motivated by the success of Transformer in computer vision (CV), excellent models based on Transformer are emerging endlessly. However, patches with a fixed scale limit the further improvement of the model performance. To address this problem, a dual-resolution road segmentation network (DCS-TransUperNet) with a features fusion module (FFM) was proposed for road segmentation. Firstly, the encoder of DCS-TransUperNet was designed based on CSwin Transformer, which uses dual subnetwork encoders of different scales to obtain the coarse and fine-grained feature representations. Secondly, a new FFM was constructed to build enhanced feature representation with global dependencies, using different scale features from the subnetwork encoders. Thirdly, a mixed loss function was designed to avoid the local optimum caused by the imbalance between road and background pixels. Experiments using the Massachusetts dataset and DeepGlobe dataset showed that the proposed DCS-TransUperNet could effectively solve the discontinuity problem and preserve the integrity of the road segmentation results, achieving a higher IoU (65.36% on Massachusetts dataset and 56.74% on DeepGlobe) of road segmentation compared to other state-of-the-art methods. The considerable performance also proves the powerful generation ability of our method.

Keywords: remote sensing image; road segmentation; CSwin Transformer; dual scales; long-range contextual dependencies

Citation: 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. https://doi.org/10.3390/app12073511

Academic Editors: Qizhi Xu, Jin Zheng and Feng Gao

Received: 26 December 2021
Accepted: 21 March 2022
Published: 30 March 2022

1. Introduction

High precision and real time are crucial for obtaining road network information in specific application scenarios, for instance, automatic navigation, urban transportation management and military combat planning [1–3]. It is costly and laborious for obtaining the road information by relying on visual interpretation [4]. Therefore, automatic road segmentation [5–7], from remote sensing images, has become a promising method. Nevertheless, this has always been a difficulty because the roads in the different regions have irregular physical properties, such as size, geometrical shape and geographic distribution. Moreover, there is the complexity of the background beside the road in high-resolution remote sensing imagery. For example, trees and vehicles will block the road. The above factors will increase the difficulty of road segmentation. Accordingly, extracting high-quality road information is still a challenge.

In recent years, a large number of novel segmentation methods for road extraction have been proposed by research workers. These methods are normally classified as traditional methods and semantic segmentation methods. In traditional methods, the researchers put forward a reasonable hypothesis that the grayscale value belonging to the same category is relatively consistent, but the grey values between classes will form a certain contrast. This assumption ensures that road pixels are distinguishable and divisible. Researchers

divide the road extraction methods into two categories: pixel-based and object-oriented. In method-based pixels, the differences between spectral features are fully utilized to identify roads. For example, Mu et al. [8] introduced a method-based Otsu threshold to distinguish the road, the output is a binary image composed of road and non-road road pixels. Coulibaly et al. [9] proposed a method based on the spectral angle for road extraction, which distinguished the road pixels by the spectral angle threshold.

In object-oriented road segmentation, the remote sensing imagery was clustered and split into different areas, regarded as the target of road segmentation. A method based on region-growing was designed by Lu et al. [10] to extract roads from synthetic aperture radar imageries. Furthermore, Yu and Yi [11] integrated Markov random field based on object-oriented methods. The methods mentioned above have attained excellent performance under certain circumstances. However, these manual methods designed for specific scenarios have poor generalization ability, and many thresholds' parameters demand a tune up.

Unlike the traditional method, the method based on deep learning depends on the superiority of feature representation learning and the ability for parameter sharing to distinguish roads automatically and efficiently. For instance, Gao et al. [12] introduced a CNN-based network with the multifeatured pyramid to obtain more road information details because this structure validly expands the receptive field of the features. A network-in-network structure combined with a full convolution neural network (FCN) successfully was proposed by Mendes et al. [13], which utilized the superiority of large context windows to achieve faster detection. In addition, Li et al. [14] introduced a hybrid CNN combining multiple subnetworks to detect multi-scale roads. These methods based on CNNs demonstrated superior performance in terms of automatic road segmentation. However, due to the intrinsic inductive biases, the specific convolution layer fails to obtain a global receptive field. So, their performance will face great resistance to further breakthroughs. It is a sub-optimal choice for CNNs to expand the receptive field by stacking convolution layers and down-sampling. The reason is that these operators make the model more complex and easier to over-fit. However, many scholars have attempted to build long-distance dependencies in CNNs, such as attention mechanisms. There is still great potential in obtaining the global receptive field.

Lately, the new structure Transformer [15], proposed for sequence-to-sequence tasks in natural language processing (NLP), has aroused heated discussion in the CV community. Transformer performs well on many NLP tasks. The reason is that the multi-head self-attention mechanism can obtain a global connection between the tokens. Surprisingly, the researchers found that this ability to build long-distance dependencies also applies to pixel-based cv tasks. For example, DETR [16] is the pioneering work of object detection based on Transformer. ViT [17] is the first image classification model based on native Transformer. In order to optimize computational complexity, Swin Transformer and CSwin Transformer were proposed successively and achieved comparable performance. SETR [18] proved that Transformer as an encoder was able to obtain state-of-the-art (SOTA) performance in semantic segmentation. However, the potential of models based on Transformer has not been fully tapped in the remote sensing field.

In addition, multiscale feature representations are also a crucial role in vision Transformer. The latest research results are CrossViT [19] applied in image classification, MViT [20], designed for video recognition, and M2TR [21] designed for object detection. Generally, multi-scale feature representation can improve performance in vision Transformer, but has rarely been applied in remote sensing.

In order to alleviate the intrinsic inductive bias of CNN, we introduced a new framework based on the Transformer encoder, which mainly took into account the superiority of CSwin Transformer and multi-scale vision Transformer to efficiently combine the architecture of UperNet for automatic remote sensing image segmentation. The proposed encoder of DCS-TransUperNet uses dual subnetworks to learn feature representation from different scales inputs. In particular, we first obtained many overlapping patches by cutting

the input imagery, the scales of which are large scales and small scales, respectively. In order to fully utilize these features, we proposed a specific FFM to fuse the multi-scale features from the two encoder subnetworks. Furthermore, we adopted the UperNet as the decoder. Finally, we obtained the pixel-level predictions whose resolutions are the same as the input images by progressive up-sampling. In addition, to alleviate the problems caused by sample imbalance, we proposed a novel loss function L_{MIX}, which maintained the stability of the gradient and prevented falling into the local optimum. Profiting from these refinements, the proposed DCS-TransUperNet can improve the performance of road segmentation. We evaluated the effectiveness of DCS-TransUperNet on two remote sensing road datasets. Sufficient experiments show that our proposed DCS-TransUperNet for road extraction achieves superior performance in the Massachusetts and the DeepGlobe datasets.

2. Relate Work

This section first reviewed the most classic-methods-based CNN in road extraction from network structure and loss function. Then, we summarized the related work on vision Transformer, particularly in segmentation. Finally, we gave an overview of the typical methods used for multi-scale feature extraction. Furthermore, we made a comparison between these methods and the one we proposed.

2.1. Road Extraction Based on CNNs

Super performance [22–24] about CNN has been shown in remote sensing image segmentation, especially U-Net with an encoder–decoder structure, e.g., in order to enhance the effectiveness of road feature representation, a variant of U-Net with residual was presented by Zhang et al. [25]. The model absorbed the strong points of U-Net and residual networks so that the convergence property and information propagation were better. Likewise, Xin et al. [26] designed a new model that mainly consists of skip contacts and densely connected blocks. The model was able to utilize road features representation from different layers fully. Moreover, a model based on LinkNet was introduced by Zhou et al. [27] in which an elaborate dilation convolution structure was applied to capture multi-scale context information. The methods mentioned above perform well, but they are still CNN based.

In addition, many researchers also have obtained accurate segmentation performance by designing an effective loss function besides modifying the structure of the model. Wei et al. [28] introduced a novel loss based on the cross-entropy loss function. The loss skillfully combined the geometrical features of road characteristics. In order to avoid blurriness segmentation results, He et al. [29] defined a loss function based on structural similarity. Mosinska et al. [30] believe that the topological structure also affects the final result, so they proposed a loss function, synthesizing the topological awareness. These loss functions fully consider the structural characteristics of roads. However, they neglect the influence of the imbalance between background information and road samples.

2.2. Vision Transformer

Motivated by the excellent performance of Transformer in miscellaneous NLP tasks, increasing the number of methods based on Transformer was proposed in various computer vision tasks. In the recent research on vision Transformer, a novel model based on pure Transformer was introduced by ViT [17]. The model achieved the SOTA performance on the image classification task. However, such performance must depend on pre-training on large datasets, such as JFT-300M and ImageNet-22K. In order to solve the above problem, knowledge distillation was applied in DeiT [28]. The model with an effective training strategy allows ViT to promote performance on smaller datasets, such as ImageNet-1K. Swin Transformer [31], a hierarchical architecture model, has linear computational complexity relative to the image size. This model, through limited self-attention to non-overlapping local windows, proposed a shifted window scheme to achieve higher efficiency. Furthermore, CSwin Transformer [32] effectively strengthened the information

interaction between patches. This model obtained SOTA performance in various CV tasks and achieved lower computational complexity than Swin Transformer. Motivated by these models, we introduced an UperNet-like structure that used CSwin Transformer blocks to feature extraction.

2.3. Multi-Scale Transformer

In computer vision, multi-scale feature representation based on CNN has been widely applied. In particular, the representative FPN (feature pyramid network) [33] promoted the progress of the classic cv tasks [34–37], such as object detection and semantic segmentation. Nevertheless, we have not fully explored the potential of this typical structure in vision Transformer. Recent relevant work about multi-scale Transformer is as follows. A new dual-branch visual Transformer was proposed by CrossViT [19] to extract multi-scale feature representation. This model made an efficient token fusion strategy based on cross attention. Accordingly, the performance is better than ViT, and DeiT on image classification. M2TR [21] designed a multi-scale feature representation based on different patch sizes. Moreover, by connecting the multi-scale feature with the Transformer model, Haoqi Fan et al. proposed a multi-scale visual Transformer (MViT) [20] for video and image recognition. Inspired by the enormous potentiality of multi-scale vision Transformer, we designed the dual branch encoder, which profits from the powerful feature representation capability of CSwin Transformer. In addition, we proposed a special module called FFM to achieve the effective interaction between different scale features.

3. Method

This section presents a novel segmentation network, DCS-TransUperNet. Firstly, we introduce the standard Transformer [38] and CSwin Transformer [32] used in DCS-TransUperNet. Secondly, we propose the encoder and decoder based on CSwin Transformer. Then, we describe the benefits of the dual design and explain how to realize the interaction effects between different scale features by FFM. Finally, we propose the new loss function to assist network training.

3.1. CSwin Transformer Block

The standard Transformer encoder consists of a stack of N identical blocks. As shown in Figure 1a, the core compositions of the Transformer block are multi-head-attention (MSA) and multilayer perceptron (MLP). In addition, in order to ensure the stability of the data feature distribution, the layer norm is introduced. Meanwhile, the residual structure is used to ensure the effectiveness of deeper network training. So, the output z_l of the L layer in the Transformer is defined as

$$\begin{aligned} \widetilde{z}_l &= \mathrm{MSA}(LN(z_{l-1})) + z_{l-1} \\ z_l &= MLP(LN(\widetilde{z}_l)) + \widetilde{z}_l \end{aligned} \tag{1}$$

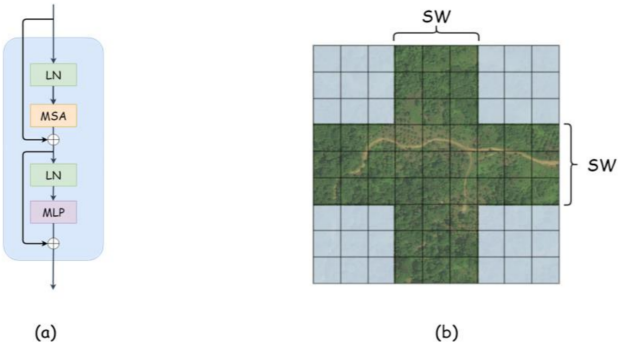

Figure 1. (**a**) The structure of a standard Transformer block. (**b**) The cross-shaped window.

One of the challenges of the standard Transformer is that the calculation of the global attention mechanism is computationally expensive. In particular, for high-resolution pictures and intensive tasks, the cost of such calculations is unacceptable. For more efficient modeling, the cross-shaped window self-attention mechanism is proposed by CSwin Transformer. This module reduces the computational complexity and effectively strengthens the information interaction between patches. Specifically, as shown in Figure 1b, we divide the input features into equal width stripes, which will form a cross-shaped window. Cross-shaped window MSA will calculate self-attention weight in both vertical and horizontal directions. The formulation of cross-shaped window MSA is given as [29]

$$\begin{aligned} CSWin - Attention(X) &= Concat(head_1, \ldots, head_k)W^O \\ where\ head_k &= \begin{cases} H - Attention_k(X)\ k = 1, \ldots, K/2 \\ V - Attention_k(X)\ k = K/2 + 1, \ldots, K \end{cases} \end{aligned} \tag{2}$$

3.2. Encoder

We adopt a similar UPerNet structure in our model on the whole. For the encoder section, we apply CSwin Transformer as the backbone for feature representation. As shown in Figure 2, many overlapping patches are obtained by cutting the input image. Here, the length of patches is $(W/s) \times (H/s)$, where s means the patch size, and each encoded patch is called a "token". Through the linear embedding layer, the patch is projected onto dimension C. Specifically, the convolution operation generates the overlapping patches. The convolution kernel is 7, the stride is 4, and the number of output channels is C. After adding position information, the tokens are fed to CSwin Transformer for feature extraction, which goes through four stages. Among the stages, it is composed of some CSwin Transformer blocks. In order to increase the receptive field of the attention mechanism, convolution is introduced to reduce the resolution of the features with the Transformer blocks getting deeper. Except for the last stage, the number of channels of the feature will also increase exponentially. Specifically, the feature resolution is reduced to half, and the number of channels is doubled after going through the convolution layer. So, the output resolutions of the four stages are $(W/s) \times (H/s)$, $(W/2s) \times (H/2s)$, $(W/4s) \times (H/4s)$, $(W/8s) \times (H/8s)$ and the numbers of channels are C, 2C, 4C, and 8C.

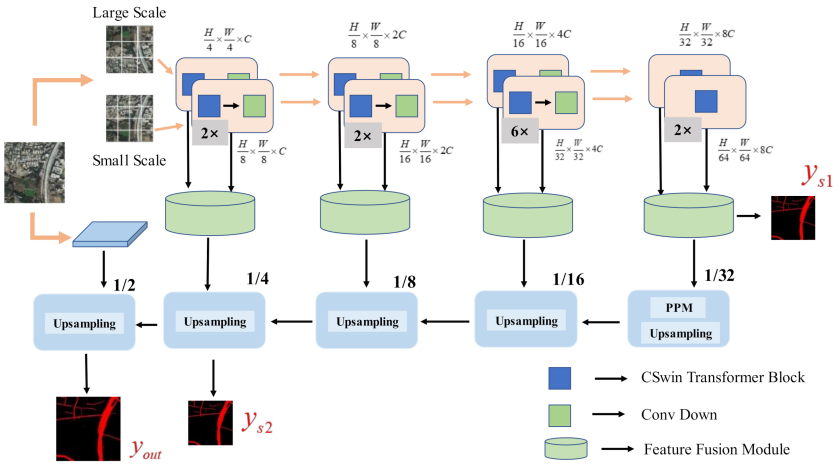

Figure 2. Illustration of the designed DCS-TransUperNet. We obtain overlapping patches at different scales by splitting the remote sensing image and then feed these patches into dual branches of the CSwin Transformer decoder. Next, to fully utilize these features, we propose a specific FFM to fuse the multi-scale features from the two encoder subnetworks. Finally, the fused feature representations are restored to the same resolution as the input image by up-sampling based on the feature pyramid. Therefore, we obtain the final mask predictions.

3.3. Multi-Scale Feature Representations and Fusion

Although MSA effectively establishes long-range dependencies between patches, the pixel-level features in the patch are disregarded. This detailed information is particularly important for the prediction of dense small roads. In addition, ViT achieves better performance through the fine-grained patch size. In order to enhance the performance and robustness of road extraction, a multi-scale CSwin Transformer is used for feature extraction.

It is complementary to the information from different patches in feature extraction. Coarse-grained features are easier to be captured by large patches, while the small scale has a faculty for obtaining fine-grained characteristics. Although multi-layer perceptron introduces the local information between batches, the pixel-level information inside the patch is still missing. In DCS-TransUperNet, these problems are alleviated to some extent. Inspired by this, a multi-scale dual CSwin Transformer encoder is proposed. Specifically, two branch structures are adopted to obtain the features at different spatial levels. One is the main branch (patch size is 4), and the other is the auxiliary (patch size is 8). As a result, the resolutions of the output features obtained from the large-scale branch are $(H/4) \times (W/4)$, $(H/8) \times (W/8)$, $(H/16) \times (W/16)$ and $(H/32) \times (W/32)$ while those of the other branch are $(H/8) \times (W/8)$, $(H/16) \times (W/16)$, $(H/32) \times (W/32)$ and $(H/64) \times (W/64)$.

Feature representations containing abstract information are obtained from dual branch encoders. The following work is how to realize the information interaction between them effectively. A simple way is to straightforwardly concatenate the features from different branches followed by convolution layers. However, this direct way cannot obtain long-distance dependence and global connections from different scale features. Hence, a new FFM is introduced, which relies on the advantages of MSA to make the interaction more effective between multi-scale features. In particular, the standard Transformer is chosen rather than CSwin Transformer in FFM. The main reason is that the standard Transformer works on the feature map based on a rectangle block. A token is generated to be size specific based on one branch's features map. The next step is to calculate the attention weight of the tokens sequence reshaped by the other branch. In addition, the number of the Transformer block in FFM is only two, which will not cause computational complexity. A more intuitive explanation is shown in Figure 3.

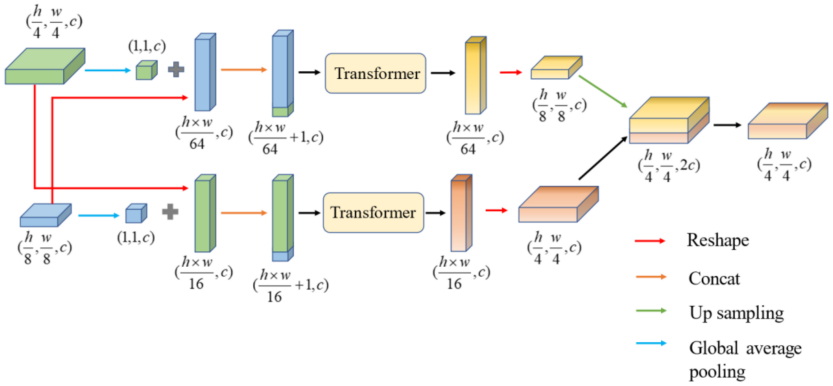

Figure 3. Illustration of FFM plays an important role as a core component in features interaction from different scales.

FFM can aggregate information from different scale features. Let us take the output features from the first CSwin Transformer stage as an example and explain them in detail, and the same measure is suitable for the large-scale branch.

In particular, the two branches' output features from the same stage j (j: 1, 2, 3, 4) are represented by

$$H^j = [h_1^j, h_2^j, \ldots, h_{h \times w/16}^j] \in \mathbb{R}^{C \times (\frac{h}{4} \times \frac{w}{4})} \quad (3)$$

$$F^j = [f_1^j, f_2^j, \ldots, f_{h \times w/64}^j] \in \mathbb{R}^{C \times (\frac{h}{8} \times \frac{w}{8})} \quad (4)$$

Then the transformation output of H^j is obtained by the global average pooling layer.

$$\widetilde{h}^j = Flatten(Gavgpool(H^j)) \tag{5}$$

where $\widetilde{h}^j \in \mathbb{R}^{C\times 1}$, Gavgpool indicates global average pooling, and then the flatten operation is used. For further description, the global abstract information is expressed by token $\widetilde{h}^j$, making the information fusion with F^j at the pixel level. In the meantime, $\widetilde{h}^j$ is concatenated with F^j to form a sequence of $1 + C \times ((H/8) \times (W/8))$ tokens, followed by a standard Transformer to calculate the global self-attention:

$$\begin{aligned} \widetilde{F}_j &= \mathrm{MSA}([\widetilde{h}^j, f_1^j, f_2^j, \ldots, f_{h\times w/64}^j]), \\ &= [\hat{f}_0^j, \hat{f}_1^j, \ldots, \hat{f}_{h\times w/64}] \in \mathbb{R}^{C\times(1+h\times w/64)} \end{aligned} \tag{6}$$

$$F_{out}^j = [\widetilde{f}_1^j, \widetilde{f}_2^j, \ldots, \widetilde{f}_{h\times w/64}^j] \in \mathbb{R}^{C\times(h\times w/64)} \tag{7}$$

where MSA is the key component of the standard Transformer, and $\widetilde{F}_j$ is fed into the MLP layer for linear transformation. Finally, F_{out}^j is the ultimate feature of the smaller-scale branch in FFM. This method skillfully integrates information between each token in $F^j = [f_1^j, f_2^j, \ldots, f_{h\times w/64}^j] \in \mathbb{R}^{C\times(\frac{h}{8}\times\frac{w}{8})}$ and the whole H^j, so that coarse-grained features from the larger-scale branch are obtained by the fine-grained ones. Consequently, FFM can accomplish better segmentation performance due to this effective multi-scale feature fusion mechanism.

3.4. Decoder

The UperNet-like architecture is adopted, mainly consisting of two partials: pyramid pooling module and multi-scale feature pyramids. In detail, the output feature of stage 4 in FFM is adopted as the first input of the decoder. The bilinear interpolation is introduced as being up-sampled by 2 in each step of the decoder. Meanwhile, these features are concatenated with the corresponding skip connection from FFM in the same step. The reasons for this are as follows: (1) The pyramid pooling module extracts features from different scales and then aggregates them. This approach improves the robustness of the algorithm. (2) Multi-scale feature pyramids enhance the interaction between features with different granularity for better segmentation performance.

After the two partials above, the quarter size features of the original image are obtained. In order to prevent the loss of shallow features, the input image is down-sampled to obtain a low-level feature with half the resolution of the original image, followed by up-sampling and fusion of the above features step by step instead of using a four up-sampling operator directly. Finally, the output mask predictions are obtained by all these operators.

3.5. Mixed Loss Function

Road extraction is often considered a pixel-level classification problem. Each pixel in the image is classified as road or non-road. Generally, we solve this problem by a binary cross-entropy loss function, whose formula is expressed as follows:

$$L_{BEC} = -\frac{1}{N}\sum_{n=1}^{N}(y_n \log(y'_n) + (1-y_n)\log(1-y'_n)) \tag{8}$$

However, most of the pixels are non-road in high-resolution remote sensing images, and the pixels of roads only account for a small proportion. This means that the two types of pixel distributions are very uneven. In this case, if the loss function treats road pixels and non-road pixels equally, the model constrained by the loss function is optimized toward non-road categories to fall into the local optima. Furthermore, the road's pixel accuracy is declined under the constraint of the loss function. Therefore, in order to solve the problem

of classes imbalance, we propose a new loss function L_{MIX} based on binary cross-entropy loss and the Dice loss. The L_{DICE} L_{MIX} formulaic expressions are as follows:

$$L_{DICE} = 1 - \frac{1}{N}\left(\frac{2\sum_{n=1}^{N} y_n y'_n + o}{\sum_{n=1}^{N} y_n + \sum_{n=1}^{N} y'_n + o} + \frac{2\sum_{n=1}^{N} (1-y_n)(1-y'_n) + o}{\sum_{n=1}^{N} (1-y_n) + \sum_{n=1}^{N} (1-y'_n) + o}\right) \quad (9)$$

where L_{DICE} represents Dice loss and y_n is the ground truth label, whose value is either 0 or 1. y'_n is the probability of the road.

$$L_{MIX} = L_{BEC} + \lambda L_{DICE} \quad (10)$$

where L_{DICE} is the Dice loss, L_{BCE} is the binary cross-entropy, and λ is a coefficient to balance between L_{BCE} and L_{DICE}.

From Formula (10), it can be concluded that the similarity between the predicted value and the real label value only determines Dice, which is independent of the number of training samples. Therefore, it can alleviate the problems caused by sample imbalance. However, Dice loss has some flaws, and strong gradient changes, which will lead to unstable training. Considering the above, we propose the new Loss function L_{MIX}, which not only maintains the stability of the gradient but also prevents falling into the local optimum.

4. Experiment

This section first introduces two widely used data sets: the Massachusetts roads dataset and the DeepGlobe dataset. Then, we briefly describe some evaluation metrics, such as precision, recall, F-1 score and IoU. Next, the implementation details of the experiment are introduced in detail. Finally, we demonstrate the optimal experimental results and perform ablation experiments.

4.1. Datasets

In order to estimate the performance of the designed DCS-TransUperNet, we perform enough comparative experiments on the Massachusetts road dataset and the DeepGlobe dataset.

The Massachusetts roads dataset [36] widely is applied for model verification of various road extraction. The dataset consists of 1171 images with a resolution of 1 m. Each one has a corresponding binary label whose resolution is 1500 × 1500 pixels, like the image. We randomly split the dataset into three parts, including 1108 training images, 14 validation images and 49 test images. Before training, we perform data enhancements, such as random clipping and random rotation with different angles. Finally, we obtain 110,800 training samples at 256 × 256 pixels in size, 1400 validation samples and 49 original testing samples after data pre-processing.

The DeepGlobe dataset [37] comes from the CVPR DeepGlobe 2018 road extraction challenge. It contains 8570 images with the size of 1024 × 1024 pixels and a resolution of 0.5 m, which are split into 6626 images for training, 1243 images for validation and 1101 images for testing. Since the only binary labels of the training set are available after the challenge, we select 6626 training set images and corresponding labels as our experimental data. We randomly divide the dataset according to the ratio of 8:1:1 and then perform the same data enhancements as the Massachusetts road dataset.

4.2. Evaluation Metrics

We use four standard metrics [38] to estimate the segmentation performance of the model, including precision, recall, F1-score and IoU. The precision represents the proportion of correctly classified road pixels in the total prediction result, while the recall reflects the percentage of predicted road pixels in the ground truth. F1-score and IoU are comprehensive metrics that weigh precision and recall. The formulas of these metrics are shown as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (11)$$

$$\text{Re}call = \frac{TP}{TP + FN} \tag{12}$$

$$F1 - score = \frac{2TP}{2TP + FN + FP} \tag{13}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{14}$$

4.3. Implementation Details

We also applied multi-scale training strategies in our experiments apart from data augmentations. Motivated by [39], we found that in-depth supervision is helpful for model training. So, the final loss function L_{total} can be expressed as

$$L_{total} = \lambda L_{MIX}(y, y_{out}) + \mu L_{MIX}(y, y_{s1}) + \eta L_{MIX}(y, y_{s2}) \tag{15}$$

where y is the ground truth label, y_{out} is the final output, y_{s1} is the output of the final stage in the encoder, and y_{s2} is the output of the first stage in the decoder, which was additionally used to supervise deeper training. In addition, λ, μ, η are weight coefficients set to 0.7, 0.15, and 0.15, empirically. The SGD optimizer was applied to train our model, where the learning rate is equal to 0.01, momentum is 0.9, and the weight decay is 0.0001.

Our models are built by Pytorch and trained using 4 NVIDIA RTX 2080Ti GPUs. The trained epoch of our models was set at 100 and used early stopping and a cosine annealing schedule. We provided two versions of the model: the DCS-TransUperNet-B adopts CSwin-Base as the primary branch encoder, while DCS-TransUperNet-L applies CSwin-Large. Both the above models use CSwin-Tiny as a complementary branch decoder. The pre-trained weights of the CSwin Transformer were obtained from [29]. Further, Table 1 shows the detailed parameters of the model.

Table 1. Hyperparameters of CSwin Transformer model variants.

Methods	Hidden Size	Layer Number	Head Number	Window Size	Param
CSwin-T	64	[1,2,21,1]	[2,4,8,16]	[1,2,7,7]	23M
CSwin-B	96	[2,4,32,2]	[2,4,8,16]	[1,2,7,7]	78M
CSwin-L	144	[2,4,32,2]	[6,12,24,48]	[1,2,7,7]	173M

4.4. Result on Massachusetts Roads Dataset

In order to evaluate the performance of our proposed model, we conduct adequate comparative experiments. We select many SOTA segmentation networks as a reference, including D-LinkNet [24], ELU-SegNet [40], GL-Dense-U-Net [41], ResUnet [22] and D-ResUnet [42].

The visual comparison results are shown in Figure 4. On the whole, DCS-TransUperNet is significantly improved, such as the highlighted place in the green boxes. Specifically, our model has a more powerful ability to extract small roads in the first image. The second image we select is a scene with a large amount of vegetation occlusion. DCS-TransUperNet and D-ResUnet overcome the problem of vegetation cover to a certain extent, but DCS-TransUperNet has better continuity, as shown in the blue box. The last image represents a dense scene, which shows that our recall has increased significantly.

Table 2 shows the quantitative comparison results of different methods. DCS-TransUperNet with mixed loss achieved superior performance in precision (82.44), recall (78.43), F1-score (80.39) and IoU (65.36). The experimental results show that the proposed DCS-TransUperNet has superior performance in road extraction. However, the disadvantage is that the inference speed is slower compared with the method based on CNN.

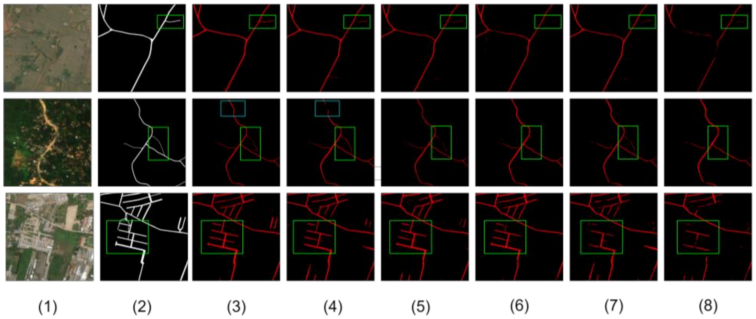

Figure 4. Visual analysis of road segmentation results using different methods. (**1**) The test image. (**2**) The ground truth. (**3**) Results with DCS-TransUperNet. (**4**) Results with D-ResUnet. (**5**) Results with ResUnet. (**6**) Results with GL-Dense-U-Net, (**7**) ELU-SegNet. (**8**) Results with D-LinkNet.

Table 2. Quantitative analysis results using different methods on Massachusetts roads dataset.

Method	Precision	Recall	F1-Score	IoU	Inference
D-LinkNet	75.57	72.63	74.07	58.80	1.75
ELU-SegNet	77.98	69.50	73.50	58.04	**1.02**
GL-Dense-U-Net	75.37	77.32	76.33	61.59	2.52
ResUnet	83.64	71.05	76.83	62.69	1.35
D-ResUnet	**85.28**	71.34	77.69	63.44	1.46
DCS-TransUperNet	82.44	**78.43**	**80.39**	**65.36**	2.94

In order to prove the effectiveness of Dual-scales, we design a comparative experiment between CSwin TransUperNet and DCS-TransUperNet. Note that the former does not add Dual-scales, while the latter adds it. The experimental results are shown in Table 3. The IoU of road extraction is increased from 64.53 to 65.36 by adding the Dual-scales module. The visual comparison results are shown in Figure 5. The robustness and recall of the road segmentation are superior in DCS-TransUperNet. Therefore, the effectiveness of the Dual-scales and FFM is further verified. In addition, our model also makes comparative experiments on the selection of encoder backbone and decoder.

From Table 3, we find that the performance in road segmentation of CSwin Transformer is better than Swin Transformer, and the UperNet decoder is better than U-Net. In Tables 2 and 3, we find that the Transformer-based model is generally better than the CNNs-based model in IoU, while the inference is slower. As shown in Table 4, we design different variants of DCS-TransUperNet, and DCS-TransUperNet-L has better performance in the F1-score and IoU but also has higher parameters and slower inference speed.

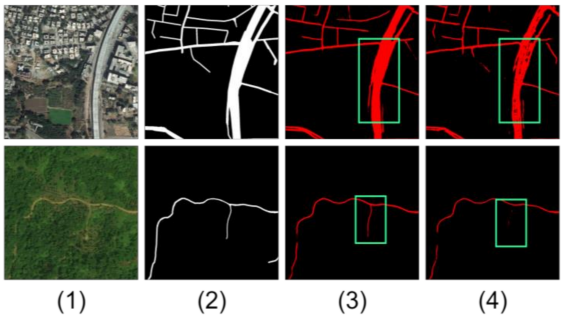

Figure 5. Visual analysis of road segmentation results using Dual-scales and not using them. (**1**) The test image. (**2**) The ground truth. (**3**) Results with DCS-TransUperNet. (**4**) Results with CSwin TransUperNet.

Table 3. Quantitative analysis results with different encoder, decoder and Dual-scales.

Method	Precision	Recall	F1-Score	IoU	Inference
CSwin TransU-Net	81.01	77.04	78.97	64.05	1.98
CSwin TransUperNet	81.44	77.62	79.48	64.53	2.51
Swin TransUperNet	80.79	76.91	78.80	63.78	3.27
DCS-TransUperNet	**82.44**	**78.43**	**80.39**	**65.36**	**2.94**

Table 4. Quantitative analysis results with different DCS-TransUperNet variants.

Method	Precision	Recall	F1-Score	IoU	Inference
DCS-TransUperNet-B	81.63	77.42	79.46	64.22	**2.01**
DCS-TransUperNet-L	**82.44**	**78.43**	**80.39**	**65.36**	2.94

In order to evaluate the effects of the hyperparameter λ in the mixed loss function We set the λ to [0, 0.001, 0.01, 0.1, 1, 10], respectively, to perform the experiments as shown in Figure 6; when λ = 0, the score of F1-score and IoU are the lowest, 78.01% and 64.07%, respectively. In this case, the mixed loss function has degenerated to the binary cross-entropy loss. With the increase in λ, the score is gradually improving. When λ = 1, the score of the F1-score and IoU reaches its maximum (80.39% and 65.36%) and then decreases rapidly. Therefore, the λ is set to 1 in our experiments.

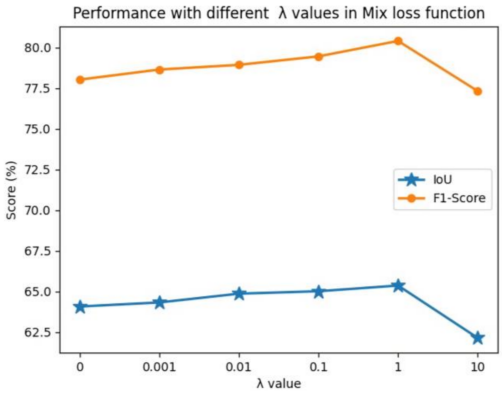

Figure 6. Performance of DCS-TransUperNet with different λ values in the mixed loss function.

4.5. Result on DeepGlobe Dataset

In order to further evaluate the DCS-TransUperNet, we conducted a comparative experiment between our models and other SOTA road extraction methods on the DeepGlobe dataset, such as FCN-8s, U-Net, D-LinkNet, ELU-SegNet, and D-ResUnet. The experimental environment remains the same as previous experiments.

Figure 7 illustrates a representative comparison result of road extraction on the DeepGlobe dataset from a globe view. There is more intensive road segmentation in the DeepGlobe dataset, which also increases the difficulty of road detection. Intuitively, our proposed result of road segmentation has better continuity and fewer false positives than others, such as the sub-regions marked by red boxes shown in Figure 8.

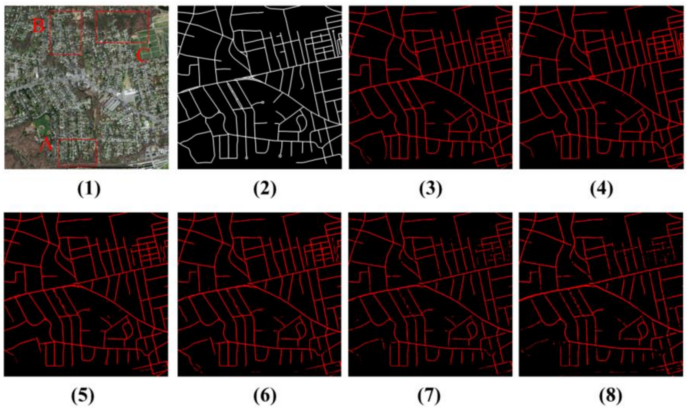

Figure 7. Visual analysis of road segmentation results using different networks. (**1**) Testing remote sensing image. (**2**) The ground truth of this image. (**3**) Segmentation results with DCS-TransUperNet. (**4**) Segmentation results with D-ResUnet. (**5**) Segmentation results with ELU-SegNet. (**6**) Segmentation results with D-LinkNet. (**7**) Segmentation results with U-net. (**8**) Segmentation results with FCN-8s.

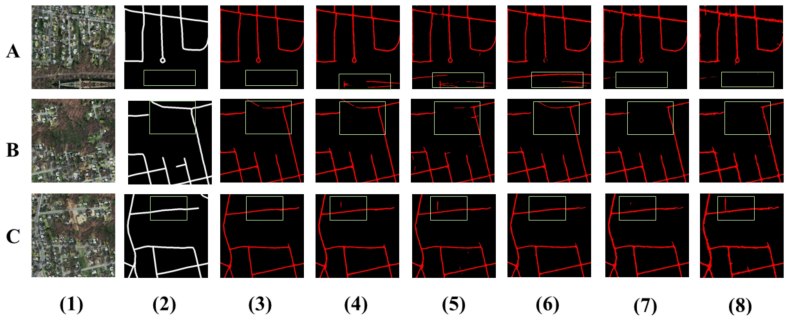

Figure 8. Road segmentation results of subareas. (**A**–**C**) represent road segmentation results of subareas with different networks, respectively. (**1**) Input remote sensing image of the subareas. (**2**) The ground truth of the subareas. (**3**) Segmentation results with DCS-TransUperNet. (**4**) Segmentation results with D-ResUnet. (**5**) Segmentation results with ELU-SegNet. (**6**) Segmentation results with D-LinkNet. (**7**) Segmentation results with U-net. (**8**) Segmentation results with FCN-8s.

Figure 8 shows more details. These three images are from the red box areas of Figure 7. The second column is the corresponding ground truth of this sub-region. The other columns are road segmentation results. From Figure 8A, there is obvious road false detection in (4), (5) and (6). Compared with (7) and (8), the segmentation result of (3) has better continuity and integrity. From Figure 8B, the road extraction results of (3), (4), and (6) are better than those of (5), (7), and (8). From Figure 8C, the overall segmentation result of (3) has made great progress, including continuity and accuracy compared with other methods. To sum up, our proposed model has strong road extraction ability.

Table 4 shows the quantitative comparison results of different methods on the DeepGlobe dataset. DCS-TransUperNet with mixed loss achieved superior performance in precision (77.94), recall (69.52), F1-score (73.48) and IoU (56.74) in Table 5.

Table 5. Quantitative analysis results using different methods on DeepGlobe dataset.

Method	Precision	Recall	F1-Score	IoU	Inference
FCN-8s	65.08	63.12	63.56	47.23	1.66
U-Net	72.33	65.84	68.93	52.86	1.83
D-LinkNet	70.79	68.95	69.85	53.62	1.75
ELU-SegNet	78.02	55.63	64.94	48.01	**1.02**
D-ResUnet	**79.38**	63.94	70.82	54.76	1.46
DCS-TransUperNet	77.94	**69.52**	**73.48**	**56.74**	2.94

5. Discussion

Our model has two main advantages in innovation. Firstly, the proposed encoder of DCS-TransUperNet uses dual subnetworks to learn features representation from different scale inputs, which avoids the limitation of the fixed scale patch. Coarse-grained features are more easily captured by large patches, while small scales have the faculty of obtaining fine-grained features. Secondly, to obtain enhanced features, we propose a specific FFM to fuse the coarse- and fine-grained feature representations from the two encoder subnetworks. FFM differs from other fusion modules in that we integrate a two-layer native Transformer to increase the interaction of information instead of just using concat and convolution operations.

Through visual analysis, the road segmentation results of our proposed model have better continuity and integrity. Furthermore, as shown in the green box in the second image in Figure 4, our model can also detect the roads blocked by trees because of Transformer's powerful global context modeling ability.

We give four evaluation metrics, which are precision, recall, F1-score and IoU. Among them, precision and recall are a pair of reverse indicators. In other words, too high recall will inevitably lead to low precision. In order to better evaluate the performance of the model, the F1-score and IOU are the comprehensive embodiment of precision and recall. Through experimental analysis, the DCS-TransUperNet achieves state-of-the-art results on the Massachusetts roads dataset (F1-score 80.39%, IoU 65.36%) and DeepGlobe dataset (F1-score 73.48%, IoU 56.74%).

In the experiment, we found that D-resunet achieves higher precision than our model. The reason is that D-resunet has a weak recall. From Table 2, compared with D-resunet, the recall of our model is improved by 7.09%, and the precision is only reduced by 2.84%. On the whole, our F1-score and IOU are still the best.

In terms of inference speed, the Transformer-based model has obvious disadvantages compared with the CNNs model. The reason is that the calculation of the attention mechanism consumes a lot of computational power. In the future, we will try to reduce the time complexity of the attention mechanism and improve the inference speed of our model.

6. Conclusions

In this paper, we introduce a new framework based on the CSwin Transformer, which mainly takes into account the superiority of multi-scale vision Transformer to efficiently combine the architecture of UperNet for automatic remote sensing image segmentation.

The main contributions are as follows: firstly, the proposed encoder of DCS-TransUperNet uses dual subnetworks to learn feature representation from different scale inputs, which avoids the limitation of the fixed scale patch. Coarse-grained features are more easily captured by large patches, while small scales have the faculty of obtaining fine-grained features. Secondly, to obtain enhanced features, we propose a specific FFM to fuse the coarse and fine-grained feature representations from the two encoder subnetworks. Thirdly, to alleviate the problems caused by sample imbalance, we propose a new mixed loss function, which maintains the stability of the gradient and prevents falling into the local optimum. Profiting from these refinements, the proposed DCS-TransUperNet can improve the performance of road segmentation.

Experiments using the Massachusetts dataset and DeepGlobe dataset showed that the proposed DCS-TransUperNet could effectively solve the discontinuity problem and preserve the integrity of the road segmentation results, achieving a higher IoU (65.36% on Massachusetts dataset and 56.74% on DeepGlobe) of road segmentation, compared to other state-of-the-art methods. The considerable performance also proves the powerful generation ability of our method.

In the next work, we will pay attention to building lighter models based on Transformer to enhance the speed of road segmentation. As we all know, Transformer's computing time is mainly spent on the attention mechanism. The memory occupied by attention

increases with the square of the calculation. So, we plan to reduce the time complexity by designing new attention computing methods.

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. and C.M.; original draft preparation, C.M.; visualization, Z.Z.; supervision, C.L. 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. 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]
2. Hinz, S.; Baumgartner, A.; Ebner, H. Modeling Contextual Knowledge for Controlling Road Extraction in Urban Areas. In *IEEE/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas (Cat. No. 01EX482)*; IEEE: Piscataway, NJ, USA, 2001; pp. 40–44.
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. Xu, Y.; Xie, Z.; Wu, L.; Chen, Z. Multilane Roads Extracted from the OpenStreetMap Urban Road Network Using Random Forests. *Trans. GIS* **2019**, *23*, 224–240. [CrossRef]
5. Bastani, F.; He, S.; Abbar, S.; Alizadeh, M.; Balakrishnan, H.; Chawla, S.; Madden, S.; DeWitt, D. Roadtracer: Automatic Extraction of Road Networks from Aerial Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 19–23 June 2018; pp. 4720–4728.
6. Wu, Q.; Luo, F.; Wu, P.; Wang, B.; Yang, H.; Wu, Y. Automatic Road Extraction from High-Resolution Remote Sensing Images Using a Method Based on Densely Connected Spatial Feature-Enhanced Pyramid. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 3–17. [CrossRef]
7. Lu, X.; Zhong, Y.; Zheng, Z.; Liu, Y.; Zhao, J.; Ma, A.; Yang, J. Multi-Scale and Multi-Task Deep Learning Framework for Automatic Road Extraction. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9362–9377. [CrossRef]
8. Mu, H.; Yun, Z.; Li, H.; Guo, Y.; Yuan, Z. Road Extraction Base on Zernike Algorithm on SAR Image. In Proceedings of the IGARSS 2016—2016 IEEE International Geoscience and Remote Sensing Symposium, Beijing, China, 10–15 July 2016.
9. Coulibaly, I.; Spiric, N.; Lepage, R.; St-Jacques, M. Semiautomatic Road Extraction from VHR Images Based on Multiscale and Spectral Angle in Case of Earthquake. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *11*, 238–248. [CrossRef]
10. Lu, P.; Du, K.; Yu, W.; Wang, R.; Deng, Y.; Balz, T. A New Region Growing-Based Method for Road Network Extraction and Its Application on Different Resolution SAR Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *7*, 4772–4783. [CrossRef]
11. Yu, C.; Yi, Y. Object-Based Road Extraction in Remote Sensing Image Using Markov Random Field. *Geomat. Inf. Sci. Wuhan Univ.* **2011**, *36*, 544–547.
12. 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]
13. Mendes, C.C.T.; Fremont, V.; Wolf, D.F. Exploiting fully convolutional neural networks for fast road detection. In Proceedings of the IEEE International Conference on Robotics & Automation, Stockholm, Sweden, 16–21 May 2016; pp. 3174–3179. [CrossRef]
14. Li, Y.; Guo, L.; Rao, J.; Xu, L.; Jin, S. Road Segmentation Based on Hybrid Convolutional Network for High-Resolution Visible Remote Sensing Image. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 613–617. [CrossRef]
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762. Available online: https://arxiv.org/abs/1706.03762v5 (accessed on 25 December 2021).
16. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
17. 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.
18. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.

19. Chen, C.-F.R.; Fan, Q.; Panda, R. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. *arXiv* **2021**, arXiv:2103.14899.
20. Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; Feichtenhofer, C. Multiscale Vision Transformers. *arXiv* **2021**, arXiv:2104.11227.
21. Wang, J.; Wu, Z.; Chen, J.; Jiang, Y.-G. M2TR: Multi-Modal Multi-Scale Transformers for Deepfake Detection. *arXiv* **2021**, arXiv:2104.09770.
22. Zhang, Z.; Liu, Q.; Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
23. Chen, Y.S.; Hong, Z.J.; He, Q.; Bin Ma, H. Road Extraction from High-Resolution Remote Sensing Images Based on Synthetical Characteristics. *Appl. Mech. Mater.* **2013**, *333–335*, 828–831. [CrossRef]
24. 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 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 182–186. [CrossRef]
25. Wei, Y.; Wang, Z.; Xu, M. Road Structure Refined CNN for Road Extraction in Aerial Image. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 709–713. [CrossRef]
26. He, H.; Yang, D.; Wang, S.; Wang, S.; Li, Y. Road Extraction by Using Atrous Spatial Pyramid Pooling Integrated Encoder-Decoder Network and Structural Similarity Loss. *Remote Sens.* **2019**, *11*, 1015. [CrossRef]
27. Mosinska, A.; Marquez-Neila, P.; Kozinski, 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, 18–22 June 2018; pp. 3136–3145. [CrossRef]
28. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *arXiv* **2021**, arXiv:2103.14030.
29. 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.
30. 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.
31. Nah, S.; Hyun Kim, T.; Mu Lee, K. Deep Multi-Scale Convolutional Neural Network for Dynamic Scene Deblurring. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3883–3891.
32. Cai, Z.; Fan, Q.; Feris, R.S.; Vasconcelos, N. A unified multi-scale deep convolutional neural network for fast object detection. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 354–370. [CrossRef]
33. Chen, C.-F.; Fan, Q.; Mallinar, N.; Sercu, T.; Feris, R. Big-Little Net: An Efficient Multi-Scale Feature Representation for Visual and Speech Recognition. *arXiv* **2018**, arXiv:1807.03848.
34. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. Higherhrnet: Scale-Aware Representation Learning for Bottom-up Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5386–5395.
35. Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.; Ku, A.; Tran, D. Image Transformer. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 4055–4064.
36. Mnih, V.; Hinton, G.E. Learning to Detect Roads in High-Resolution Aerial Images. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 210–223. [CrossRef]
37. 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, 18–23 June 2018; pp. 172–181.
38. Wei, Y.; Ji, S. Scribble-Based Weakly Supervised Deep Learning for Road Surface Extraction from Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–12. [CrossRef]
39. Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. Pranet: Parallel Reverse Attention Network for Polyp Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 263–273.
40. Panboonyuen, T.; Vateekul, P.; Jitkajornwanich, K.; Lawawirojwong, S. An Enhanced Deep Convolutional Encoder-Decoder Network for Road Segmentation on Aerial Imagery. In *International Conference on Computing and Information Technology*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 191–201. [CrossRef]
41. Xu, Y.; Xie, Z.; Feng, Y.; Chen, Z. Road Extraction from High-Resolution Remote Sensing Imagery Using Deep Learning. *Remote Sens.* **2018**, *10*, 1461. [CrossRef]
42. 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; pp. 3927–3930.