remote sensing

MDPI

Article

Complex Mountain Road Extraction in High-Resolution Remote Sensing Images via a Light Roadformer and a New Benchmark

Xinyu Zhang [1,2,†], Yu Jiang [3,4,†], Lizhe Wang [1,2], Wei Han [1,2,*], Ruyi Feng [1,2], Runyu Fan [1,2] and Sheng Wang [1,2]

1 School of Computer Science, China University of Geosciences, Wuhan 430078, China
2 Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences, Wuhan 430078, China
3 Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
4 University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: weihan@cug.edu.cn
† These authors contributed equally to this work.

Abstract: Mountain roads are of great significance to traffic navigation and military road planning. Extracting mountain roads based on high-resolution remote sensing images (HRSIs) is a hot spot in current road extraction research. However, massive terrain objects, blurred road edges, and sand coverage in complex environments make it challenging to extract mountain roads from HRSIs. Complex environments result in weak research results on targeted extraction models and a lack of corresponding datasets. To solve the above problems, first, we propose a new dataset: Road Datasets in Complex Mountain Environments (RDCME). RDCME comes from the QuickBird satellite, which is at an elevation between 1264 m and 1502 m with a resolution of 0.61 m; it contains 775 image samples, including red, green, and blue channels. Then, we propose the Light Roadformer model, which uses a transformer module and self-attention module to focus on extracting more accurate road edge information. A post-process module is further used to remove incorrectly predicted road segments. Compared with previous related models, the Light Roadformer proposed in this study has higher accuracy. Light Roadformer achieved the highest IoU of 89.5% for roads on the validation set and 88.8% for roads on the test set. The test on RDCME using Light Roadformer shows that the results of this study have broad application prospects in the extraction of mountain roads with similar backgrounds.

Keywords: road extraction; remote sensing; high-resolution remote sensing; semantic segmentation; transformer

Citation: Zhang, X.; Jiang, Y.; Wang, L.; Han, W.; Feng, R.; Fan, R.; Wang, S. Complex Mountain Road Extraction in High-Resolution Remote Sensing Images via a Light Roadformer and a New Benchmark. *Remote Sens.* **2022**, *14*, 4729. https://doi.org/10.3390/rs14194729

Academic Editor: Shuying Li

Received: 22 August 2022
Accepted: 18 September 2022
Published: 21 September 2022

1. Introduction

Mountainous road extraction is essential in spatial geographic information databases, which is significant for traffic navigation and military road planning [1]. However, the mountainous environment is desolate and complex, accompanied by frequent sandstorms, rainstorms, blizzards, and other poor weather. These environmental conditions may cause the loss of human lives and economic loss during long-distance driving. To safeguard the economy and people's safety, the extraction of mountain roads is becoming increasingly important.

HRSIs have become increasingly critical for geographic information system applications [2–4]. HRSIs are also an effective means of road extraction in mountainous areas. Research scholars have extracted information from the spectral features, shape features, and spatial relationships of HRSIs before classifying and identifying roads [5–7]. Currently, much road extraction work has been performed on urban road datasets [8–17], but there is a lack of extraction work on mountain roads. On the one hand, there is a lack of road datasets in the complex environment of mountainous areas, and on the other hand, there is a lack of effective models for road extraction in mountainous areas.

Road extraction work used to be manually annotated; however, even though the manual process is accurate, it is time-consuming. The task of road extraction has received more attention with the advancement of computer applications. The existing methods can be divided into heuristic methods and data-driven methods. Because of the different degrees of interaction, heuristic road extraction methods can be divided into semi-automatic and automatic extraction methods. The mainstream methods of the semi-automatic methods are the active contour model [18], dynamic programming [19–21], and template matching [22–25]. The semi-automatic methods require human intervention, so these methods are less efficient. The most commonly used automatic methods are segmentation methods [26–31], edge analysis methods [32,33], object-based methods [34,35], and multispectral segmentation methods [36,37].

In the context of big data and deep learning, semantic segmentation methods in data-driven methods have become mainstream. Data-driven methods have been proposed using deep learning. The proposed methods cannot scale effectively to challenging large datasets because of the small size of neural networks and the lack of big data [9]. After deep learning was proposed [38], Minh et al. [9] first attempted to extract roads using deep learning and achieved significant improvements in precision and recall. Because distinguishing categories cannot rely on a single pixel, Mnih et al. [39] predicted every small patch of labels from one large image context by using a patch-based deep convolutional neural network (DCNN). However, the overlay of the patches and duplicates of the adjacent pixels made the prediction process time-consuming and inefficient. In 2015, the FCN was proposed using the pixel-level classification method [40]. The FCN replaced the last fully connected convolutional layer. Here, the FCN was applied to extract roads in 2016 [41]. Deconvnets was proposed based on the FCN, and Segnet [42], Deeplab [43], and U-Net [44] replaced the interpolation layers as deconvolutional layers (called the decoder). In 2014, generative adversarial nets (GANs) were designed [45], consisting of a generator and a discriminator. GANs were then used in road extraction work and achieved better extraction results [46,47]. However, the GANs models may suffer from non-convergence and vanishing gradients. Based on pixel-level segmentation, the model output is noisy. The above data-driven methods result in the poor continuity of the extracted road segments [48]. Iterative road tracking [49] and polygon detection [50] are both based on graph-based methods, which are vectorized representations and show higher connectivity, but the graph reconstruction and optimization process are complex. Attention-like mechanisms were introduced in the 1990s. Based on this mechanism, attention was added to RNNs to learn the important part of an image [51]. Attention mechanisms were also applied in NLP by Bahdanau et al. [52]. Many researchers also attempted to use attention on CNNs [53]. In 2017, self-attention was proposed to replace the RNN or CNN, which was shown to perform well on NLP [54]. Compared with the CNN and RNN, the attention mechanism has fewer parameters and lower model complexity. In 2020, the first pure transformer structure achieved outstanding results in computer vision [55], and then, various transformer variants such as T2T-ViT [56], IPT [57], PVT [58], and SwinTransformer [59] were conceived of. The self-attention and transformer modules showed better performance than the FCN, being able to capture the global information from the whole image and focusing on the crucial details. However, the above methods are more suitable for extracting urban roadsthan extracting mountain roads.

Additionally, in recent years, some road datasets have been published, mainly for urban road extraction [60–62]. Wang introduced the TorontoCity [63] benchmark, which covers the full Greater Toronto Area (GTA). RoadTracer, which covers high-resolution ground truth road network maps of 40 urban cores in six countries, was proposed in 2018 [64]. In 2019, the Toulouse Road Network Dataset was released for road network extraction from remote sensing images [65]. In 2021, Xu et al. proposed Topo-boundaries for offline topological road boundary detection [66], which contains 25,295 1000 × 21,000-sized four-channel aerial images.Although many urban road datasets have been proposed, the road datasets are not accurate enough because cities are developing and changing all the time. At the same time, there is also a lack of mountain road datasets and network

models suitable for mountain road extraction. In 2021, Zhou et al. [67] proposed a split depthwise (DW) separable graph convolutional network (SGCN) and a mountain road dataset. The SGCN achieved good accuracy on the Massachusetts road dataset. However, the classical network model was not very accurate on the mountain road dataset. Hence, the accuracy of different network models on Zhou's mountain dataset is mixed.In 2021, the DSDNet [68] also achieved good results on mountain road extraction, but it is limited by the threshold in the post-processing. The NIGAN [69] was proposed in 2022 and achieved good results on mountain road extraction. However, the resolution of the dataset used in this work is low for extracting roads, and the NIGAN does not solve the problem of poor road extraction in complex environments such as shadow areas.

Compared with urban roads, mountain roads have two characteristics. Figure 1 shows that mountain roads are small in size and have blurred edges. Figure 2 shows that mountain roads have high similarity with terrain objects with respect to the topological and morphological features. Both of these features make the extraction of mountain roads challenging.

Figure 1. Characteristics of roads in different datasets: (**a**) is a clear road example in DeepGlobe; (**b**) is a road with a smaller edge blur width in RDCME.

Figure 2. Roads in DeepGlobe are easier to identify in (**a**), while roads in (**b**) from RDCME are challenging to extract because they are similar in shape to other features.

To address the low-precision segmentation of mountain roads because of vague road edges and complex backgrounds, we first labeled RDCME collected from high-resolution remote sensors. Then, the transformer-based road extraction model Light Roadformer was used. A self-attention module and pyramid structure were employed to attend to the entire image information and focus on local details. A post-process was applied to remove incorrectly classified road segments based on road topological features. Finally, we tested our model on RDCME with other road extraction models, and two road segments from the area were extracted to test the performance.

2. Materials and Methods

2.1. Study Area

RDCME is located in the northwest of China, and it has two characteristics as mentioned above. First, RDCME has a small size and blurred edges. The road edges in DeepGlobe are clear and bulky, while the roads in RDCME are covered more, with the edges being blurred, as shown in Figure 1. In addition to complex environmental factors such as mountain weathering caused by frequent dust storms, road surfaces and road boundaries are blurred due to the limited quality of the satellite imagery. Additionally, mountain roads are smaller in size and situated in complex environments. The actual width of the roads in mountainous regions is about 4–5 m, and the road objects are 12–15 pixels in width in the HRSIs. Second, mountainous roads have high similarity in their topological and morphological features with the terrain objects. In a complex mountainous area, there are various terrain objects, such as rivers, mountains, and dunes. Geologically, geographicalfeatures such as rivers, dunes, and ridgelines have linear representations. The roads in DeepGlobe are clearly differentiated from cities, towns, and farmland. However, the roads in RDCME are morphologically similar to the geologicallines, as Figure 2 shows.

2.2. Datasets

The remote sensing image data are from the QuickBird satellite, which uses Ball Aerospaces Global Imaging System 2000 (BGIS 2000). The satellite collects panchromatic (black and white) imagery at a 0.61 m resolution and multispectral imagery at a 2.44 m to 1.63 m resolution. The main objects of our study area are roads, rivers, and mountains, and the elevation is between 1264 m and 1502 m.

The remote sensing images we chose for road extraction include red, green, and blue channels with a resolution of 0.61 m. HRSIs could appear blurry with low confinement because of the camera angles, sunlight angles, shadows from mountains, and the motion of satellites. Removing the received corrupted images, we selected 22 HRSIs containing mountain roads from the study area. The size of image samples ranges from 1536 $\times$ 2048 pixels to 12,288 $\times$ 13,312 pixels. In general, we chose an experimental area with a more complex geological environment, as shown in Figure 3. We located the roads by visual recognition and labeled the roads segments at the pixel level with painting in the program Krita. To train the models and test the performance, 20 HRSIs were used to make the datasets, and the remaining two image samples were predicted by the model.

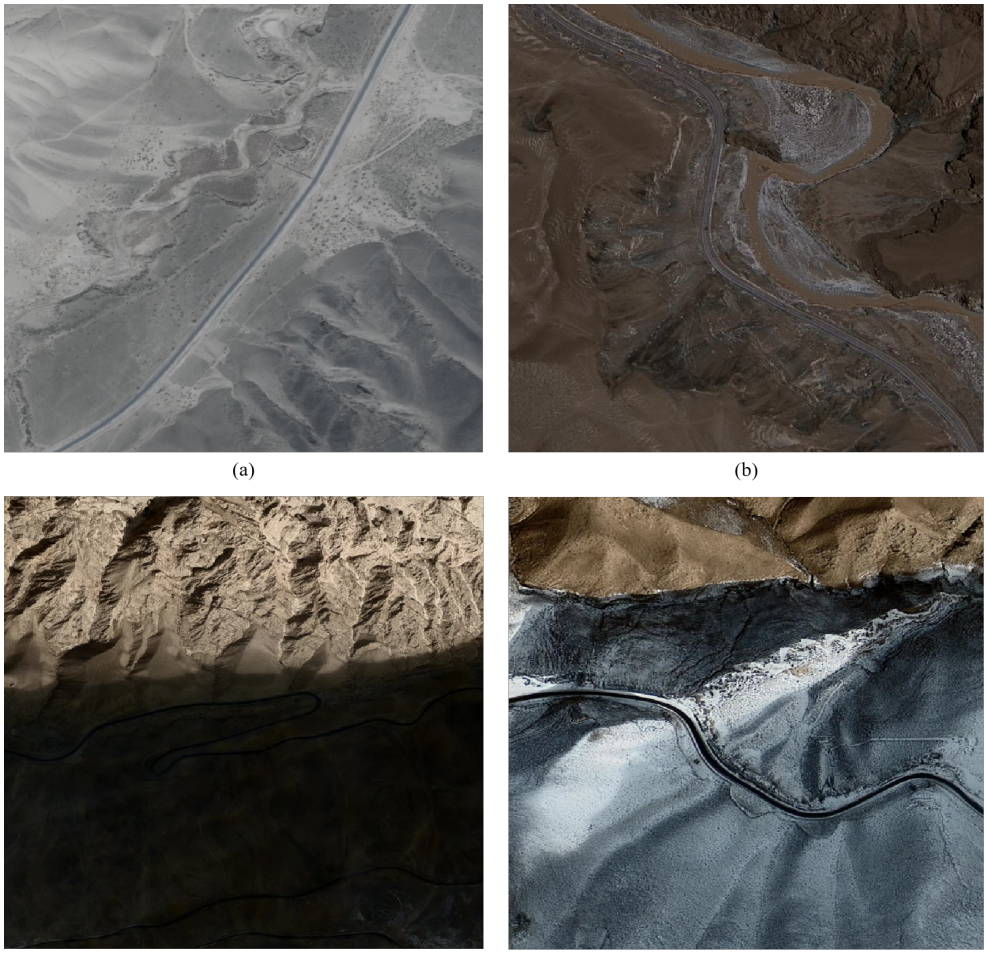

Figure 3. These pictures show the roads in a complex environment in RDCME. (**a**) is a road on a hill; (**b**) is a road near a river; (**c**) is a road in the shadow of a mountain; (**d**) is a road on a ridgeline.

2.3. Method

2.3.1. Image Pre-Processing

In the shadow of the mountains in the study area, road objects may be dark and difficult to segment from their surroundings. Contrast limited adaptive histogram equalization (CLAHE) [70] has been proposed on the basis of adaptive histogram equalization (AHE) [71] to increase the global contrast of image samples. CLAHE was used for tiles in the image, not the entire image. To remove artificial boundaries, we combined adjacent tiles using bilinear interpolation. The effect of the CLAHE image enhancement method is shown in Figure 4: as shown in the images, the low contrast problem and dim image problem are solved. In the enhanced images, road objects are easy to identify.

2.3.2. Light Roadformer Model

In the current paper, we propose Light Roadformer to extract roads for a mountainous area. The architecture of the Light Roadformer model is shown in Figure 5. The Light Roadformer model consists of two parts: an encoder and a decoder module. We also adjusted some parameters of the Light Roadformer model to improve the performance of the model. The num-layers of the segformer in Light Roadformer was adjusted to be (3, 6, 32, 3), which improved the overall performance of the model.

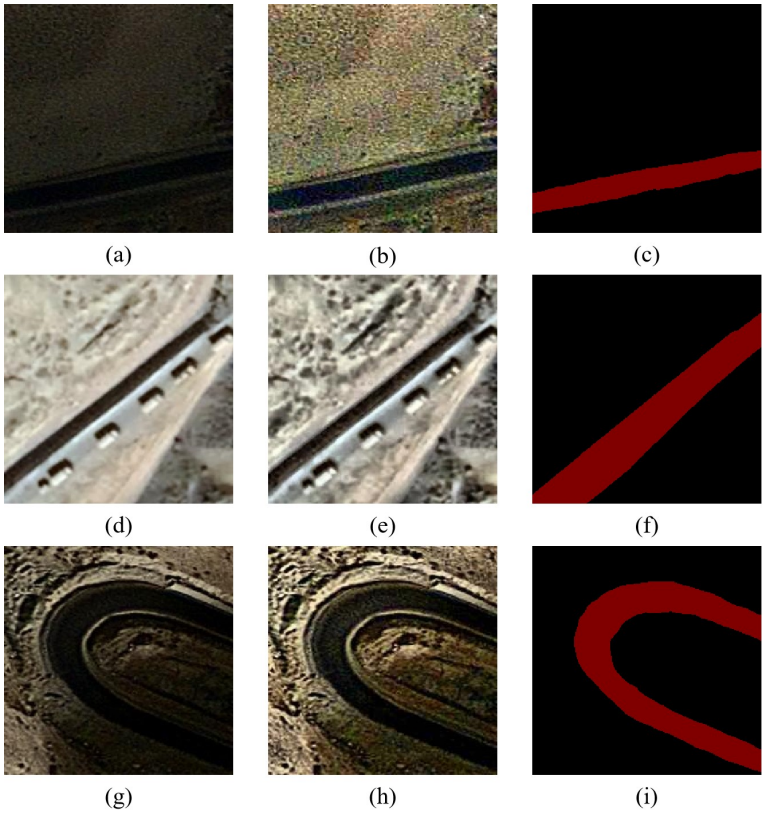

Figure 4. (**a**,**d**,**g**) is the original unprocessed image; (**b**,**e**,**h**) are the images after enhancement; (**c**,**f**,**i**) are the labels of RDCME.

- *Encoder:*
  The encoder consists of a pyramid structure, extracting high-resolution coarse features and low-resolution fine-grained features, which help enhance the performance of the segmentation: for every transformer block, there is a self-attention layer, feed-forward network, and overlap patch merging.
  The attention module maps the query and sets the key-value pairs to the output: first, a compatibility function is used to calculate the weight from the query and the corresponding key of dimension d_{head}. Then, the weight is assigned to each value, and the weighted sum of the values is computed as the output. The attention module can be described as in Equation (1) and (a) in Figure 6.

$$Attention(O, P, Q) = Softmax(\frac{OP^T}{\sqrt{k_{h(ead)}}})V \tag{1}$$

  where O represents the query, P represents the key, and Q represents the value.
  The complexity of the attention module is $O(n^2)$. The complexity of the self-attention mechanism is reduced using a reduction ratio R, which is expressed as:

$$\hat{P} = Reshape(\frac{N}{R}, C \cdot R)(P) \tag{2}$$

$$P = Linear(C \cdot R, C)(\hat{P}) \tag{3}$$

  where the shape of P is (N, C); Equation (2) reshapes P to $\hat{P}$, whose shape is $(\frac{N}{R}, C \cdot R)$; Equation (3) converts $\hat{P}$ to the origin sequence; the complexity is reduced from $O(N^2)$ to $O(\frac{N^2}{R})$.
  Instead of using single attention, multihead attention was applied in the model. The multihead attention projects queries, keys, and values multiple times, and then,

an attention function is performed in parallel. This results in the concatenated final values, as shown in (b) of Figure 6. The role of the multihead attention model structure is to jointly learn information from different representation subspaces.
The mix-feed-forward network (Mix-FFN) was used to learnthe location information, after the self-attention module. In vision transformer (ViT) [55], positional encoding (PE) is used to introduce the location information. However, the fixed PE leads to a drop in accuracy when interpolated. Here, 3×3 conv was applied to alleviate the problem in [72]. Here, Xie et al. [73] showed that a 3×3 convolution could provide positional information and replaced the PE with the FFN. The Mix-FFN is formulated as follows:

$$M_{out} = MLP(GELU(Conv_{3\times3}(MLP(m_{in})))) + M_{in} \tag{4}$$

where M_{in} is the feature from the self-attention module. The Mix-FFN mixes a 3×3 convolution and an MLP into each FFN.

- *Decoder:*
The decoder part first unifies the channel dimension by an MLP layer in Equation (5), and all the features from different transformer blocks are up-sampled and concatenated in Equation (6); then, the feature is fused by an MLP layer. Finally, the segmentation mask M is predicted from the fused feature in Equation (8).
The decoder part gathers the extracted features and draws the road extraction map, which consists of four steps:

1. Unify the channel dimension by an MLP layer:

$$\hat{F}_i = Linear(C_i, C)(F_i), \forall i \tag{5}$$

where F_i represents the feature extracted from the transformer blocks, C_i represents the channel number of the output features, C represents the channel number of the feature, and $\hat{F}_i$ is the unified feature.

2. All the features of different sizes are upsampled to the same size:

$$\hat{F}_i = Unsample(\frac{W}{4} \times \frac{H}{4})(\hat{F}_i), \forall i \tag{6}$$

where W is the image width and H is the image height.

3. The features are concatenated and fused by an MLP layer.

$$F = Linear(4C, C)(Concat(\hat{F}_i)), \forall i \tag{7}$$

4. The mask M is segmented from the fused feature by an MLP layer to produce the extracted road segments.

$$M = Linear(C, N_{cls})(F) \tag{8}$$

where N_{cls} represent the class number of objects and M is the final mask predicted.

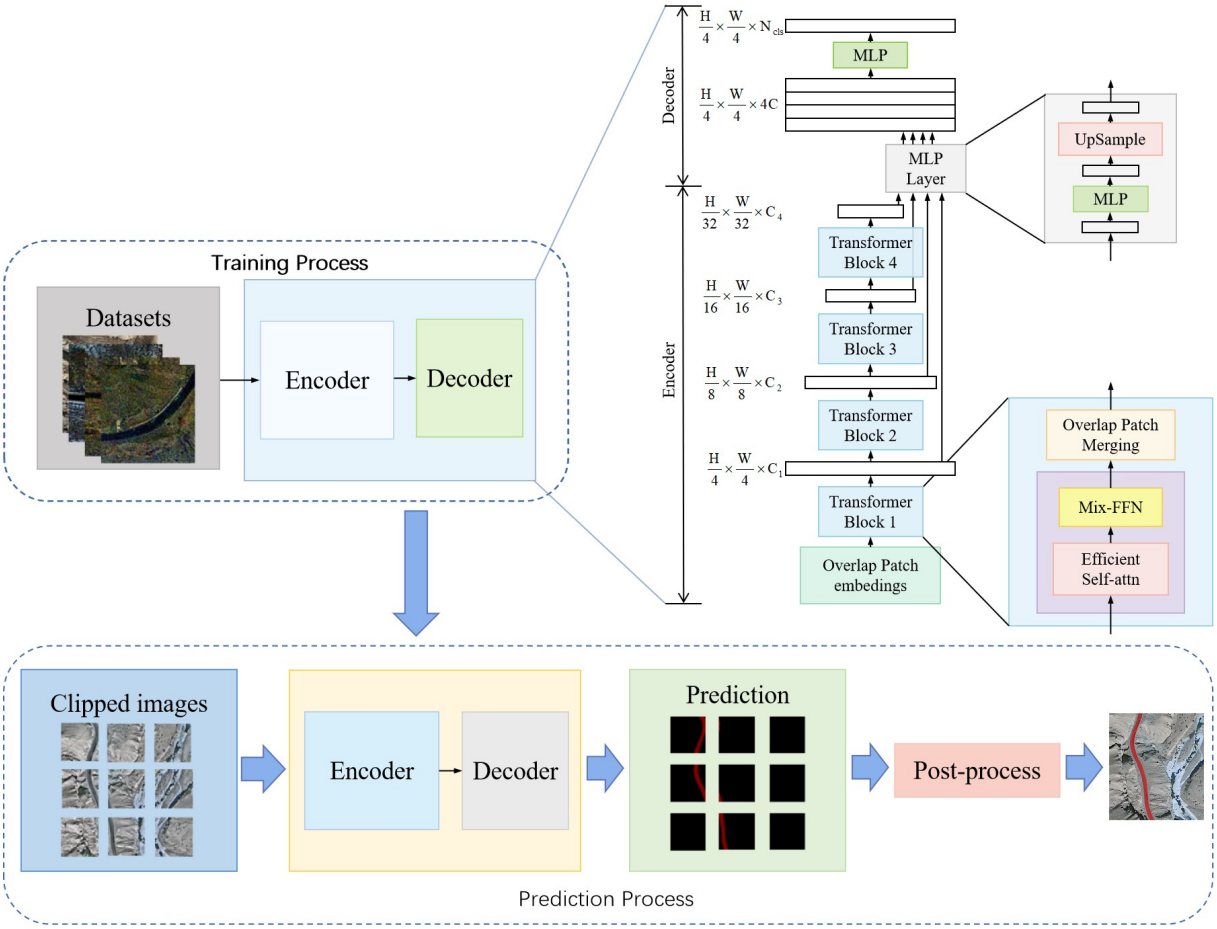

Figure 5. The Light Roadformer model consists of two parts, training and prediction, where both the training and prediction parts contain an encoding and decoding process. In the prediction module, we added post-processing steps to achieve more accurate prediction results.

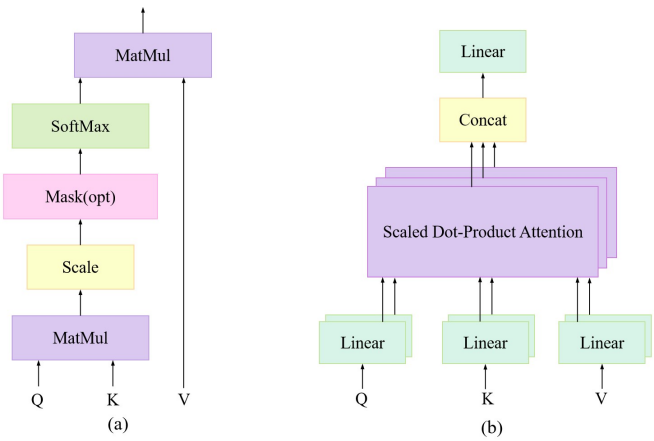

Figure 6. (**a**) is the scaled dot-product attention module, and (**b**) is the multi-head attention module.

2.4. Post-Process

Combined with the characteristics of mountain roads, we post-processed the predicted results. When the study area is predicted, the image is clipped into small image patches, called tiles. Furthermore, due to the different characteristics of mountainous and urban roads, the roads on each tile in RDCME are not as dense as the urban road dataset, which increases the probability of misclassification, so post-processing also affects the prediction results in the Light Roadformer key factor. Every tile is predicted and combined into

one whole image. Because the road segment is always connected to other road segments, the wrongly classified pixels will not be connected to other predicted road segments. Based on this characteristic, the wrongly predicted road segments can be easily removed by checking the size of road segments. Our post-processing work uses the DFS algorithm [74]. DFS traverses or searches a tree or graph, going through the nodes of the tree along the depth of the tree and searching deep into the branches. When the edge of the node v is searched or the node does not meet the conditions during the search process, the search will go back until the starting node of the edge of node v is found. The whole process of the algorithm is iteratively repeated until all nodes have been visited. This blind search method is less efficient, but more effective.

Post-processing can be divided into the following steps:

1. First, traverse the whole resulting image to find the predicted road pixel and ignore the non-road pixel.
2. Second, for each road pixel, take every pixel that is predicted as the road and connect this to itself as in the same road segment; then, search the connected pixel to calculate the size of the road segments.
3. Third, according to the ratio of the extracted road segment area to the entire image area and whether the road segment is connected to the edge of the image, the non-road segment is removed.

3. Results

3.1. Experimental Setting and Evaluation Metrics

An Intel CPU i7 11700K, NVIDIA GeForce RTX 3090 GPU, with 24 GB graphic, and 32 GB memory were employed to conduct the experiments. The operating system was Windows 11. The batch size was set to use the graphics card memory for every model fully. The graphics card memory of the model was fully utilized for the batch size settings. The models were constructed based on the MMSegmentation [75] toolbox and the PyTorch [76] framework, while a total of 40,000 iterations were performed.

The evaluating metrics of the road methods were precision and intersection over union (IoU). The evaluation index used to describe the accuracy of road area segmentation was the IoU, and the calculation formula is:

$$IoU = J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{9}$$

where A is the road reference and B is the road obtained by segmentation.

3.2. Prediction Experiment

Figure 7 shows the predicted results of Light Roadformer. The images in the left column are image samples; the middle column is the labeled ground truth; the right column is the predicted image of the model. In Figure 7a, although the road pixels take up a small proportion of the whole sample, the IoU of the road is 89%. In Figure 7d,g, the model reached a high IoU of 97% and 96% despite the cars as noise.

To test the Light Roadformer model's performance on the dataset test set, two road segments were selected from the study area, which were not included in the dataset. Firstly, the remote sensing image was divided into image patches, all of whose sizes were 256×256. Then, the road segments were extracted with Light Roadformer, and the image patches were combined into one. After that, we applied road post-processing to obtain the final predictions, and the result is shown in Figures 8 and 9.

Figures 8a and 9a are the original remote sensing images for the road segments; the labels are shown in Figure 8b,c; Figure 9b,c show the segment result of the Light Roadformer model. Since the image is divided into multiple patches, the corner of the patch is wrongly segmented, and parts of the river and mountain shadow are also classified into roads due to their striped shape, which is similar to the road objects. The road $IoUs$ are 65.46% and 79.55%. After the post-processing on the predicted roads, the wrongly classified

road segments are removed, and the *IoUs* are 84.47% and 87.11%, which is significantly boosted.

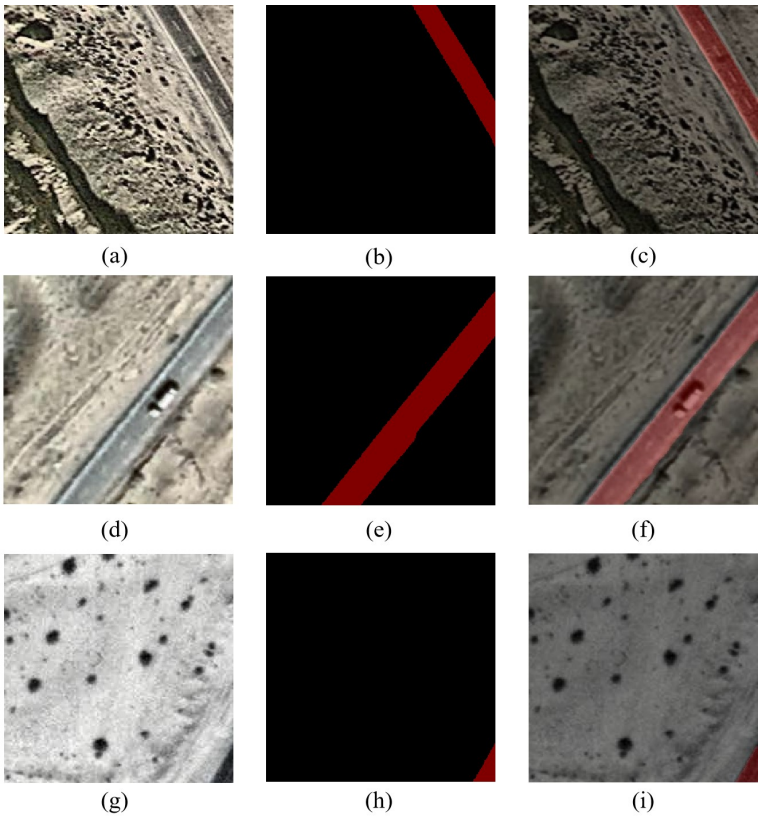

Figure 7. (**a**,**d**,**g**) are the original unprocessed images; (**b**,**e**,**h**) are the labels of RDCME; (**c**,**f**,**i**) are the predicted results.

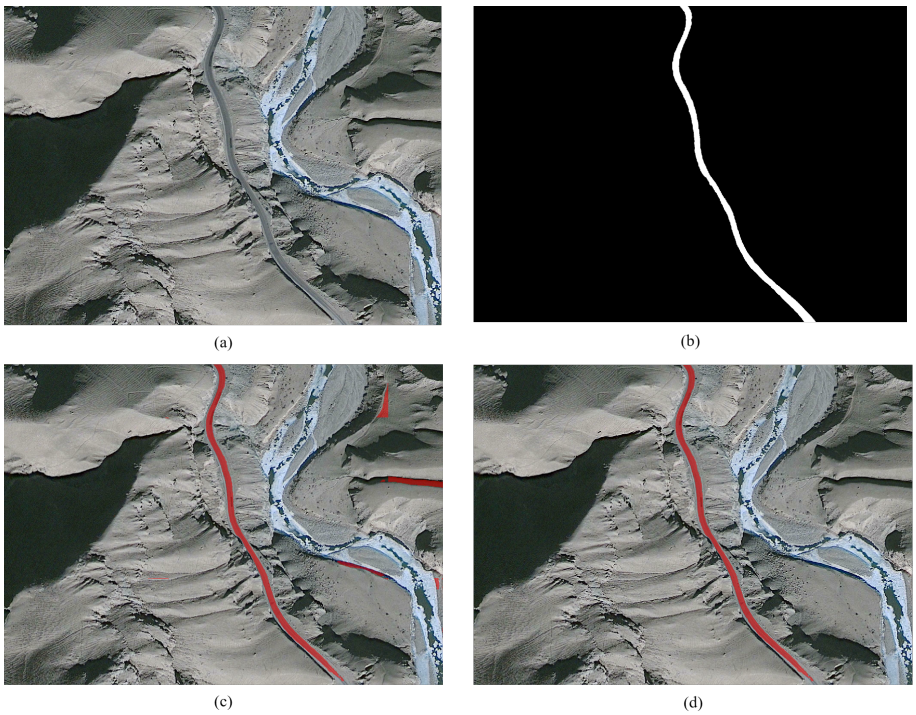

Figure 8. The first image result used for prediction. (**a**) is the original imagery; (**b**) is the manual road label; (**c**) is the prediction roads of Light Roadformer; (**d**) is the prediction after post-process.

(a) (b) (c) (d)

Figure 9. The second image result used for prediction. (**a**) is the original imagery; (**b**) is the manual road label; (**c**) is the prediction roads of Light Roadformer; (**d**) is the prediction after post-process.

4. Discussion

4.1. Comparison with the Existing Datasets

RDCME was compared with other public road extraction datasets: DeepGlobe [77] and Massachusetts [39]. The DeepGlobe road extraction dataset is in RGB, with a size of 1024 × 1024 pixels, was collected by DigitalGlobe's satellite, and has a 0.5 m resolution. The Massachusetts roads dataset covers an area of 2.25 square kilometers and includes 1,171 remote sensing images, each with a size of 1500 × 1500 pixels. The Massachusetts roads dataset mainly contains various urban, suburban, and rural areas. RDCME aims at roads in mountainous areas and excludes the roads in villages and cities. Table 1 shows the comparison of RDCME, DeepGlobe, and Massachusetts.

Table 1. Comparison of RDCME, DeepGlobe, and Massachusetts.

Dataset Name	Number of Images	Resolution	Image Width
DeepGlobe	6226	1.2 m	1024
Massachusetts	1171	0.5 m	1500
RDCME	775	0.6 m	256

4.2. Experimental Comparison

In this experiment, we compared Light Roadformer with a series of milestone semantic segmentation models to test the performance, and the semantic segmentation included a dual attention network (Danet) [78], criss-cross network (CCNet) [79], context encoding network (EncNet) [80], short-term dense concatenate network (STDC2) [81], unified perceptual parsing network (UPerNet) [82], global context network (GCNet) [83], bilateral segmentation network (BiSeNet V2) [84], and high-resolution network (HRNet) [85]. The *IoU* and parameter numbers were calculated on the dataset.

The IoU in the training process is shown in Figure 10. The Light Roadformer model reached 89.6%, surpassing other semantic segmentation models; the IoU for the Light Roadformer on the test set was 88.8%. The parameter numbers and road IoU of the models are shown in Table 2. The parameter number for Light Roadformer was 68,719,810. By comparison, Light Roadformer outperformed the other road extraction models while maintaining a moderate size of the parameters.

Figure 10. Comparison of the semantic segmentation model's training process.

Table 2. The performance of different network models on RDCME.

Model Name	Training IoU	Testing IoU	Parameters Counts
Danet	88.26	87.35	68,808,170
Light Roadformer	89.50	**88.80**	68,719,810
CCNet	87.82	86.96	49,812,965
EncNet	87.33	86.63	54,862,278
STDC2	87.73	86.59	12.596,904
UPerNet	88.42	85.93	85,394,276
BiSeNetv2	85.75	85.75	14,804,255
HRNet	88.48	86.68	65,846,402
GCNet	88.52	87.12	68,609,253

Compared to other network models, Light Roadformer achieves the highest accuracy.

5. Conclusions

In the current study, with the aim of road extraction work in mountain areas, a mountain road extraction dataset was first manually labeled. Then, we proposed Light Roadformer model to address the road extraction in the mountain environment. The model uses a self-attention module to focus on road edge details and uses a pyramid structure to obtain high-resolution coarse features and low-resolution fine features, providing better segmentation for mountain roads. A post-process module is also used to remove the incorrectly segmented road based on the road topological features. The model reached an 88.8% IoU for roads in the dataset that were manual labeled, outperforming other road extraction models. The validation of the remote sensing images showed good potential for road extraction in mountainous areas.

Author Contributions: X.Z., conceptualization, data curation, methodology, writing—original draft, and writing—review and editing; Y.J., conceptualization, data curation, methodology, and writing—original draft; L.W., conceptualization, funding acquisition, methodology, writing—review and editing, and project administration; W.H., investigation and data curation; R.F. (Ruyi Feng), writing—review and editing; R.F. (Runyu Fan), writing—review and editing; S.W., writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: The work was supported by the National Natural Science Foundation of China under Grants U21A2013, 42201415 and 41925007; the Hubei Natural Science Foundation of China (No. 2019CFA023); the Fundamental Research Founds for the Central Universities, China University of Geosciences (Wuhan) (No. 162301212697).

Data Availability Statement: The dataset is available at https://github.com/zxy1211/RDCME, accessed on 21 August 2022.

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

References

1. Zhou, M.; Sui, H.; Chen, S.; Wang, J.; Chen, X. BT-RoadNet: A boundary and topologically-aware neural network for road extraction from high-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *168*, 288–306. [CrossRef]
2. Panteras, G.; Cervone, G. Enhancing the temporal resolution of satellite-based flood extent generation using crowdsourced data for disaster monitoring. *Int. J. Remote Sens.* **2018**, *39*, 1459–1474. [CrossRef]
3. Han, W.; Feng, R.; Wang, L.; Cheng, Y. A semi-supervised generative framework with deep learning features for high-resolution remote sensing image scene classification. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 23–43. [CrossRef]
4. Han, W.; Chen, J.; Wang, L.; Feng, R.; Li, F.; Wu, L.; Tian, T.; Yan, J. A survey on methods of small weak object detection in optical high-resolution remote sensing images. *IEEE Geosci. Remote. Sens. Mag.* **2021**, *9*, 8–34. [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. Han, W.; Li, J.; Wang, S.; Zhang, X.; Dong, Y.; Fan, R.; Zhang, X.; Wang, L. Geological Remote Sensing Interpretation Using Deep Learning Feature and an Adaptive Multisource Data Fusion Network. *IEEE Trans. Geosci. Remote. Sens.* **2022**, *60*, 1–14. [CrossRef]
7. Han, W.; Li, J.; Wang, S.; Wang, Y.; Yan, J.; Fan, R.; Zhang, X.; Wang, L. A context-scale-aware detector and a new benchmark for remote sensing small weak object detection in unmanned aerial vehicle images. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102966. [CrossRef]
8. Mathibela, B.; Newman, P.; Posner, I. Reading the road: Road marking classification and interpretation. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 2072–2081. [CrossRef]
9. 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.
10. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual u-net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
11. Zhou, L.; Zhang, C.; Wu, M. D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 182–186.
12. 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]
13. Gao, L.; Song, W.; Dai, J.; Chen, Y. Road extraction from high-resolution remote sensing imagery using refined deep residual convolutional neural network. *Remote Sens.* **2019**, *11*, 552. [CrossRef]
14. Wulamu, A.; Shi, Z.; Zhang, D.; He, Z. Multiscale road extraction in remote sensing images. *Comput. Intell. Neurosci.* **2019**, *2019*, 2373798. [CrossRef] [PubMed]
15. Courtial, A.; El Ayedi, A.; Touya, G.; Zhang, X. Exploring the potential of deep learning segmentation for mountain roads generalisation. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 338. [CrossRef]
16. Kolhe, A.; Bhise, A. Modified PLVP with Optimised Deep Learning for Morphological based Road Extraction. *Int. J. Image Data Fusion* **2022**, *13*, 155–179. [CrossRef]
17. Chen, Z.; Deng, L.; Luo, Y.; Li, D.; Junior, J.M.; Gonçalves, W.N.; Nurunnabi, A.A.M.; Li, J.; Wang, C.; Li, D. Road extraction in remote sensing data: A survey. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102833. [CrossRef]
18. Kass, M.; Witkin, A.; Terzopoulos, D. Snakes: Active contour models. Inter-national. *J. Comput.* **2004**, *1*, 321–331.
19. Gruen, A.; Li, H. Road extraction from aerial and satellite images by dynamic programming. *ISPRS J. Photogramm. Remote Sens.* **1995**, *50*, 11–20. [CrossRef]
20. Dal Poz, A.; Do Vale, G. Dynamic programming approach for semi-automated road extraction from medium-and high-resolution images. *ISPrS Arch.* **2003**, *34*, W8.
21. Dal Poz, A.P.; Gallis, R.A.; Da Silva, J.F. Three-dimensional semiautomatic road extraction from a high-resolution aerial image by dynamic-programming optimization in the object space. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 796–800. [CrossRef]
22. Park, S.R. Semi-automatic road extraction algorithm from IKONOS images using template matching. In Proceedings of the 22nd Asian Conference on Remote Sensing, Singapore, 5–9 November 2001.
23. Lin, X.; Shen, J.; Liang, Y. Semi-automatic road tracking using parallel angular texture signature. *Intell. Autom. Soft Comput.* **2012**, *18*, 1009–1021. [CrossRef]
24. Fu, G.; Zhao, H.; Li, C.; Shi, L. Road detection from optical remote sensing imagery using circular projection matching and tracking strategy. *J. Indian Soc. Remote Sens.* **2013**, *41*, 819–831. [CrossRef]

25. Lin, X.; Zhang, J.; Liu, Z.; Shen, J. Semi-automatic extraction of ribbon roads from high resolution remotely sensed imagery by T-shaped template matching. In Proceedings of the Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Classification of Remote Sensing Images, Guangzhou, China, 28–29 June 2008; SPIE: Bellingham, WA, USA, 2008; Volume 7147, pp. 168–175.
26. Kirthika, A.; Mookambiga, A. Automated road network extraction using artificial neural network. In Proceedings of the 2011 International Conference on Recent Trends in Information Technology (ICRTIT), Chennai, India, 3–5 June 2011; pp. 1061–1065.
27. Miao, Z.; Wang, B.; Shi, W.; Wu, H.; Wan, Y. Use of GMM and SCMS for accurate road centerline extraction from the classified image. *J. Sens.* **2015**, *2015*, 784504. [CrossRef]
28. Li, L.; Zhang, X. A quickly automatic road extraction method for high-resolution remote sensing images. *Geomat. Sci. Technol.* **2015**, *3*, 27–33. [CrossRef]
29. Abdollahi, A.; Bakhtiari, H.R.R.; Nejad, M.P. Investigation of SVM and level set interactive methods for road extraction from google earth images. *J. Indian Soc. Remote Sens.* **2018**, *46*, 423–430. [CrossRef]
30. Li, R.; Cao, F. Road network extraction from high-resolution remote sensing image using homogenous property and shape feature. *J. Indian Soc. Remote Sens.* **2018**, *46*, 51–58. [CrossRef]
31. Zhang, J.; Chen, L.; Zhuo, L.; Geng, W.; Wang, C. Multiple Saliency Features Based Automatic Road Extraction from High-Resolution Multispectral Satellite Images. *Chin. J. Electron.* **2018**, *27*, 133–139. [CrossRef]
32. Baumgartner, A.; Steger, C.; Mayer, H.; Eckstein, W. Multi-resolution, semantic objects, and context for road extraction. In *Semantic Modeling for the Acquisition of Topographic Information from Images and Maps*; Birkhäuser Verlag: Basel, Switzerland, 1997; pp. 140–156.
33. Karaman, E.; Çinar, U.; Gedik, E.; Yardımcı, Y.; Halıcı, U. Automatic road network extraction from multispectral satellite images. In Proceedings of the 2012 20th Signal Processing and Communications Applications Conference (SIU), Mugla, Turkey, 18–20 April 2012; pp. 1–4.
34. Ding, L.; Yang, Q.; Lu, J.; Xu, J.; Yu, J. Road extraction based on direction consistency segmentation. In *Chinese Conference on Pattern Recognition*; Springer: Singapore, 2016; pp. 131–144.
35. Huang, X.; Zhang, L. Road centreline extraction from high-resolution imagery based on multiscale structural features and support vector machines. *Int. J. Remote Sens.* **2009**, *30*, 1977–1987. [CrossRef]
36. Mitra, P.; Shankar, B.U.; Pal, S.K. Segmentation of multispectral remote sensing images using active support vector machines. *Pattern Recognit. Lett.* **2004**, *25*, 1067–1074. [CrossRef]
37. Maboudi, M.; Amini, J.; Hahn, M.; Saati, M. Object-based road extraction from satellite images using ant colony optimization. *Int. J. Remote Sens.* **2017**, *38*, 179–198. [CrossRef]
38. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [CrossRef]
39. Mnih, V. *Machine Learning for Aerial Image Labeling*; University of Toronto (Canada): Toronto, ON, Canada, 2013.
40. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
41. Zhong, Z.; Li, J.; Cui, W.; Jiang, H. Fully convolutional networks for building and road extraction: Preliminary results. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1591–1594.
42. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
43. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
44. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*; Springer: Cham, Switzerland, 2015, pp. 234–241.
45. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems 27 (NIPS 2014), Montreal, QC, Canada, 8–13 December 2014; Volume 27.
46. Shi, Q.; Liu, X.; Li, X. Road detection from remote sensing images by generative adversarial networks. *IEEE Access* **2017**, *6*, 25486–25494. [CrossRef]
47. Costea, D.; Marcu, A.; Slusanschi, E.; Leordeanu, M. Creating roadmaps in aerial images with generative adversarial networks and smoothing-based optimization. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 2100–2109.
48. Batra, A.; Singh, S.; Pang, G.; Basu, S.; Jawahar, C.; Paluri, M. Improved road connectivity by joint learning of orientation and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10385–10393.
49. Lian, R.; Huang, L. DeepWindow: Sliding window based on deep learning for road extraction from remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1905–1916. [CrossRef]
50. Li, Z.; Wegner, J.D.; Lucchi, A. Topological map extraction from overhead images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 1715–1724.

51. Mnih, V.; Heess, N.; Graves, A.; et al. Recurrent models of visual attention. In Proceedings of the Advances in Neural Information Processing Systems 27 (NIPS 2014), Montreal, QC, Canada, 8–13 December 2014; Volume 27.
52. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
53. Yin, W.; Schütze, H.; Xiang, B.; Zhou, B. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 259–272. [CrossRef]
54. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
55. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
56. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.H.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 1–17 October 2021; pp. 558–567.
57. Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; Gao, W. Pre-trained image processing transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12299–12310.
58. 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.
59. 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, Virtual, 11–17 October 2021; pp. 10012–10022.
60. Lin, Y.; Xu, D.; Wang, N.; Shi, Z.; Chen, Q. Road extraction from very-high-resolution remote sensing images via a nested SE-Deeplab model. *Remote Sens.* **2020**, *12*, 2985. [CrossRef]
61. Shao, Z.; Zhou, Z.; Huang, X.; Zhang, Y. MRENet: Simultaneous extraction of road surface and road centerline in complex urban scenes from very high-resolution images. *Remote Sens.* **2021**, *13*, 239. [CrossRef]
62. Ding, L.; Bruzzone, L. DiResNet: Direction-aware residual network for road extraction in VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 10243–10254. [CrossRef]
63. Wang, S.; Bai, M.; Mattyus, G.; Chu, H.; Luo, W.; Yang, B.; Liang, J.; Cheverie, J.; Fidler, S.; Urtasun, R. Torontocity: Seeing the world with a million eyes. *arXiv* **2016**, arXiv:1612.00423.
64. 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, 18–22 June 2018; pp. 4720–4728.
65. Belli, D.; Kipf, T. Image-conditioned graph generation for road network extraction. *arXiv* **2019**, arXiv:1910.14388.
66. Xu, Z.; Sun, Y.; Liu, M. Topo-boundary: A benchmark dataset on topological road-boundary detection using aerial images for autonomous driving. *IEEE Robot. Autom. Lett.* **2021**, *6*, 7248–7255. [CrossRef]
67. Zhou, G.; Chen, W.; Gui, Q.; Li, X.; Wang, L. Split depth-wise separable graph-convolution network for road extraction in complex environments from high-resolution remote-sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–15. [CrossRef]
68. Xu, Z.; Shen, Z.; Li, Y.; Xia, L.; Wang, H.; Li, S.; Jiao, S.; Lei, Y. Road extraction in mountainous regions from high-resolution images based on DSDNet and terrain optimization. *Remote Sens.* **2020**, *13*, 90. [CrossRef]
69. Chen, W.; Zhou, G.; Liu, Z.; Li, X.; Zheng, X.; Wang, L. NIGAN: A Framework for Mountain Road Extraction Integrating Remote Sensing Road-Scene Neighborhood Probability Enhancements and Improved Conditional Generative Adversarial Network. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
70. Pizer, S.M.; Amburn, E.P.; Austin, J.D.; Cromartie, R.; Geselowitz, A.; Greer, T.; ter Haar Romeny, B.; Zimmerman, J.B.; Zuiderveld, K. Adaptive histogram equalization and its variations. *Comput. Vision, Graph. Image Process.* **1987**, *39*, 355–368. [CrossRef]
71. Ketcham, D.J.; Lowe, R.W.; Weber, J.W. *Image Enhancement Techniques for Cockpit Displays*; Technical Report; Hughes Aircraft Co Culver City Ca Display Systems Lab: Culver City, CA, USA, 1974.
72. Chu, X.; Tian, Z.; Zhang, B.; Wang, X.; Wei, X.; Xia, H.; Shen, C. Conditional positional encodings for vision transformers. *arXiv* **2021**, arXiv:2102.10882.
73. 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.
74. He, S.; Bastani, F.; Jagwani, S.; Park, E.; Abbar, S.; Alizadeh, M.; Balakrishnan, H.; Chawla, S.; Madden, S.; Sadeghi, M.A. Roadtagger: Robust road attribute inference with graph neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 10965–10972.
75. Contributors, M. MMSegmentation: Openmmlab Semantic Segmentation Toolbox and Benchmark. 2020 Availabe online: https://github.com/open-mmlab/mmsegmentation (accessed on 18 May 2022).
76. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. In Proceedings of the Advances in Neural Information Processing Systems 32 (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; Volume 32.

77. 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.
78. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
79. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 603–612.
80. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7151–7160.
81. Fan, M.; Lai, S.; Huang, J.; Wei, X.; Chai, Z.; Luo, J.; Wei, X. Rethinking BiSeNet for real-time semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9716–9725.
82. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified perceptual parsing for scene understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 418–434.
83. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019.
84. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 3051–3068. [CrossRef]
85. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3349–3364. [CrossRef] [PubMed]