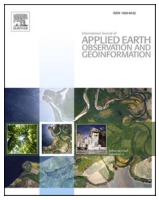
International Journal of Applied Earth Observation and Geoinformation 124 (2023) 103510

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



DPENet: Dual-path extraction network based on CNN and transformer for accurate building and road extraction

Ziyi Chen [a], Yuhua Luo [a], Jing Wang [a], Jonathan Li [b], Cheng Wang [c], Dilong Li [a,*]

[a] Department of Computer Science and Technology, Fujian Key Laboratory of Big Data Intelligence and Security, Xiamen Key Laboratory of Data Security and Blockchain Technology, Huaqiao University, 668 Jimei Road, Xiamen, FJ 361021, China
[b] Department of Geography and Environmental Management and Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada
[c] School of Informatics, Xiamen University, Xiamen, FJ 361005, China

ARTICLE INFO

Keywords:
Remote sensing
Deep learning
Building extraction
Road extraction

ABSTRACT

The acceleration of urbanization and the increasing demand for precise city planning have made the extraction of buildings and roads from remote sensing images crucial. Deep learning-based methods have propelled the progress of object extraction technology, but there are still challenges such as the missing and incomplete extraction of buildings and roads for small objects and occlusions. To address this issue, we propose a dual-path extraction network based on CNN and Transformer, combining local and global features to fully extract the semantic information of objects. To further enhance the semantic reconstruction capability of features, this paper introduces a multi-scale upsampling mechanism, thereby expanding the visual range of reconstruction. Finally, we adopt a deep supervision strategy to improve the reconstruction accuracy of objects at different resolutions. Our method has been tested on four remote sensing image datasets and has achieved excellent IoU scores on all datasets (Massachusetts Building and Roads Dataset: 76.69% and 66.41%, LRSNY and CHN6-CUG Roads Dataset: 88.96% and 61.99%). Furthermore, our method demonstrates superior performance compared to other mainstream image segmentation algorithms, fully demonstrating the effectiveness of our approach.

1. Introduction

Object extraction is a vital aspect of remote sensing image processing as it involves the precise and efficient identification and extraction of objects of interest from remote sensing imagery (Chen et al., 2022c). In recent years, the rapid advancement of high-resolution optical remote sensing technology has made object extraction from high-resolution optical images a prominent research focus (Guan et al., 2021).

High-resolution optical remote sensing images offer exceptional spatial resolution and abundant information content, allowing for the detection of even subtle changes on the Earth's surface (Li et al., 2021b; Liu et al., 2023a; Mao et al., 2023). As a result, utilizing high-resolution optical remote sensing images for object extraction becomes the preferred approach to achieve accurate object identification and spatial analysis (Chen et al., 2022b; Chen et al., 2021c; Guan et al., 2022).

Object extraction holds immense practical value and finds application in a wide range of fields (Zhu et al., 2020). During emergency earthquake response, the rapid extraction of building damage levels in affected areas provides indispensable reference information for rescue operations (Li et al., 2021c; Xu et al., 2018). In the realm of smart city development, object extraction enables automatic identification and monitoring of urban infrastructure, road traffic flow, land utilization, and other aspects, thus offering decision support for smart city planning and management (Ding et al., 2021; Yan et al., 2022). In the field of automotive navigation, precise vehicle positioning and navigation guidance can be achieved by extracting objects such as roads and buildings (Xu et al., 2021; Zhou et al., 2022). Additionally, object extraction technology plays a significant role in critical areas such as ecological environment monitoring, agricultural production, and military applications (Li et al., 2021a; Waldner and Diakogiannis, 2020).

However, object extraction from high-resolution optical remote sensing images poses several challenges, including the complexity of the imagery, object diversity, interference from lighting conditions, trees, and shadow effects, as illustrated in Fig. 1 (c) and (d). These factors often

* Corresponding author.
E-mail addresses: chenziyihq@hqu.edu.cn (Z. Chen), wroaring@hqu.edu.cn (J. Wang), junli@uwaterloo.ca (J. Li), cwang@xmu.edu.cn (C. Wang), scholar.dll@hqu.edu.cn (D. Li).

https://doi.org/10.1016/j.jag.2023.103510
Received 28 June 2023; Received in revised form 24 September 2023; Accepted 27 September 2023
Available online 10 October 2023

Fig. 1. Illustrative examples of building and road objects in remote sensing images.

lead to insufficient object extraction, resulting in issues of omission and incorrect identification. Moreover, in urban and rural environments, building objects may exhibit substantial variations in appearance and contextual surroundings. To address these challenges, researchers have proposed various algorithms and techniques, leveraging feature extraction, machine learning, and deep learning approaches. These methods drive the progress and development of object extraction technology. However, based on our investigation, the current algorithms for extracting objects from high-resolution remote sensing images still suffer from limitations in feature extraction and information reconstruction, especially when dealing with small or intricately shaped object objects, as depicted in Fig. 1 (a) and (b). The small size of building objects and the complex shapes of road objects pose significant challenges for semantic segmentation.

To overcome the above challenges, this study primarily investigates three aspects of research. Firstly, we introduce a novel dual-path object extraction network based on CNN and Transformer. The network employs the CNN branch to extract spatially detailed contextual information from the features of remote sensing images. Concurrently, the Transformer branch captures global object dependencies from the global contextual information of the remote sensing images. By combining these two paths, the network effectively integrates global information with local spatial details, facilitating the comprehensive extraction of building or road objects.

Additionally, to ensure that the decoder in the network can effectively handle diverse image features and improve the reconstruction capability of the network, we enhance the visual range of multi-sampling in the dual-path network extraction mechanism. By incorporating multi-view upsampling convolutional operations, we increase the complexity of the network's reconstruction structure, thereby augmenting its ability to extract building and road objects in complex environments.

In order to enhance object extraction at different resolutions and achieve precise reconstruction, we implement a deep supervision strategy within the framework of multiple upsampling mechanisms. Specifically, within the reconstruction branch, we incorporate multiple segmentation heads of varying sizes to supervise the object extraction process at different scales. This approach ensures the accurate extraction of building or road objects.

The paper makes significant contributions in the following three aspects:

(1) We propose a novel dual-path object extraction network that combines the strengths of CNN and Transformer. The network integrates a spatial detail branch and a global semantic branch, leading to exceptional precision in extracting building and road objects.
(2) To enhance the network's reconstruction capability for building and road objects, we introduce a multi-view fusion-based multi-sampling mechanism in the reconstruction branch. This mechanism ensures that the network captures finer object details, significantly improving its ability to reconstruct these objects.
(3) We employ a deep supervision strategy by designing segmentation heads at different levels in the reconstruction branch. This strategy allows for comprehensive supervision of the reconstruction process for building and road object features at different resolutions.

2. Related work

2.1. Building extraction

Building object extraction is a computer vision technique that aims to automatically detect and locate buildings in images or videos. It enables the identification of buildings within an image and the extraction of their bounding boxes or contours, thereby supporting applications such as building recognition, map creation, and urban planning.

In recent years, significant breakthroughs have been achieved in building segmentation and extraction based on Convolutional Neural Networks (CNN) Considering the rapid advancements in deep learning and computer vision. Ding et al. (2022) introduced Adversarial Shape Learning Network (ASLNet) to model the shape patterns of buildings and enhance the precision of building object extraction. However, challenges arise in the form of missing and incomplete objects, especially when extracting larger objects, due to variations in color and texture within buildings. To overcome these issues, Shao et al. (2020) designed a method called Building Region Refinement Network (BRRNet), comprising a prediction module and a residual refinement module. This method demonstrates better handling of building object extraction and reduces the occurrence of missing and incomplete objects. For achieving highly accurate building object extraction, Guo et al. (2020) introduced an attention-based multi-loss neural network. By incorporating attention modules, they improved the sensitivity of the model to critical features and suppressed the influence of irrelevant feature regions. To fully utilize features at different levels in object extraction, Liu et al. (2019) introduced a novel Fully Convolutional Network (FCN). This network captures and aggregates multi-scale contextual information by progressively fusing multi-level features, thereby achieving semantic understanding of object images. Xu et al. (2022) utilized rich background features in remote sensing images to assist in object extraction, preserving the shape of the extracted results. Their method effectively enhances the accuracy of building extraction. Zhu et al. (2020) proposed a novel Multi-Participation Path Neural Network (MAPNet) for accurate extraction of multi-scale building contours and boundaries, demonstrating excellent performance in capturing building shapes.

Deng et al. (2021) successfully differentiated building objects from complex environmental objects by designing grid-based attention gating and dilated convolutional pyramid modules. However, insufficient utilization of multi-scale building object features resulted in issues such as blurry edges when extracting complex-shaped building objects. To

address this problem, Guo et al. (2022) proposed a coarse-to-fine Building Boundary Extraction Network (CBR-Net). Feng et al. (2022) enhanced the extraction of high-frequency and low-frequency information in remote sensing images and proposed a Spectrum Intensity Attention Network (FSIANet), thereby enhancing the expressive power of building semantic features. Due to domain differences among different remote sensing image datasets, the generalization ability of the network on unknown datasets is relatively poor. To overcome this, Peng et al. (2021) presented a global domain adaptive extraction network that effectively enhanced the network's generalization capability across different domain images. Liu et al. (2022) obtained vector components of building objects by establishing an additional edge segmentation branch and performed semantic segmentation based on the obtained vector components as guidance, ensuring the accuracy and shape of the extraction results. Wei and Ji (2021) proposed the utilization of Graph Convolutional Networks (GCN) to generate building vector maps automatically from aerial images. This approach facilitates the polygonal prediction of building objects.

On the other hand, Transformer-based techniques have demonstrated significant advancements in the field of natural language processing (NLP), and in recent years, they have also been applied to computer vision tasks, including building image segmentation and extraction. Chen et al. (2021a) investigated the application of Transformers in building extraction and devised a streamlined dual-channel Transformer architecture to maximize efficiency. This architecture enables the model to capture long-range dependencies in both spatial and channel dimensions, thereby enhancing the performance of building extraction. Chen et al. (2021b) introduced a U-Net network that combines Transformer self-attention and reconstruction deviation modules, effectively enhancing the capability of semantic segmentation and achieving extraction of complex building objects. In their work, Wang et al. (2022) introduced a novel approach named BuildFormer, which is based on Vision Transformer (ViT). It features a dual-path structure and allows for the application of large windows to capture global context, greatly improving its potential in handling large-scale remote sensing images. However, inadequate integration of global and local information can still result in incomplete, false, or missing extraction results. To mitigate these issues, Xu et al. (2023) presented a novel segmentation approach based on the Bi-branch Cross-fusion Transformer Network (BCTNet).

2.2. Road extraction

Road object extraction plays a significant role in computer vision applications, aiming to accurately identify and extract road-related objects from images or videos. Numerous studies have employed deep learning models such as Convolutional Neural Networks (CNNs) and Transformers to enhance the accuracy and robustness of object extraction.

In recent years, Convolutional Neural Networks (CNNs) have emerged as a powerful tool for road extraction tasks. Tan et al. (2021) designed a novel end-to-end road segmentation method that effectively leverages convolutional layers at different levels, enhancing the model's accurate perception of road edges and shapes, and mitigating the imbalance between CNN network depth and spatial resolution. To address challenges such as complex backgrounds, high density, insufficient training data, or high manual annotation costs in road extraction, Shamsolmoali et al. (2021) introduced domain adaptation-based methods for synthesizing images to meet the requirements of road extraction. To reduce the need for a large training dataset, efforts have been made to address the challenge of data acquisition, Hu et al. (2021) presented an improved Generative Adversarial Network called WSGAN, which utilizes weakly supervised methods for efficient road object extraction. Inspired by human pose estimation work, Lian and Huang (2020) proposed an automated road object extraction method called DeepWindow. They employed a CNN-based decision function to guide the sliding window for highly accurate road object extraction. Wei et al. (2020) presented a novel multi-level framework based on deep learning, which utilizes accelerated segmentation, multi-start tracking, and fusion mechanisms for extracting objects such as road surfaces and centerlines. To improve road connectivity and maintain precise alignment between images and real roads, Tan et al. (2020) designed an iterative graph exploration approach guided by point and segment cues. Dai et al. (2021) developed a model-driven to sample-driven road extraction method based on road geometry features. Zhou et al. (2020) introduced a boundary and topology-aware road extraction framework (BT-RoadNet) to address issues such as boundary quality, noise, and occlusion in existing automatic road extraction methods. Zhang et al. (2019)presented an approach for road extraction that leverages Fully Convolutional Networks (FCNs) and incorporates strategies to address the issue of imbalanced road-background data distribution.

Chen et al. (2021c)designed an asymmetric road extraction network that utilizes an end-to-end CNN model. By introducing multi-level upsampling biases, they improved the segmentation accuracy and performance of road objects. Chen et al. (2022a) converted binary classification maps into continuous symbolic distance maps for input transformation and proposed a distance-based road extraction method, which effectively alleviates the issue of discontinuity in road extraction. By combining the dual-branch multi-task structure and integrating road boundary details with road intersection information, Chen et al. (2022b) designs a road extraction framework that learns collaborative feature representations and enhances road connectivity. Zhou et al. (2021) introduced a deep separable graph convolutional network (SGCN) that enhances the expression capability of road features by capturing global contextual information of channel and spatial features. Lu et al. (2022) introduced a cascaded multi-task framework for road extraction, which effectively extracts road surfaces, centerlines, and edges while enhancing road connectivity. To reduce dependence on annotated data, Chen et al. (2023) presented a semi-supervised road extraction framework, known as SemiRoadExNet, which utilizes Generative Adversarial Networks (GANs) for improved performance. By utilizing multiple discriminators, to ensure coherence in feature distributions between annotated and unannotated data, they enforce consistency in the distribution of features, thereby enhancing the model's generalization ability.

Similarly, research on road extraction methods based on Transformers has achieved remarkable improvements. CNNs struggle to effectively capture global representations. To address this, Luo et al. (2022) presented a bidirectional Transformer network called BDTNet, based on a hybrid encoder-decoder architecture, this enhances the capture of both global and local information in aerial imagery, improving the extraction process. To overcome the challenge of CNNs in capturing contextual information effectively, Yang et al. (2022) introduced a road extraction approach for remote sensing images that integrates high-level semantic features with foreground context information, enhancing the accuracy of the extraction process. Jiang et al. (2022) developed a pyramid-based vision Transformer network specifically designed for the extraction of roads in remote sensing images. By adopting a multi-view contextual observation strategy to obtain higher-quality token embeddings, they effectively enhance the quality and robustness of feature representation. Tao et al. (2023) introduced a road model called Seg-Road, which improves road connectivity. While a convolutional neural network (CNN) structure is employed to extract local context information for improved road detail segmentation, they utilize a Transformer architecture to capture long-range dependencies and incorporate global contextual information, further enhancing road segmentation in the images. Zhang et al. (2022) developed a mountain road extraction network called Light Roadformer, based on Transformer and self-attention modules, to accurately extract road objects in environments with blurred road edges and sand coverage. Given the limitations of CNN convolutional kernels in capturing long-range information and global context, their performance is suboptimal in

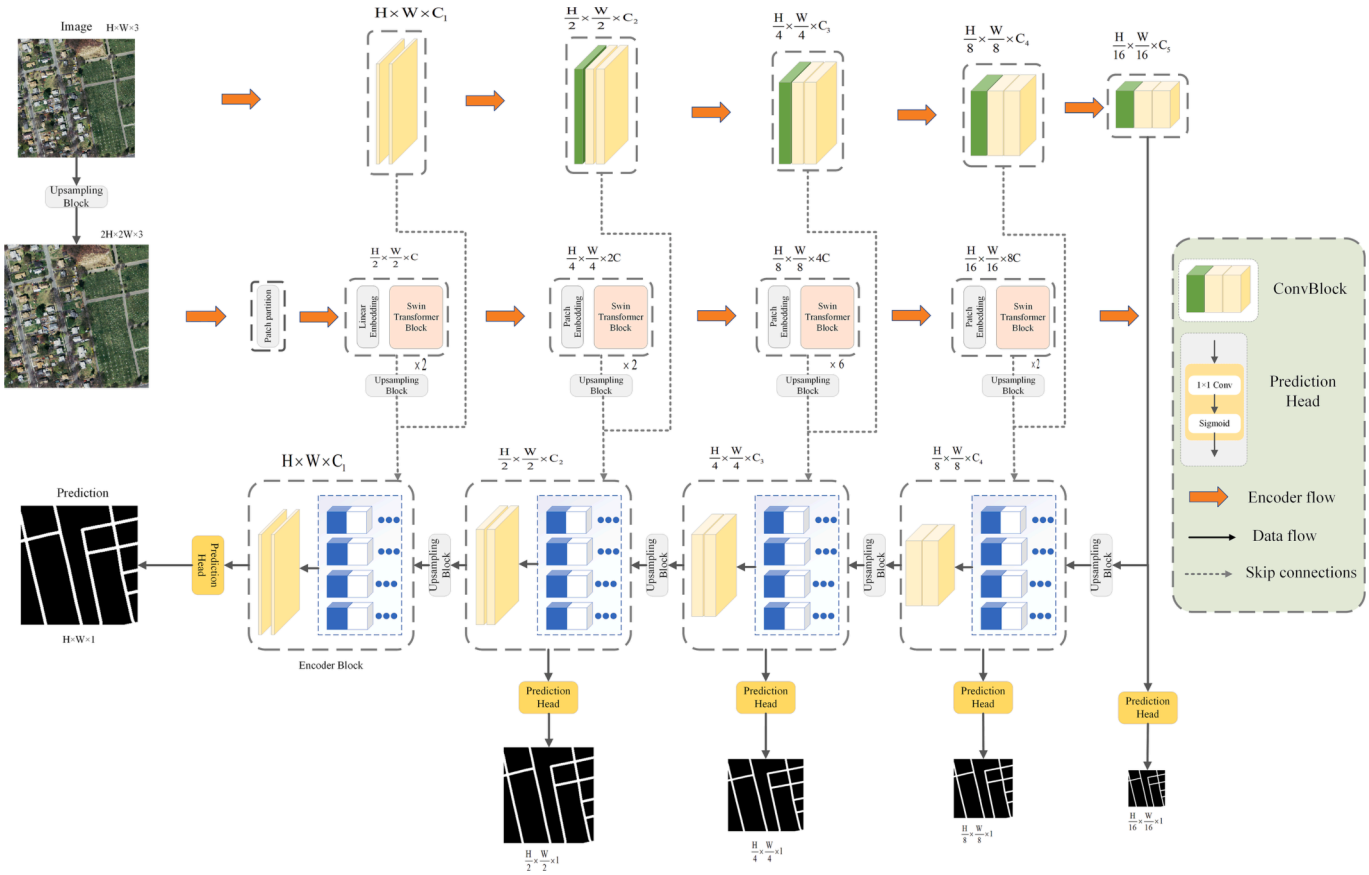


Fig. 2. Overall architecture of the designed DPENet.

scenarios involving road objects distributed over long distances and highly structured environments. To overcome this challenge, Liu et al. (2023b) introduced an innovative model named RoadFormer, which utilizes the Swin Transformer as its backbone. This approach effectively captures complex road structures and improves the performance of road extraction tasks.

3. Method

In this section, we will initially introduce the general framework of our proposed Dual-Path Extraction Network (DPENet). Subsequently, we will introduce the key modules employed in the framework through various branches. Finally, we will provide a brief overview of the loss functions used in this paper.

3.1. Network structure

As shown in Fig. 2, our proposed Dual-Path Extraction Network (DPENet) consists of three main components: a CNN-based spatial detail extraction branch, a Transformer-based global information extraction branch, and a feature reconstruction branch. The network architecture we employ is based on an encoder-decoder framework, with two distinct encoder parts: the spatial detail branch and the global information branch.

3.1.1. Spatial detail extraction branch based on CNN

The CNN-based spatial detail extraction branch consists of five extraction stages. In the first stage of this branch, the image is processed through two 3X3 convolutional layers, followed by batch normalization (BN) and ReLU activation function, transforming the channel size of the image features from 3 to 64. To prevent difficulties in gradient propagation caused by an excessive number of convolutional layers, we incorporate a residual connection with a 1X1 convolutional kernel between the two convolutional modules. This process is illustrated in Fig. 3 (a). Next, the image features are passed to the second stage of the extraction branch, which includes a 2X2 max pooling layer and two 3X3 convolutional layers with BN and ReLU activation function, also

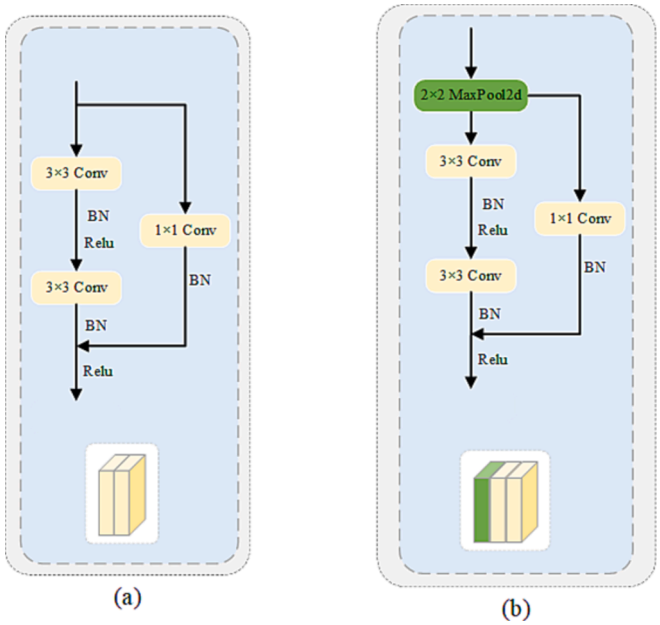


Fig. 3. Two types of Convolutional Blocks used in the spatial detail extraction branch.

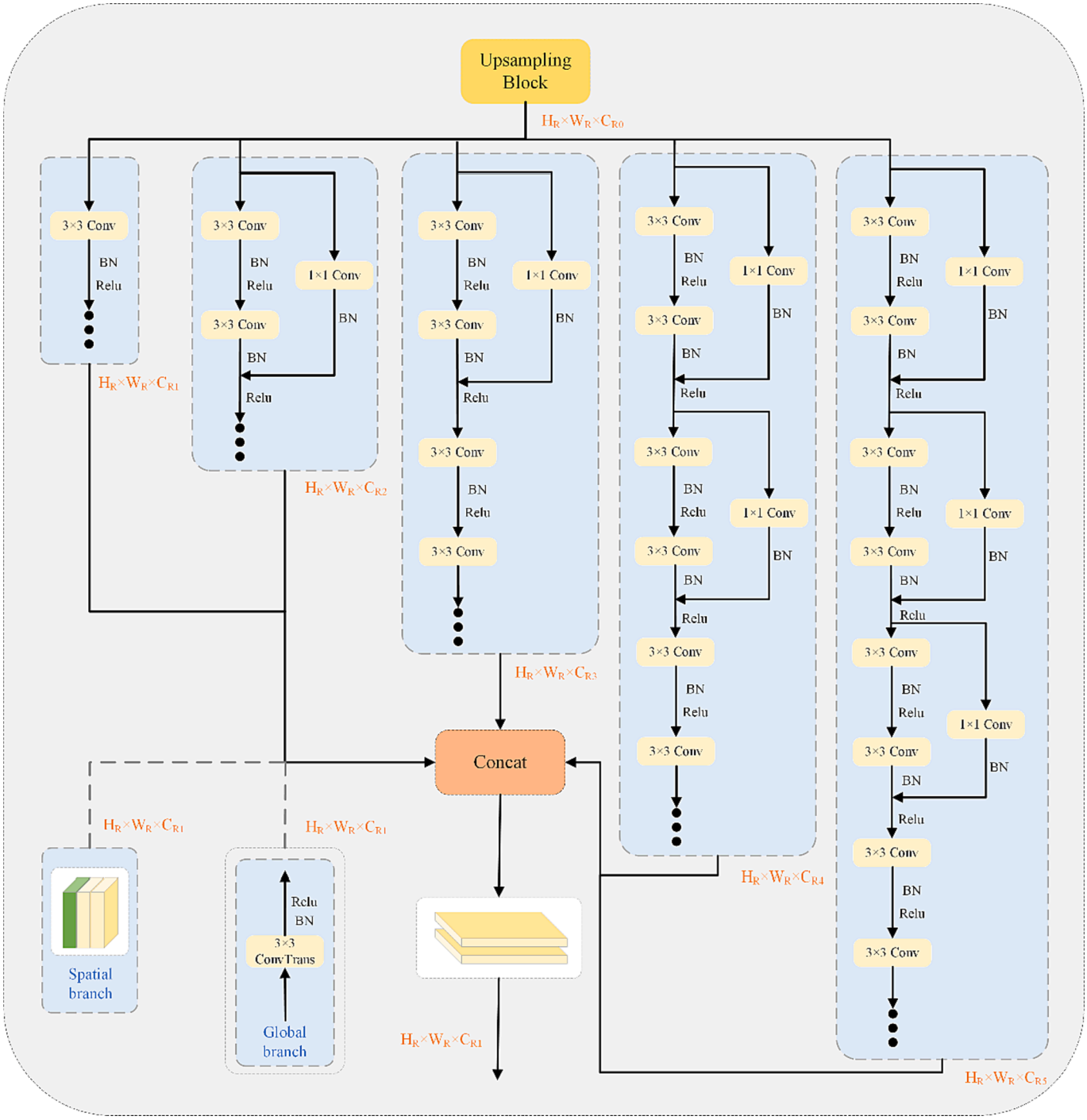


Fig. 4. A multi-residual combination mechanism.

utilizing a residual connection. This process is depicted in Fig. 3(b). The third, fourth, and fifth stages follow a similar extraction process as the second stage, progressively increasing the channel size of the image features and reducing the resolution to enhance the extraction capability of spatial details. In the first to fifth stages, the channel size of the image features increases from 64 to 128, 256, 512, and 1024, respectively, while the resolution of the image features gradually decreases from 256X256 to 128X128, 64X64, 32X32, and 16X16.

3.1.2. *Global information extraction branch based on Transformer*

The global information extraction branch in our design is based on the fine-tuned Swin Transformer(Liu et al., 2021). Swin Transformer possesses powerful capabilities in extracting global information, which is particularly crucial in semantic segmentation tasks. Therefore, it meets the requirements of our global information extraction branch. The branch can be divided into four stages. What we use is tiny Swin Transformer, as shown in Fig. 2. The network is divided into four stages, and its depths are [2,2,6,2] respectively. Each stage is Patch embedding and Swin Transformer block except the operation of the first stage is patch partition, Liner embedding and Swin Transformer block, and the remaining stages are patch embedding and Swin Transformer block. Moreover, the size of feature extraction in the first stage is H/2*W/2*C, and the feature size is reduced by two times and the number of channels is increased by two times in each stage after that.

In the first stage, we perform initial processing on the original 256X256 image by upsampling its resolution to 512X512. This is done to enable high-resolution image input into the global information extraction branch. In this stage, the high-resolution image undergoes block-wise processing using the Patch Partition module and then undergoes Patch Embedding. The resulting features are then fed into the Swin Transformer Block, producing output features with a resolution of 128X128.

Subsequently, these features enter the second, third, and fourth stages of the extraction branch. These stages consist of Patch Embedding and multiple Swin Transformer Blocks. Each stage produces image features with resolutions of 64X64, 32X32, and 16X16, respectively.

The window-based attention mechanism is computed as follows:

$$Attention(Q,K,V) = Soft\max(\frac{QK^T}{\sqrt{d}} + B)V \tag{1}$$

where Q, K, V denote Query, Key, Value, d denotes scaling factor, and B denotes relative position encoding, respectively.

3.1.3. Feature reconstruction branch

Object extraction networks typically consist of two components: a feature extraction part and a reconstruction output part. The reconstruction output part faces more challenging tasks compared to the feature extraction part, but it is often not enhanced in terms of structure, making it difficult to handle long-range dependencies between rich spatial details and global information. To enhance the reconstruction capability of our network under the dual-path feature extraction mechanism and improve the semantic perception ability of the extraction network, we introduce a multi-residual combination mechanism in the reconstruction part to improve the network's semantic reconstruction ability. The Encoder Block in Fig. 2 showcases the design of this network module.

Furthermore, in the multi-residual combination at each level, we employ a multi-branch mechanism and attempt to simulate semantic reconstruction under multiple perspectives by using residual convolution combinations of different depths. By enhancing the field of view through multiple sampling and increasing the complexity of the network reconstruction structure, we further enhance the ability of the extraction network to handle building and road objects in complex environments.

Assuming each upsampling layer adopts a 5-branch mechanism, the structural design of this layer is illustrated in Fig. 4. In the first stage, the Multi-Residual Group module simultaneously receives three input features: the input feature from the previous level, the input feature from the spatial branch, and the input feature from the global information. The main objective of the second stage is to process these three types of features.

Firstly, the input feature from the previous level is upsampled by a factor of 2, resulting in a feature map of size, and different depths of residual convolution combinations in the multi-branch mechanism are utilized to capture reconstructed semantic information from different perspectives. The basic module of the employed residual convolution combination consists of a 3X3 convolution, BN, ReLU activation function, and residual connection, with varying feature channel numbers in each residual convolution branch.

Secondly, the input feature from the global information branch undergoes deconvolution, BN, and ReLU processing to meet the requirements of size.

Next, the output features from different branches of the residual convolution combination, the spatial detail features of the same size, and the global information features are stacked together to obtain a fused feature map of size. This fused feature map not only incorporates the detail features provided by the spatial branch and the contextual dependencies from the global branch but also accommodates the high-level semantic information extracted from the dual-branch network.

Finally, the fused feature map is inputted into the Convolutional Blocks and then passed to the next layer of the encoder. The feature channel can be computed using the following formula:

$$C_{Rn} = \frac{C_{R0}}{2^n} \tag{2}$$

$$C_M = \sum_0^n \frac{C_{R0}}{2^n} + C_s + C_G \tag{3}$$

Here, $C_{R1}, C_{R2}, \ldots, C_{Rn}$ represents the feature channel number in the nth branch of the Rth residual combination group, C_M denotes the total channel sum in the entire module, C_s, C_G represent the channel numbers in the spatial branch and the global branch, respectively.

3.1.4. Deep monitoring strategy

The multi-residual module enhances the complexity of image features during reconstruction and enables multi-scale sampling of image features at different resolutions in various stages of the reconstruction part. However, solely outputting the reconstruction results in the final stage may not fully meet our requirements for accurate multi-scale reconstruction of objects at different resolutions. Therefore, under the mechanism of multi-scale upsampling, we employ a deep supervision strategy. Specifically, different scales of prediction heads are used in different stages of the reconstruction part to achieve supervision for object extraction at different scales, enabling precise extraction of road and building objects at different resolutions. As shown in Fig. 2, the prediction head structure in the low-resolution stage of the reconstruction part remains consistent with the final prediction head. It mainly consists of a 1X1 convolution layer and a Sigmoid activation function, and the output low-resolution reconstruction image maintains the same resolution as the current stage.

3.2. Loss function

During the learning process of the model, binary cross-entropy (BCE) loss function is one of the most commonly used loss functions. However, this loss function does not take into account the issue of sample imbalance. In semantic segmentation, there is an extreme imbalance between positive and negative samples, which requires a loss function capable of handling such cases. The Dice loss function is precisely designed to address this problem and is well-suited for our network's need to extract building and road objects. Therefore, in this study, we adopt a deep-supervised multi-level joint loss function that combines binary cross-entropy loss and Dice loss. Specifically, the binary cross-entropy loss can be expressed as follows:

$$L_{bce} = \frac{1}{N}\sum_i^N (g_i \log p_i + (1-g_i)\log(1-p_i)) \tag{4}$$

Among them, L_{bce} represents the BCE loss, p_i represents the predicted probability value of the i-th pixel in the image, g_i represents the ground truth of the i-th pixel in the image, and N represents the total number of pixels in the image.

The loss function used during the training of the network is Dice loss, which is calculated using the following formula:

$$L_{dice} = 1 - Dice \tag{5}$$

where L_{dice} represents the Dice loss and Dice is the coefficient of the loss.

The calculation of the Dice loss coefficient is given by the following formula:

$$Dice = \frac{\sum_i^N p_i \times g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2} \tag{6}$$

where p_i represents the predicted probability value of the i-th pixel in the image, g_i represents the ground truth value of the i-th pixel in the image, and N represents the total number of pixels in the image. The proposed

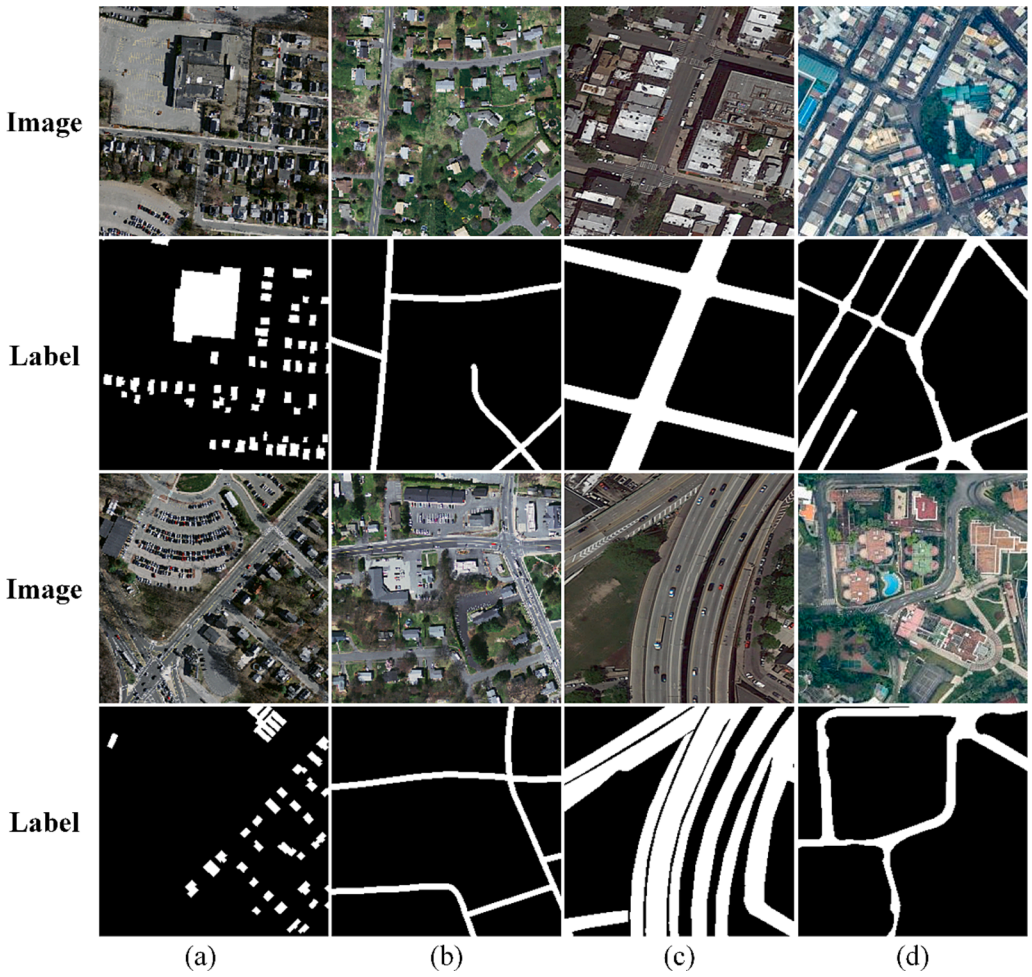


Fig. 5. Examples of remote sensing image datasets.

multi-level cascaded loss function is as follows:

$$L_{all} = \lambda_1 L_{bce}(P_1, G_1) + \lambda_2 L_{bce}(P_2, G_2) + \lambda_3 L_{bce}(P_3, G_3) + \lambda_4 L_{bce}(P_4, G_4) + \lambda_5 (L_{bce}(P_5, G_5) + L_{dice}(P_5, G_5)) \quad (7)$$

where P_1, P_2, P_3, P_4, P_5 represent the object prediction outputs at different resolutions in the reconstruction part. G_1, G_2, G_3, G_4, G_5 represent the corresponding ground truth objects at different resolutions, where the ground truth objects at different resolutions are obtained by down-sampling the original resolution ground truth using bilinear interpolation.$\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ represent the loss coefficients for different resolutions in the multi-level cascaded loss.

4. Experiments

4.1. Datasets

To evaluate the performance of our proposed DPENet in extracting building and road objects, we conducted experimental analysis using four publicly available remote sensing image datasets. These datasets include the Massachusetts Building Dataset(Hinton and Mnih, 2013), Massachusetts Roads Dataset(Hinton and Mnih, 2013), LRSNY Roads Segmentation Dataset(Chen et al., 2021d), and CHN6-CUG Roads Dataset(Zhu et al., 2021), as shown in Fig. 5(a), (b), (c), and (d) respectively.

1)Massachusetts Building Dataset:

The dataset is a collection of aerial building images captured from the Boston area in the United States. It consists of 151 aerial images with a resolution of 1500X1500 pixels, providing a spatial resolution of 1 m. This dataset covers buildings of various sizes and scales in both urban and suburban areas of Boston, with a total coverage area of 340 square kilometers. Due to the uniqueness of the dataset, evaluating building extraction models poses significant challenges. We followed the official data partition provided with the dataset, where 137 images were used for training, 10 images for validation, and an additional 4 images for testing. To adapt to the input size of our model, we cropped the images into 256X256 patches with a 64-pixel overlap region. As a result, we obtained a dataset comprising 8,768 training images, 256 validation images, and 640 testing images.

2)Massachusetts Roads Dataset

The road dataset consists of 1,171 satellite images with a resolution of 1500X1500 pixels. It includes 1,108 images for training, 14 images for validation, and 49 images for testing. This dataset covers a wide range of urban, suburban, and rural areas, with a total coverage area of approximately 2,600 square kilometers. The image resolution is 120 cm/pixel. Similar to the Massachusetts Building Dataset, we cropped the images into 256X256 patches without using any overlap regions. As a result, we obtained a dataset comprising 27,700 training images, 350 validation images, and 1,225 testing images.

3)LRSNY Roads Segmentation Dataset

The dataset is a large-scale road segmentation dataset sourced from optical remote sensing images of New York City. It covers a significant portion of the city center, with a spatial resolution of approximately 0.5 m. The dataset consists of 1,368 road object images with dimensions of 256X256 pixels. Among these, 716 images are used for training, 220 images for validation, and 432 images for testing purposes.

4)CHN6-CUG Roads Dataset

The dataset is sourced from Google Earth and includes road remote sensing images of six major cities in China. It consists of a total of 4,511 images, covering areas such as Chaoyang District in Beijing, Yangpu District in Shanghai, the central area of Wuhan, Nanshan District in Shenzhen, Sha Tin District in Hong Kong, and Macau. Each image has dimensions of 512X512 pixels with a resolution of 50 cm/pixel. The dataset comprises 3,608 images for training and 903 images for testing. Similarly, we have cropped the dataset to a size of 256X256 pixels, ensuring that there is no overlap between the cropped images.

4.2. Experimental settings

To improve the generalization ability of the DPENet network, we applied data augmentation techniques during the training process, including rotation, horizontal flipping, and vertical flipping. For model training, we used the Adam optimizer with a batch size of 16 and a learning rate of 0.001. If the test loss did not decrease after 5 epochs of training, we would halve the learning rate. The experiments were conducted under the PyTorch framework, and the training was performed for 100 epochs on the building and road datasets. We saved the model with the highest accuracy during the training process. The experimental setup included an Intel(R) Core(TM) i9-12900KF CPU running at 3.20 GHz, 128 GB of RAM, and two GPUs (NVIDIA GeForce RTX 3090) for training.

4.3. Evaluation indicators

The extraction of building and road objects from high-resolution satellite imagery is commonly regarded as a binary semantic segmentation problem. To evaluate the object extraction performance of the network framework, we employed five widely used evaluation metrics, including overall accuracy, precision, recall, F1 score, and intersection over union (IoU). Overall accuracy measures the overall correctness by considering the classification accuracy of all samples without considering the classes. Precision represents the percentage of pixels predicted as building or road objects and correctly classified within the correct regions, while recall represents the proportion of correctly predicted pixels in the ground truth of building or road objects. The F1 score is the harmonic mean of precision and recall, and IoU measures the ratio of the intersection area between the predicted building or road objects and the ground truth to the union area. These evaluation metrics can be expressed using the following formulas after calculating the number of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) of pixels:

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

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1 = \frac{2 \times recall \times precision}{recall + precision} \tag{11}$$

$$IoU = \frac{TP}{TP + FN + FP} \tag{12}$$

4.4. Quantitative analysis

In this section, we will compare the performance of DPENet with several state-of-the-art methods for building and road object extraction on four remote sensing image datasets, namely the Massachusetts Building Dataset, Massachusetts Roads Dataset, LRSNY Roads Segmentation Dataset, and CHN6-CUG Roads Dataset. We trained on these four datasets for about 14 h, 36 h, 6 h and 23 h respectively, and spent 0.3 s/batch on the test dataset.

Table 1
The proposed DPENet was compared with 5 other methods on the Massachusetts Building Dataset.

Method	OA(%)	Pre(%)	Recall(%)	IoU(%)	F1(%)
Residual U-Net	92.54	81.69	75.74	64.75	78.60
DANet	92.60	82.76	74.67	64.62	78.51
PSPNet101	92.83	82.35	76.86	65.99	79.51
SAB U-Net	94.18	84.35	83.29	72.14	83.82
CBRNet	94.67	86.59	84.61	74.81	85.59
BuildFormer	/	**87.52**	84.90	75.74	86.19
DPENet	**95.06**	86.84	**86.77**	**76.69**	**86.81**

On the Massachusetts Building Dataset, we compared five methods that have shown remarkable performance in building object extraction: Residual U-Net(Zhang et al., 2017), DANet(Fu et al., 2018), PSPNet101 (Zhao et al., 2016), SAB U-Net(Chen et al., 2021b), CBRNet(Guo et al., 2022), and BuildFormer(Wang et al., 2022). inspired by ResNet, combines this concept with U-Net to propose a novel approach for building object extraction. DANet enhances the model's understanding of global semantic relationships by incorporating Position Attention Module (PAM) and Channel Attention Module (CAM) into an expanded FCN. PSPNet101 aggregates contextual information from different regions, enabling the model to have a stronger global contextual understanding. SAB U-Net achieves high-precision building object extraction through self-attention mechanisms and large receptive fields. CBRNet refines building boundaries from coarse to fine, enhancing the model's edge-awareness capability.

On the Massachusetts Roads Dataset, LRSNY Roads Segmentation Dataset, and CHN6-CUG Roads Dataset, we compared several methods that have demonstrated outstanding performance in road object extraction, including U-Net(Ronneberger et al., 2015), SegNet(Badrinarayanan et al., 2017), Residual U-Net(Zhang et al., 2017), DANet(Fu et al., 2018), PSPNet50(Zhao et al., 2016), PSPNet101(Zhao et al., 2016), DeepLabV3(Chen et al., 2017), DeepLabV3Plus(Chen et al., 2018), Bias U-Net(Chen et al., 2021d), GCBNet(Zhu et al., 2021), Swin-Unet(Cao et al., 2022) and D-LinkNet(Zhou et al., Jun 2018). Among them, DeepLabV3 enhances the ability to capture multi-scale information of the objects by introducing convolutions with different dilation rates into ASPP. DeepLabV3Plus treats the DCNN part of DeepLabV3 as the encoder and utilizes upsampling of output feature maps as the decoder, strengthening the edge regions of semantic segmentation. Bias U-Net improves the reconstruction capability of the objects through upsampling operations with multiple convolutional kernels. GCBNet combines global contextual awareness modules with batch-independent mechanisms to enhance the integrity and continuity of road object extraction results. In addition, the bold font in each table in the experimental section indicates the best score achieved on that evaluation metric.

According to Table 1, our method achieved the highest OA, Recall,

Table 2
The proposed DPENet was compared with 7 other methods on the Massachusetts Roads Dataset.

Method	OA(%)	Pre(%)	Recall(%)	IoU(%)	F1(%)
SegNet	97.91	81.15	72.89	62.34	76.80
PSPNet50	97.80	82.46	68.06	59.45	74.57
Residual U-Net	97.89	79.81	74.16	62.45	76.88
DeepLabV3	97.74	81.81	67.37	58.59	73.89
DANet	97.87	**85.26**	66.58	59.70	74.77
D-LinkNet	97.99	81.93	73.92	63.56	77.72
Swin-Unet	97.88	77.57	71.5	63.42	77.62
Bias U-Net	/	79.14	78.53	65.06	78.83
DPENet	**98.09**	80.17	**79.46**	**66.41**	**79.81**

Table 3
The proposed DPENet was compared with 7 other methods on the LRSNY Roads Segmentation Dataset.

Method	OA(%)	Pre(%)	Recall(%)	IoU(%)	F1(%)
SegNet	97.54	92.80	91.29	85.25	92.04
PSPNet101	97.84	93.40	92.64	86.94	93.01
Residual U-Net	97.53	92.72	91.27	85.16	91.99
DeepLabV3	97.51	93.23	90.59	85.00	91.89
DeepLabV3plus	97.80	93.09	92.75	86.78	92.92
DANet	97.66	94.42	90.32	85.74	92.32
Swin-Unet	97.22	90.96	91.19	83.61	91.07
Bias U-Net	97.83	93.42	92.57	86.90	92.99
DPENet	**98.16**	**94.64**	**93.67**	**88.96**	**94.16**

Table 4
The proposed DPENet was compared with 7 other methods on the CHN6-CUG Roads Dataset.

Method	OA(%)	Pre(%)	Recall(%)	IoU(%)	F1(%)
U-Net	96.36	74.61	55.34	46.58	63.54
SegNet	96.69	76.24	61.46	51.58	68.06
Residual U-Net	96.39	71.39	61.78	49.52	66.24
DeepLabV3	96.92	77.63	64.93	54.70	70.72
DeepLabV3plus	96.86	74.86	68.27	55.54	71.41
D-LinkNet	96.91	**79.06**	62.72	53.79	69.95
Swin-Unet	96.85	73.46	70.48	56.18	71.94
GCBNet	/	/	/	60.44	72.70
DPENet	**97.37**	74.79	**78.37**	**61.99**	**76.54**

IoU, and F1 scores on the Massachusetts Building Dataset. Compared to CBRNet, our DPENet shows comparable performance in terms of OA and Precision scores, but it outperforms in Recall, IoU, and F1 scores by 2.16%, 1.88%, and 1.22%, respectively. Although our DPENet is lower than the BuildFormer method in Pre score, we are higher than it in other metrics. This improvement can be attributed to the stronger object recognition and extraction capabilities of our dual-path network in building extraction.

According to Table 2, Table 3, and Table 4, our proposed method achieved the highest IoU and F1-Score on the Massachusetts Roads Dataset, LRSNY Roads Segmentation Dataset, and CHN6-CUG Roads Dataset, demonstrating the outstanding performance of our DPENet in road object extraction. On the Massachusetts Roads Dataset, DANet achieved the highest Precision score of 85.25%, but its performance in Recall score was not satisfactory, indicating the difficulty of distinguishing road objects from the environment. On the LRSNY Roads Segmentation Dataset, our method showed the best performance across all five evaluation metrics, with improvements of 0.33%, 1.22%, 1.10%, 2.06%, and 1.17% over Bias U-Net, highlighting the effectiveness of our approach in the reconstruction part. According to Table 3, our proposed method achieved the highest OA, Recall, IoU, and F1 scores on the CHN6-CUG Roads Dataset. D-LinkNet performed the best in terms of Precision evaluation, but it did not excel in other evaluation metrics. Additionally, our method showed improvements of 1.55% and 3.84% in the two most important metrics, IoU and F1 scores, compared to GCBNet. We also compare with the novel Swin-Unet network, but the performance on the three road datasets is not very optimistic. We also compare with the novel Swin-Unet network, but the performance on the three road datasets is not very optimistic.

4.5. Visualization analysis

To further compare and analyze the advantages and limitations of our proposed method, in this section, we will present the extraction results of DPENet and other comparative methods on four building and road datasets. These datasets include the Massachusetts Building Dataset, Massachusetts Roads Dataset, LRSNY Roads Segmentation Dataset, and CHN6-CUG Roads Dataset.

As shown in the red boxes in Fig. 6, our method demonstrates excellent capture and extraction of smaller building objects on the Massachusetts Building Dataset, with extracted results exhibiting regular shapes close to the Ground Truth. The comparison of extraction results in the second row also indicates the high integrity of our method in extracting building objects. On road remote sensing image datasets such as the Massachusetts Roads Dataset, LRSNY Roads Segmentation Dataset, and CHN6-CUG Roads Dataset, our method also showcases outstanding object extraction capabilities. When comparing our method to Swin-Unet, it becomes evident that the road objects extracted by our approach exhibit a notably higher degree of regularity along their edges. In Fig. 7, the extraction results in the first row reveal that other methods struggle to capture small forks within the red boxes, and both the DANet and D-LinkNet methods exhibit incomplete object extraction. However, our method successfully extracts such fork-like objects comprehensively. The object extraction results of several comparative methods presented in Fig. 8 are satisfactory, as they capture road objects effectively. The only potential drawback is the difficulty in distinguishing subtle gap areas between multiple roads. Nonetheless, compared to the DeepLabV3Plus method, which performs well in gap distinction, our method still maintains an advantage. In Fig. 9, the red box in the first row highlights the triangular flower bed in the middle of the intersection, which is a challenging area to differentiate in road object extraction and can be misleading. However, compared to other comparative methods, our method is capable of separating the triangular flower bed

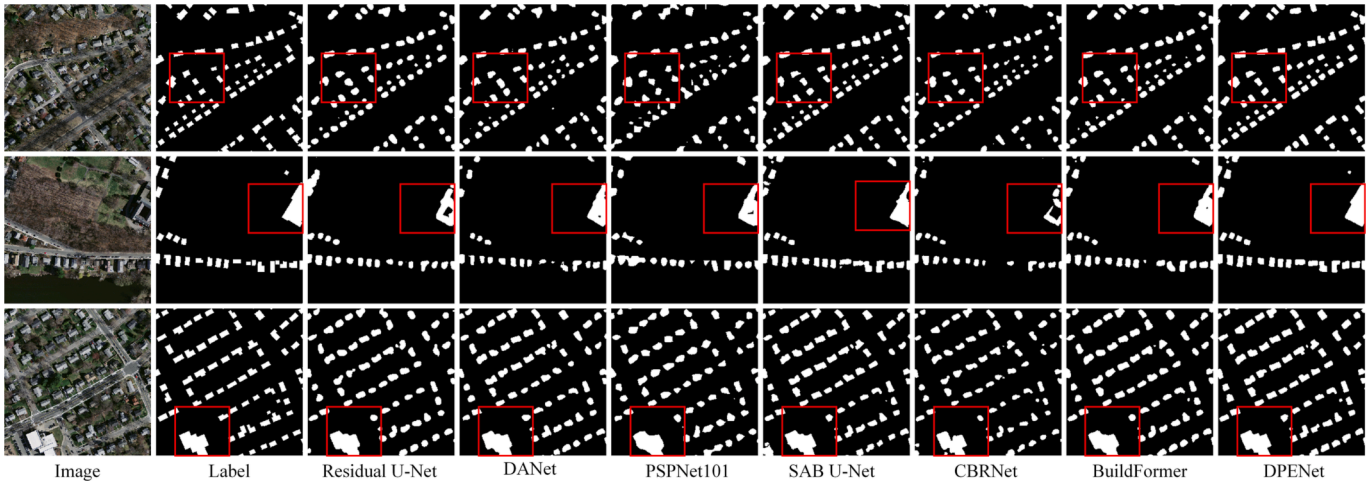


Fig. 6. An example of the extraction results of DPENet and several other methods on the Massachusetts Building Dataset.

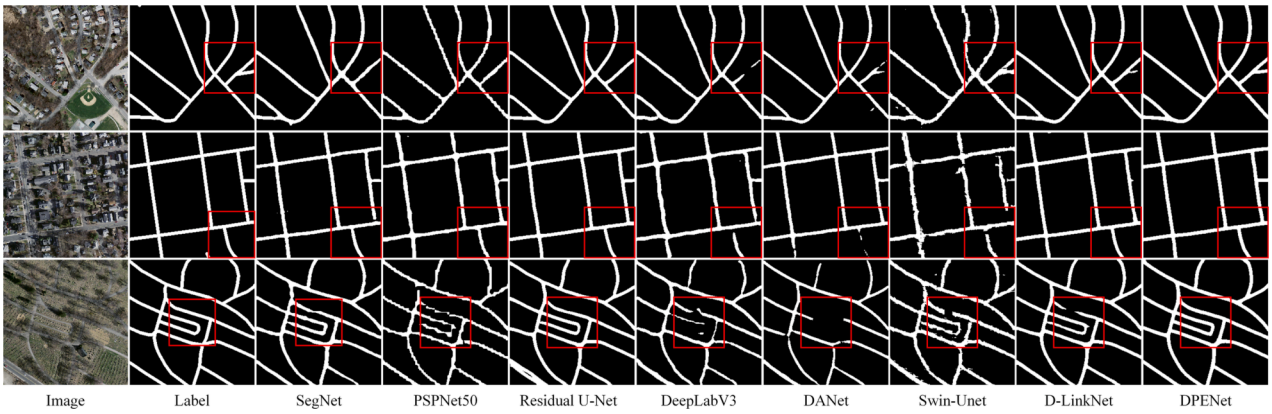


Fig. 7. An example of the extraction results of DPENet and several other methods on the Massachusetts Roads Dataset.

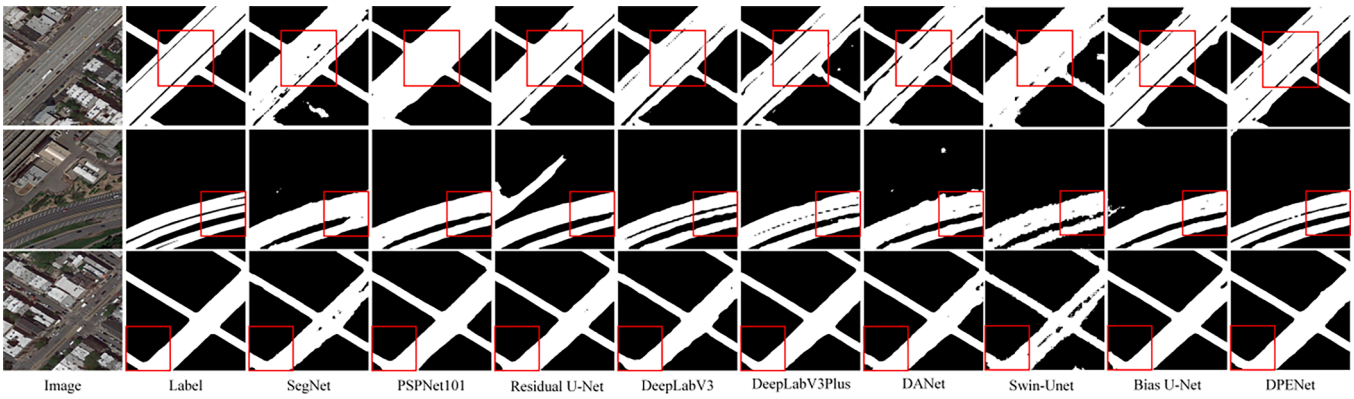


Fig. 8. An example of the extraction results of DPENet and several other methods on the LRSNY Roads Segmentation Dataset.

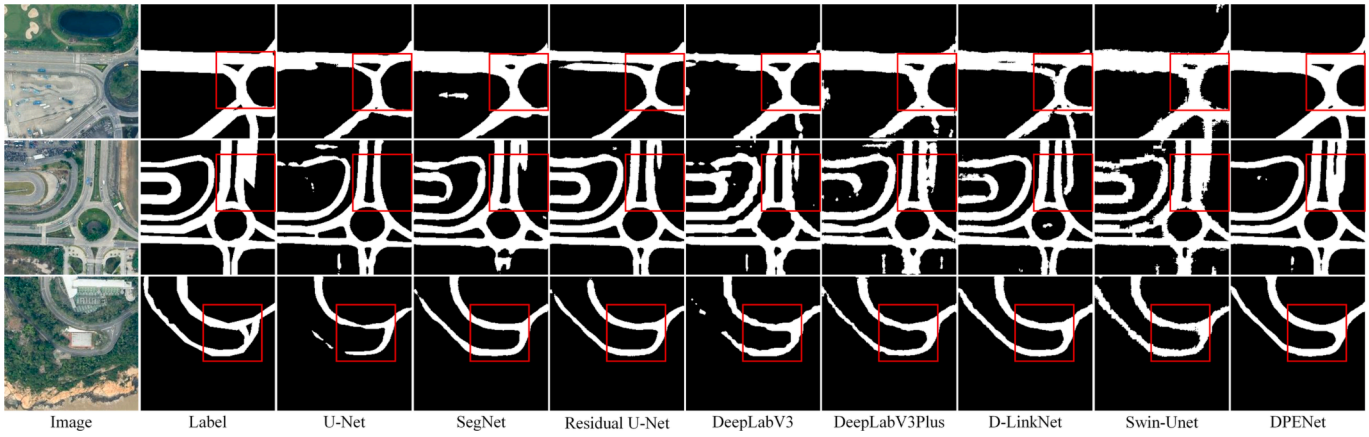


Fig. 9. An example of the extraction results of DPENet and several other methods on the CHN6-CUG Roads Dataset.

more completely. The extraction results within the red box in the second row of Fig. 9 demonstrate that our method also excels in complex road environments. However, it may slightly struggle in extracting auxiliary lanes and differentiating open spaces with colors similar to the road, leading to errors in detecting small regions. In addition, compared with Swin-Unet, our method excels in extracting road objects with a higher degree of regularity along their edges. This enhanced regularity results in smoother and more precisely defined road boundaries, leading to improved accuracy and consistency in road object segmentation. Our approach carefully preserves the intricate details and fine-grained features of road edges, ensuring that the extracted road objects closely adhere to their actual shapes and contours. This level of precision is particularly valuable for applications such as autonomous driving and urban planning, where accurately delineated road boundaries are critical for safe and effective navigation and analysis.

4.6. Ablation experiment analysis

In this section, we primarily analyze the effectiveness of the components of the proposed DPENet model in extracting building and road objects. We conducted ablation experiments and evaluated them on four datasets. Through quantitative and qualitative analyses, we explored the

Table 5
Experimental analysis of ablation on the Massachusetts Building Dataset.

Method	OA(%)	Pre(%)	Recall(%)	IoU(%)	F1(%)
Baseline	94.77	86.13	85.89	75.45	86.01
Baseline + M	94.91	**86.97**	85.63	75.90	86.30
Baseline + M + G	94.93	86.50	86.38	76.12	86.44
Baseline + M + G + D	**95.06**	86.84	**86.77**	**76.69**	**86.81**

Table 6
Experimental analysis of ablation on the Massachusetts Roads Dataset.

Method	OA(%)	Pre(%)	Recall(%)	IoU(%)	F1(%)
Baseline	98.05	79.38	79.59	65.95	79.48
Baseline + M	98.06	79.15	**80.07**	66.13	79.61
Baseline + M + G	98.07	79.55	79.94	66.32	79.75
Baseline + M + G + D	**98.09**	**80.17**	79.46	**66.41**	**79.81**

impact of different sub-components on the model's performance in order to gain a better understanding of their roles. M (Multi-residual combination mechanism), G (Global information extraction branch), D (Deep monitoring strategy).

According to Table 5, we introduced a residual mechanism in each convolutional operation of the U-Net network and used it as our baseline model. We performed supervised training on images and labels using BCE loss and Dice loss, and introduced multiple losses during the training process. As shown in Table 6, these improved models demonstrated significant performance improvements. On the Massachusetts Building and Roads Dataset, the IoU and F1 scores increased to 75.45% and 86.01%, respectively. After introducing the multiple residual mechanism, the IoU score on the Massachusetts Building Dataset increased by 0.45%, and the Recall on the Massachusetts Road Dataset increased by 0.48%. Subsequently, we further introduced the global branch, which slightly improved the IoU scores by 0.22% and 0.19% respectively. Finally, by introducing the deep supervision strategy, the IoU score on the Massachusetts Building Dataset significantly improved from 76.12% to 76.69%. On the Massachusetts Road Dataset, the Pre score increased from 79.55 to 80.17. This is because the deep supervision strategy effectively supervises the lower-resolution and smaller-scale Massachusetts road and building datasets. As shown in Fig. 10, in the extraction results of Massachusetts roads and buildings, the completeness of the objects gradually improved, especially within the closed-loop regions of the building and road objects. Our proposed comprehensive network effectively controls the boundaries of the extraction results and obtains more refined results.

We performed ablation experiments and analysis on the LRSNY Roads Segmentation Dataset and CHN6-CUG Roads Dataset to provide additional evidence of the effectiveness of our proposed method. As demonstrated in Table 6 , Table 7 and Table 8, our comprehensive network (Baseline + M + G + D) achieved the highest IoU and F1 scores on both datasets, validating the effectiveness of our method. After introducing the multiple sampling module on the LRSNY Roads Segmentation Dataset, the IoU and F1 scores increased from 88.09% and

Table 7
Experimental analysis of ablation on LRSNY Roads Segmentation Dataset.

Method	OA(%)	Pre(%)	Recall(%)	IoU(%)	F1(%)
Baseline	98.04	93.90	93.44	88.09	93.67
Baseline + M	98.14	94.29	**93.72**	88.69	94.00
Baseline + M + G	**98.19**	94.60	93.70	88.94	94.14
Baseline + M + G + D	98.16	**94.64**	93.67	**88.96**	**94.16**

Table 8
Experimental analysis of ablation on CHN6-CUG Roads Dataset.

Method	OA(%)	Pre(%)	Recall(%)	IoU(%)	F1(%)
Baseline	97.24	77.43	73.05	60.23	75.18
Baseline + M	97.29	77.76	73.79	60.93	75.72
Baseline + M + G	97.33	77.81	74.62	61.52	76.18
Baseline + M + G + D	**97.37**	**74.79**	**78.37**	**61.99**	**76.54**

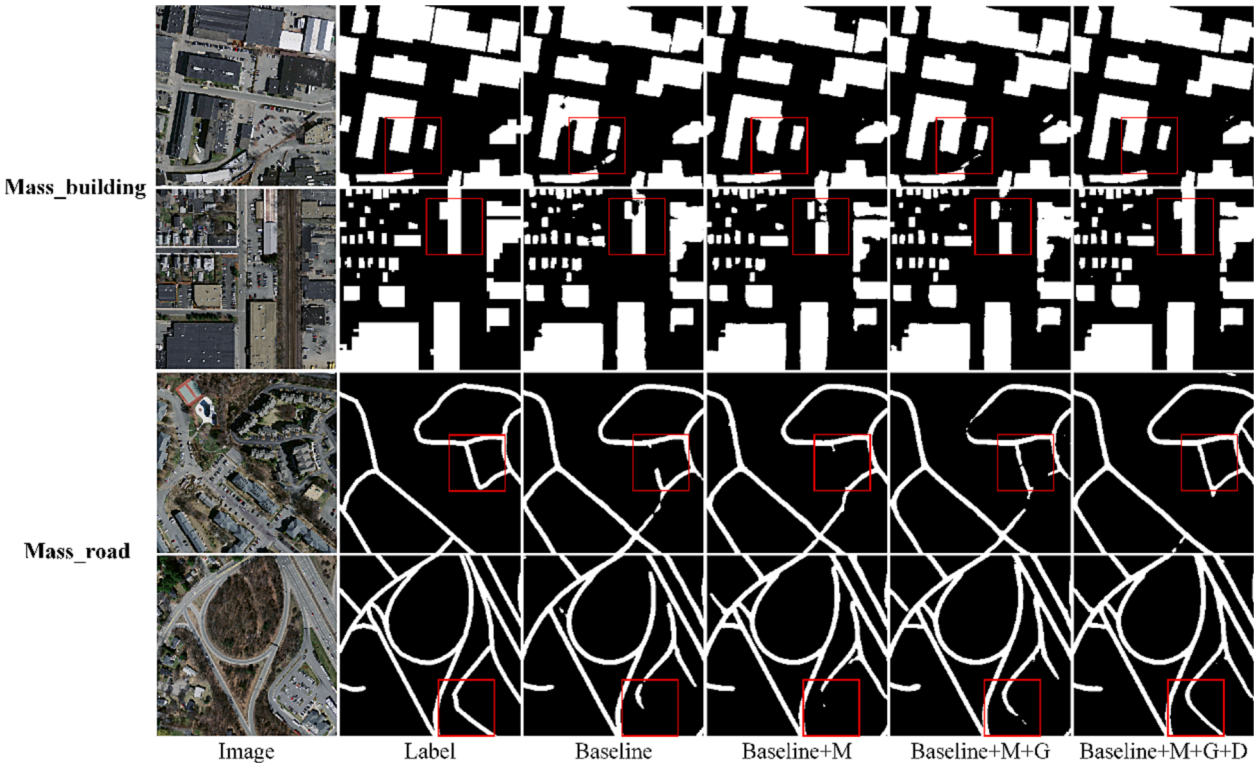


Fig. 10. An example of the results obtained from the ablation experiments conducted on the Massachusetts Building and Roads Dataset.

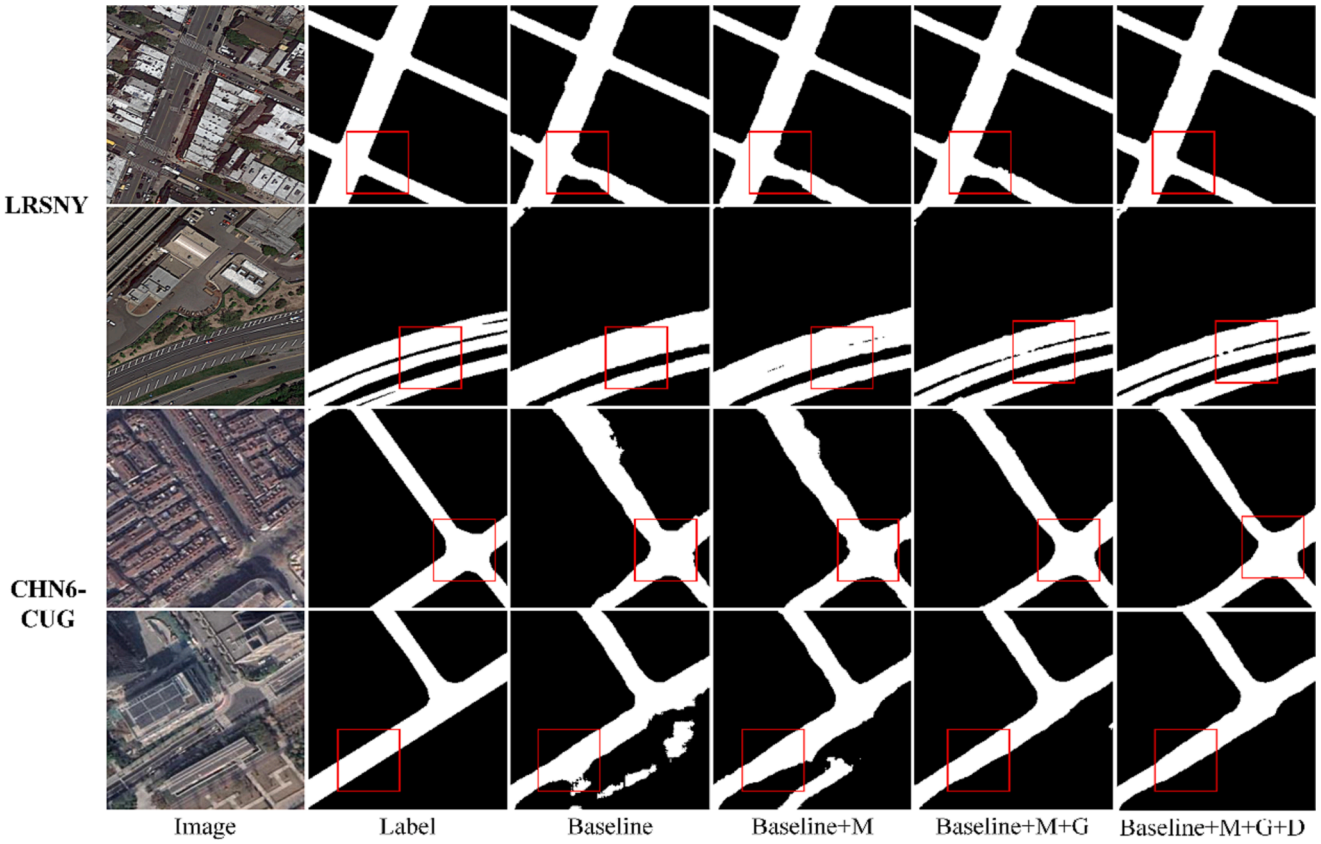


Fig. 11. An example of the results obtained from the ablation experiments conducted on the LRSNY and CHN6-CUG Roads Dataset.

93.67% to 88.69% and 94.00%, respectively. The introduction of the global branch resulted in a 0.25 improvement in the IoU score. The introduction of the deep supervision mechanism on the LRSNY dataset showed only slight improvements. By observing the extraction results in Fig. 11, we can observe a significant improvement in the comprehensive network's ability to control the spacing areas between roads.

After introducing the multiple sampling mechanism to the Baseline, the IoU and F1 scores on the CHN6-CUG dataset increased by 0.7% and 0.54% respectively. As shown in Fig. 11, the results with multiple sampling are more refined. By adding the global branch and deep supervision, significant improvements in the IoU and F1 scores of the extraction results can be observed in Table 8. Fig. 11 also demonstrates noticeable enhancements in object recognition and shape control in the extraction results, further confirming the effectiveness of our proposed multiple sampling, global branch, and deep supervision.

5. Conclusion

This paper presents a dual-path extraction network based on CNN and Transformer for extracting building and road objects from optical remote sensing images. To enhance the utilization of image features, we construct a CNN-based local spatial information branch and a Transformer-based global information branch from the perspectives of global dependencies and local spatial information. This effectively integrates global and local information to enhance the network's extraction capability. In addition to the dual-path extraction structure, we propose a multi-view multi-sampling mechanism for the reconstruction part of the network, increasing the complexity of the reconstruction and promoting the reconstruction of image features. To improve the accurate reconstruction of objects at different resolutions, we employ a deep supervision strategy combined with a multi-level upsampling mechanism.

During the testing phase, we utilize four publicly available building and road datasets: Massachusetts Building Dataset, Massachusetts Roads Dataset, LRSNY Roads Segmentation Dataset, and CHN6-CUG Roads Dataset. Additionally, we compare our network with several advanced extraction methods through quantitative and qualitative comparisons. The results on all four datasets demonstrate highly satisfactory performance. On the Massachusetts Building and Roads Dataset, we achieve an IoU of 76.69% and 66.41% and an F1-score of 86.81% and 79.81% respectively. On the LRSNY and CHN6-CUG Roads Dataset, we achieve an IoU of 88.96% and 61.99% and an F1-score of 94.16% and 76.54% respectively. To further validate the effectiveness of the proposed components, we conduct ablation experiments and perform quantitative and qualitative analyses of the results. The experimental results show significant improvements brought by each component on the building and road datasets. The qualitative analysis of the extraction result examples further confirms the constraining roles of different components in the recognition and extraction of building and road objects.

CRediT authorship contribution statement

Ziyi Chen: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Project administration, Funding acquisition. **Yuhua Luo:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft. **Jing Wang:** Resources, Writing – review & editing. **Jonathan Li:** Investigation, Formal analysis. **Cheng Wang:** Conceptualization, Supervision. **Dilong Li:** Formal analysis, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial

interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This study was financially supported by the Natural Science Foundation of Fujian Province (No.2023J01135), National Natural Science Foundation of China (No.62001175), Fundamental Research Funds for the Central Universities of Huaqiao University (No.ZQN-911), the National Natural Science Foundation of China (No.42201475, 61972168), the Natural Science Foundation of Fujian Province (NO.2022J01317, 2021J05059), the Fundamental Research Funds for the Central Universities of Huaqiao University (ZQN-1114), and in part by the Major Science and Technology Project of Xiamen (Industry and Information Technology Area) (NO.3502Z20231007).

References

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2022. Swin-Unet: Unet-like pure transformer for medical image segmentation. In: Proc. Eur. Conf. Comput. Vis. Springer, pp. 205-218. https://doi.org/10.1007/978-3-031-25066-8_9.

Chen, Z., Li, D., Fan, W., Guan, H., Wang, C., Li, J., 2021b. Self-attention in reconstruction bias U-Net for semantic segmentation of building rooftops in optical remote sensing images. Remote Sens. 13 (13), 2524. https://doi.org/10.3390/rs13132524.

Chen, Z., Deng, L., Luo, Y., Li, D., Junior, J.M., Gonçalves, W.N., Nurunnabi, A.A.M., Li, J., Wang, C., Li, D., 2022c. Road extraction in remote sensing data: A survey. Int. J. Appl. Earth Obs. Geoinf. 112, 102833 https://doi.org/10.1016/j.jag.2022.102833.

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. ArXiv abs/1706.05587.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: Proc. Eur. Conf. Comput. Vis. https://doi.org/10.1007/978-3-030-01234-2_49.

Chen, R., Li, X., Hu, Y., Wen, C., Peng, L., 2022a. Road Extraction From Remote Sensing Images in Wildland-Urban Interface Areas. IEEE Geosci. Remote Sens. Lett. 19, 3000705. https://doi.org/10.1109/lgrs.2020.3028468.

Chen, H., Li, Z., Wu, J., Xiong, W., Du, C., 2023. SemiRoadExNet: A semi-supervised network for road extraction from remote sensing imagery via adversarial learning. ISPRS J. Photogramm. Remote Sens. 198, 169–183. https://doi.org/10.1016/j.isprsjprs.2023.03.012.

Chen, X., Sun, Q., Guo, W., Qiu, C., Yu, A., 2022b. GA-Net: A geometry prior assisted neural network for road extraction. Int. J. Appl. Earth Obs. Geoinf. 114, 103004 https://doi.org/10.1016/j.jag.2022.103004.

Chen, Z., Wang, C., Li, J., Fan, W., Du, J., Zhong, B., 2021c. Adaboost-like End-to-End multiple lightweight U-nets for road extraction from optical remote sensing images. Int. J. Appl. Earth Obs. Geoinf. 100, 102341 https://doi.org/10.1016/j.jag.2021.102341.

Chen, Z., Wang, C., Li, J., Xie, N., Han, Y., Du, J., 2021d. Reconstruction bias U-Net for road extraction from optical remote sensing images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 2284–2294. https://doi.org/10.1109/JSTARS.2021.3053603.

Chen, K., Zou, Z., Shi, Z., 2021a. Building Extraction from Remote Sensing Images with Sparse Token Transformers. Remote Sens. 13, 4441. https://doi.org/10.3390/rs13214441.

Dai, J., Ma, R., Gong, L., Shen, Z., Wu, J., 2021. A Model-Driven-to-Sample-Driven Method for Rural Road Extraction. Remote Sens. 13, 1417. https://doi.org/10.3390/RS13081417.

Deng, W., Shi, Q., Li, J., 2021. Attention-Gate-Based Encoder–Decoder Network for Automatical Building Extraction. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 2611–2620. https://doi.org/10.1109/JSTARS.2021.3058097.

Ding, L., Tang, H., Liu, Y., Shi, Y., Zhu, X.X., Bruzzone, L., 2021. Adversarial shape learning for building extraction in VHR remote sensing images. IEEE Trans. Image Process. 31, 678–690. https://doi.org/10.1109/TIP.2021.3134455.

Ding, L., Tang, H., Liu, Y., Shi, Y., Zhu, X., Bruzzone, L., 2022. Adversarial Shape Learning for Building Extraction in VHR Remote Sensing Images. IEEE Trans. Image Process. 31, 678–690. https://doi.org/10.1109/TIP.2021.3134455.

Feng, D., Chu, H., Zheng, L., 2022. Frequency Spectrum Intensity Attention Network for Building Detection from High-Resolution Imagery. Remote Sens. 14 (21), 5457. https://doi.org/10.3390/rs14215457.

Fu, J., Liu, J., Tian, H., Fang, Z., Lu, H., 2018. Dual Attention Network for Scene Segmentation. In Proc. IEEE Conf. Comput. Vis. Patt. Recog., 3141-3149. https://doi.org/10.1109/CVPR.2019.00326.

Guan, H., Yu, Y., Li, D., Wang, H., 2021. RoadCapsFPN: Capsule Feature Pyramid Network for Road Extraction From VHR Optical Remote Sensing Imagery. IEEE Trans. Intell. Transp. Syst. 23 (8), 11041–11051. https://doi.org/10.1109/tits.2021.3098855.

Guan, H., Lei, X., Yu, Y., Zhao, H., Peng, D., Junior, J.M., Li, J., 2022. Road marking extraction in UAV imagery using attentive capsule feature pyramid network. Int. J. Appl. Earth Obs. Geoinf. 107, 102677 https://doi.org/10.1016/j.jag.2022.102677.

Guo, H., Du, B., Zhang, L., Su, X., 2022. A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 183, 240–252. https://doi.org/10.1016/j.isprsjprs.2021.11.005.

Guo, M., Liu, H., Xu, Y., Huang, Y., 2020. Building Extraction Based on U-Net with an Attention Block and Multiple Losses. Remote Sens. 12, 1400. https://doi.org/10.3390/rs12091400.

Hinton, G.E., Mnih, V., 2013. Machine Learning for Aerial Image Labeling.

Hu, A., Chen, S., Wu, L., Xie, Z., Qiu, Q., Xu, Y., 2021. WSGAN: An Improved Generative Adversarial Network for Remote Sensing Image Road Network Extraction by Weakly Supervised Processing. Remote Sens. 13, 2506. https://doi.org/10.3390/rs13132506.

Jiang, X., Li, Y., Jiang, T., Xie, J., Wu, Y., Cai, Q., Jiang, J., Xu, J., Zhang, H., 2022. RoadFormer: Pyramidal deformable vision transformers for road network extraction with remote sensing images. Int. J. Appl. Earth Obs. Geoinf. 113, 102987 https://doi.org/10.1016/j.jag.2022.102987.

Li, W., Zhao, W., Zhong, H., He, C., Lin, D., 2021c. Joint semantic-geometric learning for polygonal building segmentation. In: Proc. AAAI Conf. Artif. Intell. 35(3), 1958-1965. https://doi.org/10.1609/aaai.v35i3.16291.

Li, C., Fu, L., Zhu, Q., Zhu, J., Fang, Z., Xie, Y., Guo, Y., Gong, Y., 2021a. Attention enhanced u-net for building extraction from farmland based on google and worldview-2 remote sensing images. Remote Sens. 13 (21), 4411. https://doi.org/10.3390/rs13214411.

Li, P., He, X., Qiao, M., Miao, D., Cheng, X., Song, D., Chen, M., Li, J., Zhou, T., Guo, X., Yan, X., Tian, Z., 2021b. Exploring multiple crowdsourced data to learn deep convolutional neural networks for road extraction. Int. J. Appl. Earth Obs. Geoinf. 104, 102544 https://doi.org/10.1016/j.jag.2021.102544.

Lian, R., Huang, L., 2020. DeepWindow: Sliding Window Based on Deep Learning for Road Extraction From Remote Sensing Images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 1905–1916. https://doi.org/10.1109/JSTARS.2020.2983788.

Liu, X., Chen, Y., Wang, C., Tan, K., Li, J., 2023a. A lightweight building instance extraction method based on adaptive optimization of mask contour. Int. J. Appl. Earth Obs. Geoinf. 122, 103420 https://doi.org/10.1016/j.jag.2023.103420.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin Transformer: Hierarchical vision transformer using shifted windows. In Proc. IEEE Int. Conf. Comput. Vis, 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986.

Liu, P., Liu, X., Liu, M., Shi, Q., Yang, J., Xu, X., Zhang, Y., 2019. Building Footprint Extraction from High-Resolution Images via Spatial Residual Inception Convolutional Neural Network. Remote Sens. 11 (7), 830. https://doi.org/10.3390/rs11070830.

Liu, Z., Shi, Q., Ou, J., 2022. LCS: A Collaborative Optimization Framework of Vector Extraction and Semantic Segmentation for Building Extraction. IEEE Trans. Geosci. Remote Sens. 60, 1–15. https://doi.org/10.1109/TGRS.2022.3215852.

Liu, X., Wang, Z., Wan, J., Zhang, J., Xi, Y., Liu, R., Miao, Q., 2023b. RoadFormer: Road Extraction Using a Swin Transformer Combined with a Spatial and Channel Separable Convolution. Remote Sens. 15 (4), 1049. https://doi.org/10.3390/rs15041049.

Lu, X., Zhong, Y., Zheng, Z., Chen, D., Su, Y., Ma, A., Zhang, L., 2022. Cascaded Multi-Task Road Extraction Network for Road Surface, Centerline, and Edge Extraction. IEEE Trans. Geosci. Remote Sens. 60, 1–14. https://doi.org/10.1109/tgrs.2022.3165817.

Luo, L., Wang, J.-X., Chen, S.-B., Tang, J., Luo, B., 2022. BDTNet: Road Extraction by Bi-Direction Transformer From Remote Sensing Images. IEEE Geosci. Remote Sens. Lett. 19, 1–5. https://doi.org/10.1109/lgrs.2022.3183828.

Mao, Z., Huang, X., Niu, W., Wang, X., Hou, Z., Zhang, F., 2023. Improved instance segmentation for slender urban road facility extraction using oblique aerial images. Int. J. Appl. Earth Obs. Geoinf. 121, 103362 https://doi.org/10.1016/j.jag.2023.103362.

Peng, D., Guan, H., Zang, Y., Bruzzone, L., 2021. Full-level domain adaptation for building extraction in very-high-resolution optical remote-sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–17. https://doi.org/10.1109/TGRS.2021.3093004.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. ArXiv. https://doi.org/10.48550/arXiv.1505.04597.

Shamsolmoali, P., Zareapoor, M., Zhou, H., Wang, R., Yang, J., 2021. Road Segmentation for Remote Sensing Images Using Adversarial Spatial Pyramid Networks. IEEE Trans. Geosci. Remote Sens. 59, 4673–4688. https://doi.org/10.1109/TGRS.2020.3016086.

Shao, Z., Tang, P., Wang, Z., Saleem, N., Yam, S., Sommai, C., 2020. BRRNet: A Fully Convolutional Neural Network for Automatic Building Extraction From High-Resolution Remote Sensing Images. Remote Sens. 12, 1050. https://doi.org/10.3390/rs12061050.

Tan, Y., Gao, S., Li, X.-y., Cheng, M.-M., Ren, B., 2020. VecRoad: Point-Based Iterative Graph Exploration for Road Graphs Extraction. In: Proc. IEEE Conf. Comput. Vis. Patt. Recog., Seattle, WA, USA, pp. 8907-8915. https://doi.org/10.1109/cvpr42600.2020.00893.

Tan, X., Xiao, Z., Wan, Q., Shao, W., 2021. Scale Sensitive Neural Network for Road Segmentation in High-Resolution Remote Sensing Images. IEEE Geosci. Remote Sens. Lett. 18, 533–537. https://doi.org/10.1109/LGRS.2020.2976551.

Tao, J., Chen, Z., Sun, Z., Guo, H., Leng, B., Yu, Z., Wang, Y., He, Z., Lei, X., Yang, J., 2023. Seg-Road: A Segmentation Network for Road Extraction Based on Transformer and CNN with Connectivity Structures. Remote Sens. 15 (6), 1602. https://doi.org/10.3390/rs15061602.

Waldner, F., Diakogiannis, F.I., 2020. Deep learning on edge: Extracting field boundaries from satellite images with a convolutional neural network. Remote Sens. Environ. 245, 111741 https://doi.org/10.1016/j.rse.2020.111741.

Wang, L., Fang, S., Meng, X., Li, R., 2022. Building extraction with vision transformer. IEEE Trans. Geosci. Remote Sens. 60, 1–11. https://doi.org/10.1109/TGRS.2022.3186634.

Wei, S., Ji, S., 2021. Graph convolutional networks for the automated production of building vector maps from aerial images. IEEE Trans. Geosci. Remote Sens. 60, 1–11. https://doi.org/10.1109/TGRS.2021.3060770.

Wei, Y., Zhang, K., Ji, S., 2020. Simultaneous Road Surface and Centerline Extraction From Large-Scale Remote Sensing Images Using CNN-Based Segmentation and Tracing. IEEE Trans. Geosci. Remote Sens. 58, 8919–8931. https://doi.org/10.1109/TGRS.2020.2991733.

Xu, L., Li, Y., Xu, J., Zhang, Y., Guo, L., 2023. BCTNet: Bi-Branch Cross-Fusion Transformer for Building Footprint Extraction. IEEE Trans. Geosci. Remote Sens. 61, 1–14. https://doi.org/10.1109/TGRS.2023.3262967.

Xu, Y., Wu, L., Xie, Z., Chen, Z., 2018. Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. Remote Sens. 10 (1), 144. https://doi.org/10.3390/rs10010144.

Xu, Y., Chen, H., Du, C., Li, J., 2021. MSACon: Mining spatial attention-based contextual information for road extraction. IEEE Trans. Geosci. Remote Sens. 60, 1–17. https://doi.org/10.1109/TGRS.2021.3073923.

Xu, H., Zhu, P., Luo, X., Xie, T., Zhang, L., 2022. Extracting Buildings from Remote Sensing Images Using a Multitask Encoder-Decoder Network with Boundary Refinement. Remote Sens. 14, 564. https://doi.org/10.3390/rs14030564.

Yan, J., Ji, S., Wei, Y., 2022. A Combination of Convolutional and Graph Neural Networks for Regularized Road Surface Extraction. IEEE Trans. Geosci. Remote Sens. 60, 1–13. https://doi.org/10.1109/TGRS.2022.3151688.

Yang, Z., Zhou, D., Yang, Y., Zhang, J., Chen, Z., 2022. TransRoadNet: A novel road extraction method for remote sensing images via combining high-level semantic feature and context. IEEE Geosci. Remote Sens. Lett. 19, 1–5. https://doi.org/10.1109/lgrs.2022.3171973.

Zhang, X., Ma, W., Li, C., Wu, J., Tang, X., Jiao, L., 2019. Fully Convolutional Network-Based Ensemble Method for Road Extraction From Aerial Images. IEEE Geosci. Remote Sens. Lett. 17 (10), 1777–1781. https://doi.org/10.1109/LGRS.2019.2953523.

Zhang, X., Jiang, Y., Wang, L., Han, W., Feng, R., Fan, R., Wang, S., 2022. Complex Mountain Road Extraction in High-Resolution Remote Sensing Images via a Light Roadformer and a New Benchmark. Remote Sens. 14 (19), 4729. https://doi.org/10.3390/rs14194729.

Zhang, Z., Liu, Q., Wang, Y., 2017. Road Extraction by Deep Residual U-Net. IEEE Geosci. Remote Sens. Lett. 15, 749–753. https://doi.org/10.1109/LGRS.2018.2802944.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2016. Pyramid Scene Parsing Network. In: Proc. IEEE Conf. Comput. Vis. Patt. Recog., 6230-6239. https://doi.org/10.1109/CVPR.2017.660.

Zhou, G., Chen, W., Gui, Q., Li, X., Wang, L., 2021. Split depth-wise separable graph-convolution network for road extraction in complex environments from high-resolution remote-sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–15. https://doi.org/10.1109/TGRS.2021.3128033.

Zhou, M., Sui, H., Chen, S., Wang, J., Chen, X., 2020. BT-RoadNet: A boundary and topologically-aware neural network for road extraction from high-resolution remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 168, 288–306. https://doi.org/10.1016/j.isprsjprs.2020.08.019.

Zhou, M., Sui, H., Chen, S., Liu, J., Shi, W., Chen, X., 2022. Large-scale road extraction from high-resolution remote sensing images based on a weakly-supervised structural and orientational consistency constraint network. ISPRS J. Photogramm. Remote Sens. 193, 234–251. https://doi.org/10.1016/j.isprsjprs.2022.09.005.

Zhu, Q., Liao, C., Hu, H., Mei, X., Li, H., 2020. MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery. IEEE Trans. Geosci. Remote Sens. 59 (7), 6169–6181. https://doi.org/10.1109/TGRS.2020.3026051.

Zhu, Q., Zhang, Y., Wang, L., Zhong, Y., Guan, Q., Lu, X., Zhang, L., Li, D., 2021. A Global Context-aware and Batch-independent Network for road extraction from VHR satellite imagery. ISPRS J. Photogramm. Remote Sens. 175, 353–365. https://doi.org/10.1016/J.ISPRSJPRS.2021.03.016.