# RUW-Net: A Dual Codec Network for Road Extraction From Remote Sensing Images

Jingyu Yang, *Member, IEEE*, Zongliang Gu, Ting Wu, and Yousef Ameen Esmail Ahmed

***Abstract*—Road information plays an increasingly important role in applications, such as map updating, urban planning, and intelligent supervision. However, roads in remote sensing images may be shaded by trees and buildings or interfered with by farmland. These intrinsic image features can cause road extraction results to suffer from breakage and misidentification problems. To address these problems, this article improves on D-LinkNet and proposes a dual codec structure network, namely RUW-Net. Specifically, we use ReSidual U-blocks instead of ordinary residual blocks to extract more global contextual information during the encoding stage. Moreover, we propose a decoder-encoder combination (DEC) module to build a dual codec structure. The DEC module links the decoder of the first U-block and the encoder of the following U-block to narrow the semantic gap in the encoding and decoding process. The RUW-Net model can extract more multiscale contextual features and effectively use them to enhance the semantic information of road entities. Therefore, the RUW-Net model can obtain more accurate extraction results. We conducted a series of experiments on public datasets, such as DeepGlobe, including comparative, robustness, and ablation experiments. The results show that the proposed model alleviates the road extraction breakage and misidentification problems. Compared with other representative methods, the RUW-Net performs better in terms of completeness and accuracy of road extraction results; overall, its extraction results are also the best. The RUW-Net model provides a new idea for road extraction from remote sensing images.**

***Index Terms*—Multiscale feature, remote sensing (RS) image, road extraction, semantic segmentation.**

## I. Introduction

ROAD information is basic but not negligible, as it has essential value in theoretical research and practical applications. Remote sensing (RS) images are usually acquired at a distance by artificial earth satellites, aerial aircraft, etc. Such images can be collected over large spatial areas with relatively inexpensive and objective coverage methods. And refined extraction of roads from RS images can provide decision-making data for urban intelligent traffic management, rescue and disaster relief, and joint monitoring of heaven and earth intelligence [1]. In recent years, with the rapid development of RS technology and high-resolution processing technology, the resolution of RS images has reached the submeter level. Researchers can easily access high-resolution RS data, which means a more comprehensive and richer source of data for road extraction.

In order to extract accurate roads from RS images, a large number of scholars have conducted extensive and in-depth research. They have accumulated many valuable ideas and road segmentation methods [2], [3], [4]. These methods can be broadly divided into traditional methods and deep-learning-based methods. Roads in RS images have features such as geometric texture, radiation, topological, and context features [5]. Traditional road area extraction [6], [7], [8] generally uses methods based on segmentation or classification, and after the preliminary extraction of roads, mathematical morphological methods are often used for postprocessing. While these traditional methods are simple and effective for road extraction, they often require some postprocessing operations and high labor costs.

Deep-learning-based road extraction methods can effectively reduce labor costs and advance the automation of extraction. Most of the deep-learning-based road segmentation methods are based on deep convolutional neural networks (DCNNs). In recent years, the rapid development of DCNNs has provided technical support for road segmentation. In particular, FCN and U-Net [9] have played a landmark role. After that, end-to-end semantic segmentation models based on encoder-decoder structures have gained popularity among researchers. Liu et al. [10] developed a D-Resunet that combines residual learning and U-Net. The network outperforms U-Net but suffers from the problem of missed extraction. Zhou et al. [11] designed the D-LinkNet, based on LinkNet [12]. The model uses dilated convolution [13] and incorporates multiscale features, making it possible to deal with the narrow, connected, and large-span characteristics of roads to some extent. However, the model suffers from false recognition and missing extraction. The authors in [14], [15], and [16] introduced transformer into the vision task by cutting 2D RS images into 1D image patches, and achieved some improvement on segmentation effect. Similarly, Liu et al. [17] used the Swin transformer multiscale encoder in order to efficiently extract long-distance information and designed a bottleneck module to capture more complete road structures. However, such models tend to rely too much on pretrained models, and the detailing of segmentation results is lacking.

Despite the great strides made in deep learning road extraction, there are still some challenges in extracting reliable and accurate results from RS images. Compared to other images,

Manuscript received 7 June 2023; revised 15 September 2023 and 21 October 2023; accepted 28 November 2023. Date of publication 5 December 2023; date of current version 15 December 2023. This work was supported by the National Natural Science Foundation of China under Grant 62067006. *(Corresponding author: Zongliang Gu.)*

The authors are with the School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China (e-mail: yangjy@mail.lzjtu.cn; 11210741@stu.lzjtu.edu.cn; 12211818@stu.lzjtu.edu.cn; yousef132396@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2023.3339241

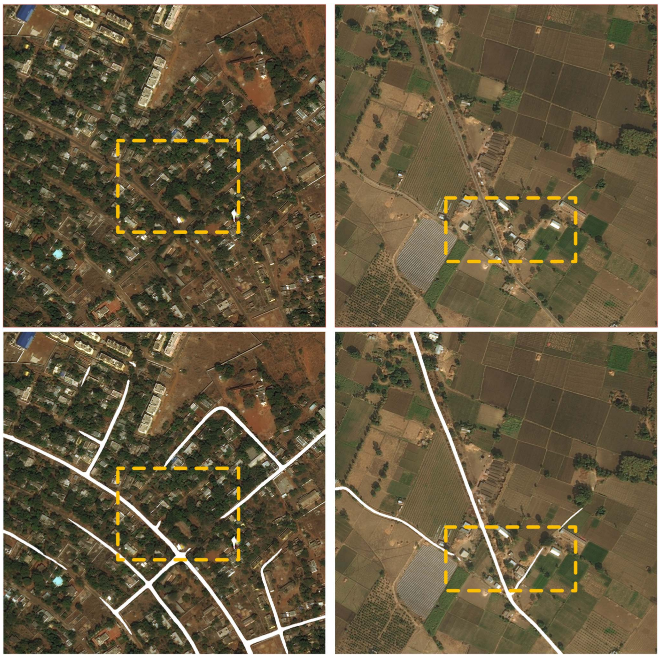

Fig. 1. Roads may be shaded by trees and buildings or interfered with by farmland.

RS images have a huge map area and contain a complex feature environment. One of the common problems is that road entities in RS images can be obscured by trees and buildings, causing breaks in the extraction (see the image on the left in Fig. 1 for an example). Another challenge is the similarity between roads and other features such as farmland, which can interfere with the extraction of roads and lead to misclassification (see the image on the right in Fig. 1 for an example). The prediction results of images in Fig. 1 are all obtained from the replicated D-LinkNet [11] model. Given the above-mentioned challenges, inspired by the ReSidual U-block [18] (RSU) and the recoding structure [19], this article improves on the D-LinkNet and proposes the RUW-Net with a dual codec structure for RS image road extraction. The basic ideas behind RUW-Net are: 1) to extract more global and multiscale contextual information to highlight road entity features in the encoding stage; 2) to construct a dual codec structure to narrow the semantic gap and predict the intermediate segmentation map after the first set of codecs to add details to the final extracted results; and 3) to design feature transfer modules for both sets of codecs to make fuller use of the extracted features. The main contributions of this article are summarized as follows.

1) First, we reconstructed the encoder of D-LinkNet by using RSU. By embedding miniature U-shaped structures, the rebuilt encoder can extract more global contextual features and multiscaled features to enhance the semantic information of the road.
2) Based on (1), we proposed a dual codec network named RUW-Net. Moreover, we designed the decoder-encoder combination (DEC) module to optimize the network's exploitation and delivery of semantic features. The DEC module connects the decoder of the first U-block to the encoder of the following U-block to narrow the semantic gap between the encoding and decoding processes. The RUW-Net model can more fully utilize the image characteristics and thus improve the completeness and accuracy of the road extraction results.
3) We have carried out extensive experiments to compare the RUW-Net model and other representative methods on publicly available datasets. The results demonstrate that our model has better performance and robustness. The RUW-Net model can obtain more complete and accurate road extraction results. More importantly, we provided a novel idea for RS road extraction.

The rest of this article is organized as follows. Section II discusses related work. Section III provides a detailed description of the methods and models in this article. Section IV contains the specific details of the experiment and the analysis of the results. Section V concludes the research and discusses future work.

## II. Related Works

This section presents the relevant work in this article. They will be briefly described in terms of semantic-segmentation-based RS road extraction, contextual information for semantic segmentation, and residual structure, respectively.

### A. Semantic-Segmentation-Based RS Road Extraction

One of the mainstream methods for road extraction from RS images is the semantic segmentation method. Yang et al. [20] developed a spatially enhanced and densely connected U-Net named SDUNet to improve road extraction by aggregating multilevel features and global priori information. Similarly, Hu et al. [21] used nested dense convolutional blocks to narrow the semantic gap for road extraction; Dai et al. [22] combined prior knowledge and variable convolution to learn the long-range dependencies of roads. Li et al. [23] designed a DC-Net model to extract rural roads using multiscale features by combining dilated convolution and ASPP structures [24]. The authors in [25], [26], and [27] used the topological features of the road to perform road extraction. Lu et al. [25] divided road extraction into three subtasks, road surface segmentation, centerline extraction, and edge detection. They explored the symbiotic relationship of the three tasks and used perceptual learning to capture the topological relationships over long distances to improve the integrity of the roads. However, the training samples of the proposed framework require three types of road samples, namely road surface, road centerline, and road edge, resulting in a restricted application scenario. Due to the complex content of RS images, the labor cost of producing high-precision road extraction datasets is high and there is a lack of high-precision datasets. Therefore, the authors in [28] and [29] used weak supervision to perform road segmentation. Similarly, Li et al. [30] designed a framework that can learn under noisy labels. In addition to this, to refine the extraction of roads, the work in [19], [31], and [32] explored road extraction algorithms in terms of the number of encoders. Jha et al. [31] proposed a new architecture called DoubleU-Net, a combination of two U-Net architectures with another U-Net added at the bottom. Wu et al.

[19] proposed a model for recoding structures with two encoders, NC-Net. Wang et al. [32] proposed a dual decoder U-shaped structure, named DDU-Net, for small-sized roads for extraction, and they improved the extraction mainly by adding a generic attention mechanism. Similar approaches that utilize attention mechanisms for road extraction can also be found in [33]. Gao et al. [33] improve road extraction model by constructing dual attention blocks within the expanded convolutional layers. Unlike them, our entry point for improvement is to build a dual decoder structure by using the DEC module for feature fusion and transfer to make fuller use of the contextual features extracted by the model.

### B. Contextual Information for Semantic Segmentation

Contextual features of images can enhance the semantic information of target classes and facilitate the extraction of roads from complex RS images. Zhou et al. [34] explored the pixel-to-pixel and pixel-to-object relationships in airborne image segmentation to learn contextual information. Yuan et al. [35] proposed a contextual representation method for semantic segmentation. They characterized the representation of corresponding object classes of pixels by calculating the relationship between each pixel and each region. The algorithm contributes some improvement to the segmentation effect on various benchmarks. The first two extract contextual features through relationships between image components. To solve the problem of contextual weakening of the overall image context modeling on the semantic level, Jin et al. [36] enhanced the pixel representation by aggregating image level and semantic level contextual information, respectively. Wu et al. [37] introduced a context-guided block to learn the joint features of local features and surrounding context. Furthermore, they designed a CGNET to capture contextual information at each stage, reducing the number of parameters and saving memory usage. Xu et al. [38] used the contextual relationship between roads and buildings to extract reliable road results. Lu et al. [39] found that capturing long-range correlations can help improve the accuracy of road recognition. Hence, they designed a Global Awareness Network to capture spatial context dependencies and interchannel dependencies. The model can establish relationships between spatial context and channels, similar to the use of GANs as [40]. He et al. [41] suggested an asymmetric encoding-decoding structure. They proposed MAE to randomly mask patches of the input image and reconstruct the pixels of the masked part. They got good results in image reconstruction with the help of contextual information. Nevertheless, MAE reconstructed the pixels of nonsemantic entities. Yang et al. [42] propose two novel modules to capture road background information in the images. By allowing different stages of the decoder to provide foreground context information to enhance the inference capability for occluded areas.

### C. Residual Structure

The residual structure was first proposed by [43] and can be a good solution to the degeneration problem of deep neural networks. Moreover, the performance of residual networks is not significantly affected by removing individual neural network layers. Therefore, combining residual structure for road extraction is also a feasible means. Eerapu et al. [44] suggested a dense refinement residual network, DRR-Net. The model can alleviate the category imbalance problem by obtaining iterative reuse of collective knowledge at different scales through dense residual connections and connectivity of DRR module problems. Liu et al. [45] developed an end-to-end residual attention local sensing network - RALC. It combined residual connectivity and attention mechanisms to design the residual attention module. Due to the positive impact of multiple feature information on road extraction, two encoders were used in the network to improve the feature extraction capability. Wu et al. [46] utilized the residual unit of ResNet, coordinate convolution, and global information enhancement module to improve the integrity and accuracy of the extracted results.

## III. Methodology

In this section, we introduce the overall RUW-Net structure and its principle first. Then we describe the RSU and the DEC module of our RUW-Net. Finally, we illustrate the feature fusion process and loss function of RUW-Net.

### A. RUW-Net Principle and Network Structure

As mentioned earlier, road entities in RS images may be affected by surface features such as trees, buildings, and farmland, making their semantics obscure. Unfortunately, weak semantics will lead to broken extraction results and false recognition. Therefore, we designed a new network model, RUW-Net. The model can utilize the contextual information of RS images to enhance the semantic information and achieve the objective of improving road extraction results. Fig. 2 illustrates the overall architecture of the RUW-Net model. The dual codec structure of the model means that the RUW-Net has two sets of encoders and decoders. The front codec is on the left, and the rear codec is on the right, making the model an overall W-shape. And the DEC module is designed to link the two sets of codecs. The decoder and encoder of the model are both five layers, En_1 to En_5 compose the front encoder, De_1 to De_5 compose the front decoder; ReE_1 to ReE_5 constitute the rear encoder, ReD_1 to ReD_5 constitute the rear decoder. Each sublayer of the front encoder consists of an RSU (RSU will be introduced in Section III-B) and a max-pooling downsample operation. The front encoder and the front decoder form the first U block through the dilated convolution at the bottom (where n equals 4) and the skip connection at each stage. The front decoder is connected to the rear encoder through the DEC module (DEC will be introduced in detail in Section III-C). The structure and connection way of the second U block's rear encoder and rear decoder are basically the same as those of the first one. The difference is that the rear encoder will combine with the output of the DEC module during downsampling, and the rear decoder will combine with the output of the DEC module for the final road segmentation

We develop the RUW-Net by improving the D-LinkNet. Unlike D-LinkNet, the encoder of RUW-Net is mainly constructed

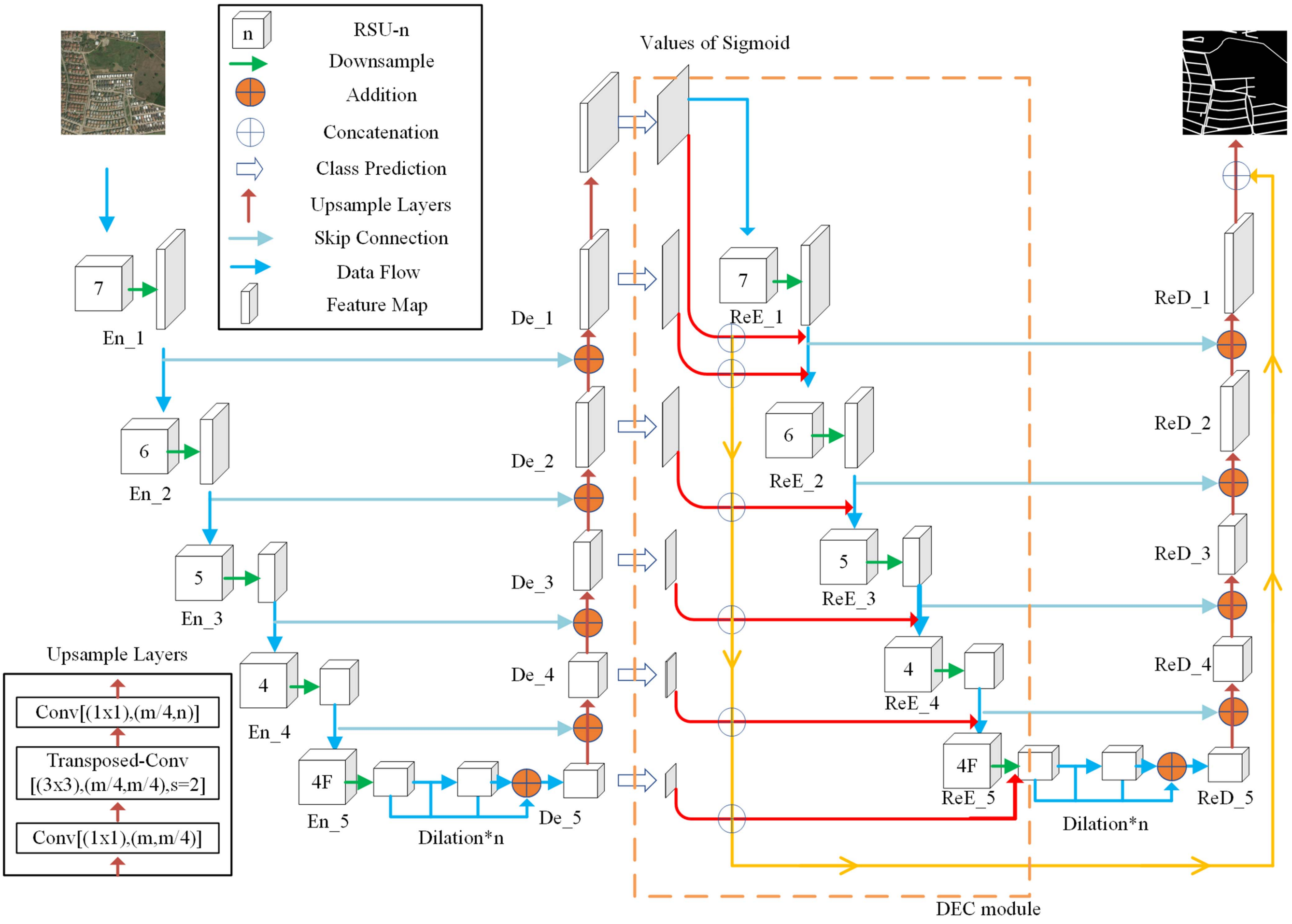

Fig. 2. Network structure of the RUW-Net.

by RSU of different depths. The reason for choosing the RSU to build the encoder is that it is significant to maintain the high resolution to extract features from the narrow and thin roads in the complex RS images. Using RSU to build the encoder, compared to the simple downsampling operation in the original D-LinkNet encoder, compensates for the impact of partial feature loss due to resolution degradation by embedding a U-shaped structure. The front encoder of the RUW-Net can be divided into five stages. The En_1, En_2, and En_3 correspond to RSU blocks with depths of 7, 6, and 5. The En_4 and En_5 are both four layers structure, and the rear encoder is similar. At the same time, considering the narrow nature, connectivity, complexity, and long span of roads, retaining more detailed feature information is necessary. Therefore, we use dilated convolution to expand the receptive field of the model in En_4 and En_5. However, performing multiple dilation convolutions on an input feature map with pristine resolution (especially in the early stages) requires too much computational and memory resources. Theoretically, it is beneficial to use dilation convolution in other modules. But in this article, we believe that the cost of applying dilation convolution is too high and the benefits are low. Therefore, considering the limited model size and computer resources, we only use multiple dilation convolutions with different dilation rates in modules En_4 and En_5. With the improved encoder, the RUW-Net can extract features at different scales directly from the RSU at different depths. Thus, the model can obtain more comprehensive contextual features, which is beneficial for enhancing semantics.

The recoding structure presented in the literature [19] restores the 16 × 16 × 512 feature map from the second encoder (H × W × C with an input image size of 512 × 512 × 3) directly to the original image size for image prediction. Although this operation reduces some of the network parameters, the final segmentation result will be affected. Because the feature map of the last layer is too small for complex RS images, the features extracted in the recoding stage cannot be fully utilized if only the “reshape” operation is performed directly. Therefore, after the feature map has been downsampled by the second encoder, we add the rear decoder part to form a dual codec structure. And the DEC module is designed to fully fuse the contextual features obtained during model training to maximize the use of the extracted semantic information. The ultimate aim is to improve the segmentation effect. The decoder structure of RUW-Net remains substantially the same as that of D-LinkNet to reduce the number of parameters and complexity of the model. Because the D-LinkNet decoder is more computationally efficient and lighter. The upsampling layers of the decoder are shown in the bottom left of Fig. 2 and consist of three layers. The input

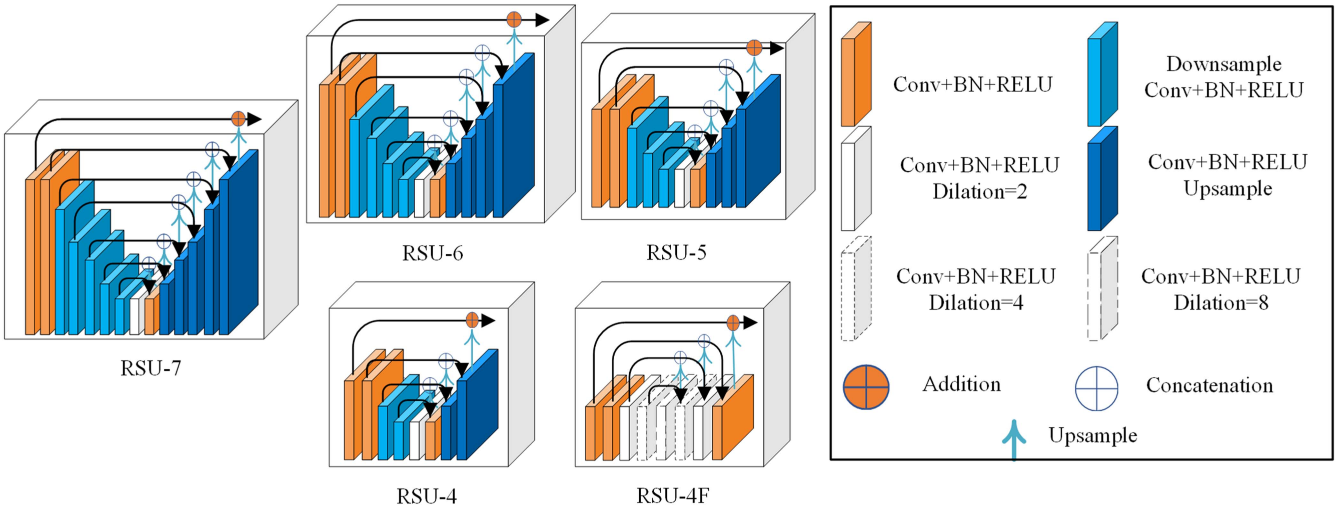

Fig. 3. Structure of RSU at different depths.

feature map of the first layer is the fused feature map of the encoder part by jump connection and De_n, which changes the number of channels by a single $1 \times 1$ convolution operation. Then, using transposed convolutional layers with stride factor equals 2. This operation increased the image size to twice the size of the input feature map while keeping the number of channels unchanged. After the transposed convolution, the number of channels is again adjusted using $1 \times 1$ convolution. Where m denotes the number of channels of the input feature map and n denotes that of the output feature map. The decoder restores the final resolution of the feature map to $1024 \times 1024$ size by upsampling layers. The DEC module is connected to each sublayer of the front decoder to extract the intermediate features of the number of categories. The rear decoder combines the fused intermediate features extracted from the DEC module when performing image segmentation prediction. The dashed rectangular area in Fig. 2 is the DEC module.

### B. ReSidual U-Block

The structure of RSU at different depths (RSU-n) is shown in Fig. 3. The parameter n represents the block depth. The deeper the depth, the more pooling operations and the larger the range of the receptive domain. Therefore, richer local and global contextual information can be extracted. The RSU is a miniature U-shaped structure that can also be divided into encoding and decoding sections. Specifically, the RSU consists of the downsampling layer (Downsample Conv + BN + RELU), the dilated convolution layer (Conv + BN + RELU Dilation = n), and the upsampling layer (Upsample Conv + BN + RELU), where the upsampling and downsampling layers transversally pass feature information through a skip connection. The upsampling layers of RSU consist of three parts, namely convolution operation, Batch Normalization operation, and RELU operation. Among them, the convolution operation is transposed convolution with a kernel size of $3 \times 3$ and a stride of 2, which expands the feature map size inside the RSU without changing the channel number. The input and output of the Batch Normalization layer are four-dimensional tensor. The Batch Normalization operation is designed to speed up the training process and improve the performance of the model by suppressing the internal covariate bias problem during the training process by normalizing each channel of the input tensor. The RELU layer is a nonlinear activation function that enhances the nonlinear representation of the mode. The input and output shapes of the RELU layer are the same as those of the convolutional and BN layers. The encoding part of the RSU has an additional feature map of the size of the input compared to the decoding part. This feature map is subjected to an addition operation with the entire downsampled processed features of RSU to retain more global information. The RSU collects multiscale contextual features from a progressively downsampled feature map. It encodes the multiscale feature map into a high-resolution feature map by progressive upsampling, concatenation, and convolution. This process mitigates the loss of detail due to direct upsampling at large scales. Fig. 4 shows the comparison of the structure of the RSU and the ordinary residual block.

Both the ordinary residual block and the RSU can be seen as consisting of three parts. First, an input convolution layer, which converts the input feature map $x$ ($\mathrm{H} \times \mathrm{W} \times \mathrm{C}$) into an intermediate feature map. It is an ordinary convolutional layer for local feature extraction, such as $3 \times 3$ convolution. Second, the intermediate feature map, which is used as the input for further extracting features. Finally, a residual concatenation operation that fuses multiscale features through an Addition operation. The difference between the ordinary residual block and the RSU lies in the way of further feature extraction in (2), which makes the features fused in (3) different too. The ordinary residual block is $\mathrm{F}_{out} = \mathrm{F}_2(\mathrm{F}_1(x)) + x$ While RSU extracts more multiscale contextual features by replacing the plain single-stream convolution with a U-shaped structure. It can be expressed as: $\mathrm{F}_{out} = \mathrm{U}(\mathrm{F}_1(x)) + x$.

The process of obtaining U (F1 (x)) is shown in Fig. 4, and the RSU-5 is used here as an example for illustration. First, the input feature map $x$($\mathrm{C} \times \mathrm{H} \times \mathrm{W}$) undergoes a $3 \times 3$ ordinary convolution to get the input F1 (x) of RSU-5. Then, adjust the

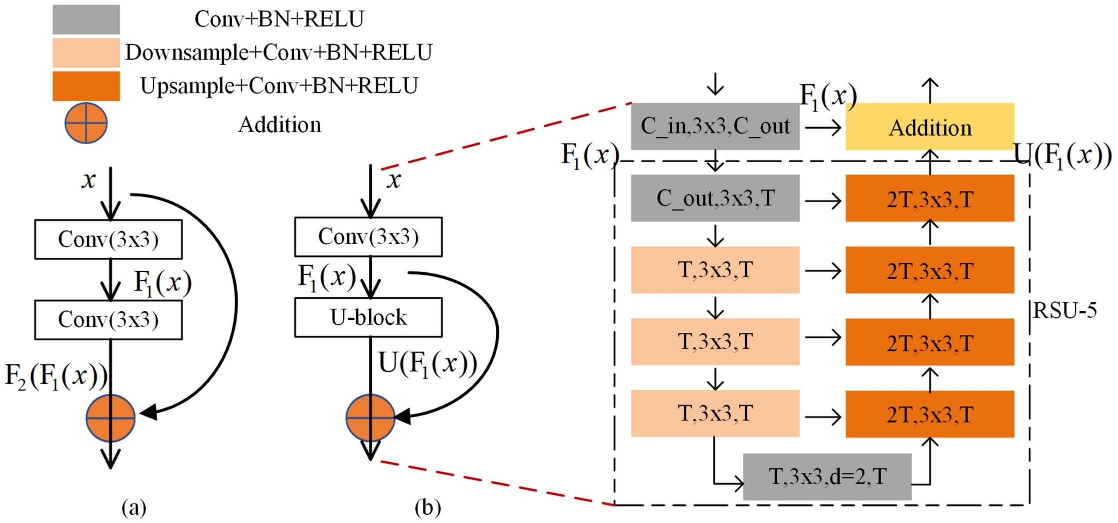

Fig. 4. Comparison Diagram of the Res-block and the ResU-block. (a) Res-block. (b) ResU-block.

channel by 3 × 3 convolution, Batch Normalization, and RELU operations. This is followed by three consecutive identical operations, each consisting of downsampling, convolution, Batch Normalization, and RELU. In the central region, a single dilation convolution with a rate of 2 is used to expand the receptive field without consuming a large amount of memory and resources. The upsampling part contains transposed convolution, Batch Normalization, and RELU operations to expand the size of the feature map and finally obtain U (F1 (x)). Finally, U (F1 (x)) will be added to F1 (x) element by element.

### *C. Decoder-Encoder Combination*

Fig. 5 is the structure of the DEC module. The DEC is mainly designed for the front decoder and the rear encoder to transfer features and fuse multiscale contextual features. Combined with Fig. 2, we add the operation of category number segmentation prediction to the feature map at the end of each sublayer of the front decoder. The purpose is to generate an intermediate binary prediction feature map. That means the input to the DEC module is the feature maps of the different sublayers of the front decoder, and the module outputs the intermediate class binary predictions of the corresponding sublayers by means of a convolution operation. As each sublayer processes a feature map of varying sizes, we can obtain the multiscale category segmentation feature map. The red curved arrows in Fig. 5 represent categorical feature transmission and the yellow arrows represent multiscale categorical feature fusion. The calculations in the DEC module come mainly from convolution and concatenation operations, so the module is lightweight and easy to use. The DEC module feeds the top two feature maps in Fig. 5 into the first sublayer of the rear encoder, while the remaining feature maps at the bottom of Fig. 5 are fed into the other sublayers of the rear encoder.

These category feature maps will be passed in two paths. Path (I) is an information fusion with the regular feature map in the rear encoder region. That will be carried through a concatenation operation after a downsampling operation. Then the rear encoder continues to filter the features through a series of RSUs. Path (II) is a contextual fusion of segmentation prediction maps at different scales. The fusion results from II are fed into the end region of the rear encoder (the region of the rear decoder pointed to by a yellow arrow in Fig. 2) to assist in the final category prediction.

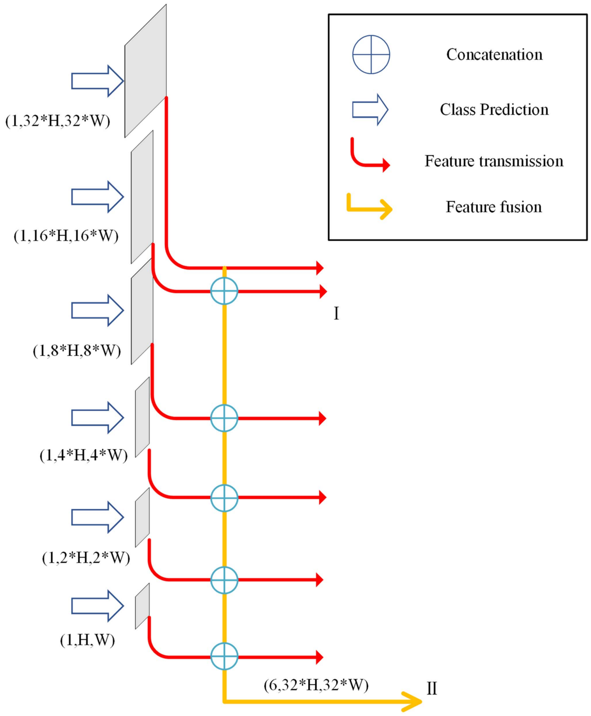

Fig. 5. Structure of the DEC module.

### *D. Feature Fusion and Loss Functions*

The overall multiscale feature fusion process of the RUW-Net model is shown in Fig. 6. This process can be divided into three parts: (a), (b), and (c). (a): Fusion of shallow and high-level contextual features by skip connections between the front and

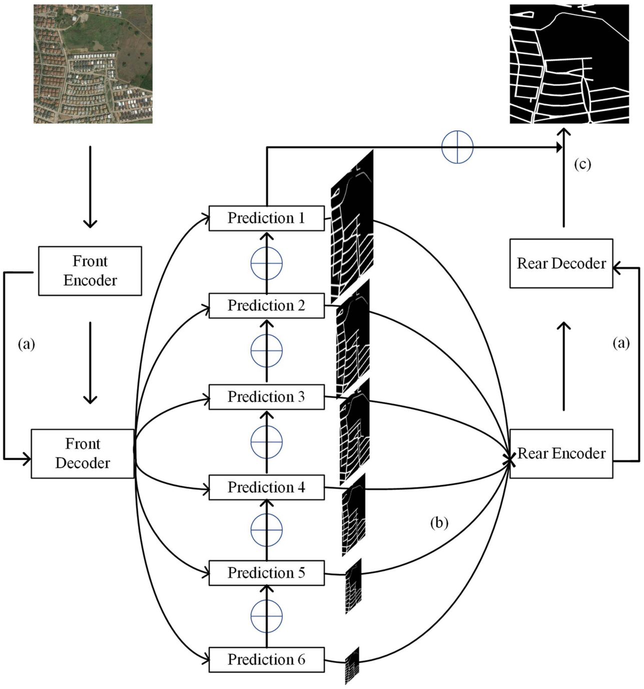

Fig. 6. Schematic diagram of feature fusion process.

rear sets of encoders and decoders. (b): With the help of path (I) of the DEC module, the front decoder can carry out efficient information transfer with the rear encoder. The front decoder passes the intermediate category prediction features to the rear encoder as additional input information. The rear encoder fuses the received features with its output features. Those features have been filtered to preserve global contextual information, which can effectively prevent information loss. (c): With the help of path (II) of the DEC module, the intermediate multiscale feature fusion results are saved and fed to the rear decoder. The rear decoder is responsible for coordinating the multiscale local and global contextual features extracted by the RUW-Net. During the fusion process, the RUW-Net is able to make fuller use of the contextual features and enhance the semantic information, thus optimizing the image segmentation results.

$$P_f = \mathrm{S}\left(P_n\right), n = 1, 2, 3, 4, 5, 6 \tag{1}$$

$$P_l = \mathrm{Concat}[P_f, \mathrm{RE}\left(x\right)], n = 1, 2, 3, 4, 5, 6. \tag{2}$$

Let $P$ be the output feature from the front decoder. $P_n$ corresponds to the feature branches generated by each sublayer of the front decoder. And the generated category segmentation features can be represented by $P_f$. These category features will then be fused with the features from the rear encoder. $P_l$ represents the features to be fed to the rear decoder after fusion, and $x$ represents the features input at different stages of the rear encoder. RE$(x)$ denotes the output feature from the rear encoder. S represents the sigmoid function.

$$L = \sum_{u=1}^{U} w_{class}^{u} l_{class}^{u} + w_{\text{fuse}} l_{fuse} \tag{3}$$

$$l = 1 - \frac{2 \times \sum_{i=1}^{N} |P_i \cap T_i|}{\sum_{i=1}^{N} |P_i \cup T_i|} + \sum_{i=1}^{N} L_{bce}(P_i, T_i). \tag{4}$$

The model uses the overall loss function shown in (3), which consists of two components. The u denotes the current image segmentation category and U denotes the total number of segmentation categories. The $l_{class}^{u}$ represents the loss function for generating intermediate category predictions in the DEC module. The $l_{fuse}$ denotes the loss for the network to fuse the output binary predictions in the rear decoder. The $w_{class}^{u}$ and $w_{\text{fuse}}$ are the weights corresponding to the loss functions. The $l_{class}^{u}$ and $l_{fuse}$ both take the form shown in (4), containing a binary cross-entropy loss function and a dice loss function. $T_i$ represents the true label value of a pixel, $P_i$ means the predicted value, and N is the total number of pixels. Road extraction can be seen as a binary classification task. The pixel values for roads are set to 1. Others are classified as backgrounds, and the pixel values are set to 0. Through the loss function, all pixel values activated in the predicted segmentation image but not in the true segmentation label image are cleared to zero. For the active pixels, it mainly punishes the prediction with low confidence. And the prediction with high confidence will get a lower loss function. The learning ability of the model is trained iteratively by minimizing the loss function.

## IV. Experiments and Analysis

In this section, we first describe the dataset and the evaluation metrics. Then, we present details of the experiments conducted and the analysis of the results. The experiments include comparative experiments, model complexity analysis, robustness experiments, and ablation experiments.

### A. Datasets

We used three datasets in the experiment, namely DeepGlobe, CHN6-CUG [47], and Massachusetts.

*DeepGlobe Road Dataset:* This dataset contains 6226 pairs of labeled 1024 × 1024 size RGB satellite RS images. The resolution of image is 0.5 m/pixel. The images are collected by Digital Globe's satellites. However, the dataset labels are not perfect, particularly in rural areas, due to the cost of annotating segmentation masks and unlabeled trails within farmland. The images cover Thailand, India, and Indonesia. And the image scenes contain a variety of environments including urban, rural, wilderness, seaside, and tropical rainforest.

*CHN6-CUG Road Dataset:* Produced and shared by the China University of Geosciences, Wuhan, China. CHN6-CUG dataset contains 4511 images of 512 × 512 size. It covers six Chinese cities, including Beijing Chaoyang District, Shanghai Yangpu District, Wuhan City Centre, Shenzhen Nanshan District, Hong Kong Shatin, and Macau. Of these images, 3608 pictures are used for model training and 903 for testing and result evaluation, with a resolution of 0.5 m/pixel.

*Massachusetts Road Dataset:* It consists of 1171 RGB aerial images of Massachusetts. Each picture is 1500 × 1500 size, with 1108 images as the training set, 14 images as the validation set, and 49 images as the test set. The dataset covers urban, suburban, and rural areas.

### B. Evaluation Indicators

In order to objectively evaluate the effectiveness of the model, Precision, Recall, F1 scores, and mIoU are used as specific evaluation indicators. The formulas for five evaluation indicators are shown in (5)–(9). The $TP$ (True Positive) represents pixels that are predicted to be roads and are real roads, $FP$ (False Positive) represents pixels that are predicted to be roads but are not roads. The $FN$ (False Negative) represents pixels that are predicted to be nonroads but are real roads. And $k+1$ denotes the total number of target categories including background. The $r^{ij}$ denotes the number of pixels predicting category *i* as category *j*. The $r^{ji}$ denotes the number of pixels predicting category *j* as category *i*. And the $r^{ii}$ denotes the number of accurately predicted pixels.

$$\text{Precision} = \frac{TP}{TP+FN} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP+FP} \tag{6}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

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

$$\text{mIoU} = \frac{\sum_{i=0}^{k} \frac{r^{ii}}{\sum_{j=0}^{k} r^{ij} + \sum_{j=0}^{k} r^{ji} - r^{ii}}}{k+1}. \tag{9}$$

### C. Experimental Results and Analysis

All experiments were conducted on Windows 10, using PyTorch 1.7 framework, two Intel(R) Xeon(R) Silver 4310 CPU, and an NVIDIA GeForce RTX 3090 GPU. We used DeepGlobe, CHN6-CUG, and Massachusetts datasets to train and evaluate the proposed RUW-Net model. The input DeepGlobe, CHN6-CUG dataset images are of default size, i.e., 1024 × 1024 and 512 × 512. For the Massachusetts dataset, the images were resized to 1024 × 1024. The learning rate was originally set 2e-4, and reduced to half while observing the training loss decreasing slowly for 4 times. The batch size during training phase was fixed as 8. We used 0.5 as our prediction threshold to generate binary outputs. Adam optimizer is used to train our network and its hyperparameters are set to default (betas = (0.9, 0.999), eps = 1e-8). We used validation dataset to train the network. Bilinear interpolation is used in resizing processes. To illustrate the effectiveness and generalization of the RUW-Net model, we performed comparative experiments, model complexity analysis, robustness experiments, and ablation experiments, respectively.

*1) Comparative Experiments:* To validate the effectiveness of the RUW-Net model, we trained the model on the DeepGlobe dataset first and compared the results with those of the other eight representative models. The comparison models include the LinkNet [12], U-Net [9], DoubleU-Net [31], NC-Net [19], D-LinkNet [11], TransUNet [14], DD-LinkNet, and U$^2$-Net [18]. The DD-LinkNet is the first version of the improved D-LinkNet, which contains the dual codec structure and the DEC module (the same below). The specific accuracy of each type of evaluation metric is shown in Table I. The bolded values are the overall optimal results, and the underlined values in the table are the second best results.

*Quantitative analysis:* From the indicators in Table I, it can be found that among the nine models, the RUW-Net model achieves the best overall performance, while the U$^2$-Net is the second best. The RUW-Net model outperforms the U$^2$-Net by 7.0%, 2.7%, 1.7%, and 2.2% in Precision, mIoU, IoU, and F1 scores, respectively. The Recall is the second best at 3.4% lower than that of the U$^2$-Net. One aspect contributing to this is the trade-off relationship between Precision and Recall metrics. RUW-Net shows a significant improvement in Precision compared to U$^2$-Net, which to some extent limits its performance in the Recall metric. Another factor is related to the network architecture of RUW-Net and U$^2$-Net. The encoder and decoder of U$^2$-Net are constructed using RSU, while the decoder part of RUW-Net, considering model complexity, does not use RSU but employs simpler and lighter upsampling layers. In comparison to the decoder part of U$^2$-Net, the capable of handling the features capability of RUW-Net decoder is slightly inferior, resulting in a second best performance in the Recall metric. In addition, the DD-LinkNet model outperforms the other six models, including the D-LinkNet. The performance of DD-LinkNet in the four metrics validates the effectiveness of the dual codec structure proposed in this article. Due to the implementation of RSU to construct the encoding module, we can train the RUW-Net model from scratch without relying on pretrained models such as the Resnet series in D-LinkNet. Although these pretrained models can speed up fine-tuning and model training to some extent, the pretrained models of Resnet and other series are not specifically trained on RS images. In other words, they cannot perfectly meet the requirements of semantic segmentation in specific RS image scenarios.

*Qualitative analysis:* Fig. 7 shows the visualization results of nine models, including RUW-Net and DD-LinkNet, on the DeepGlobe dataset. As can be seen from the ellipse-boxed area in Fig. 7, the original D-LinkNet has missing or broken content for the road extraction of RS images. The RUW-Net is better than the other eight models including the original D-LinkNet after the improvement of this article. Compared to other models, the road extraction results obtained by the RUW-Net are better in integrity and accuracy for both simple and complex RS images. From the experimental results, we can see that the RUW-Net, in the model training stage, makes up for some of the multiscale features lost in simple downsampling due to resolution degradation. Our model can extract more global contextual features by replacing ordinary residual blocks with RSUs. At the same time, the RUW-Net fuses the extracted multiscale contextual features with the help of the DEC module, making fuller use of local and global semantic information. Thus, the segmentation effect of road entities in RS images is improved.

*2) Model Complexity Analysis:* To illustrate the validity of the model structure proposed in this article, a model complexity analysis was conducted for nine models. The specific parameters are detailed in Table II. The bolded values are the overall optimal

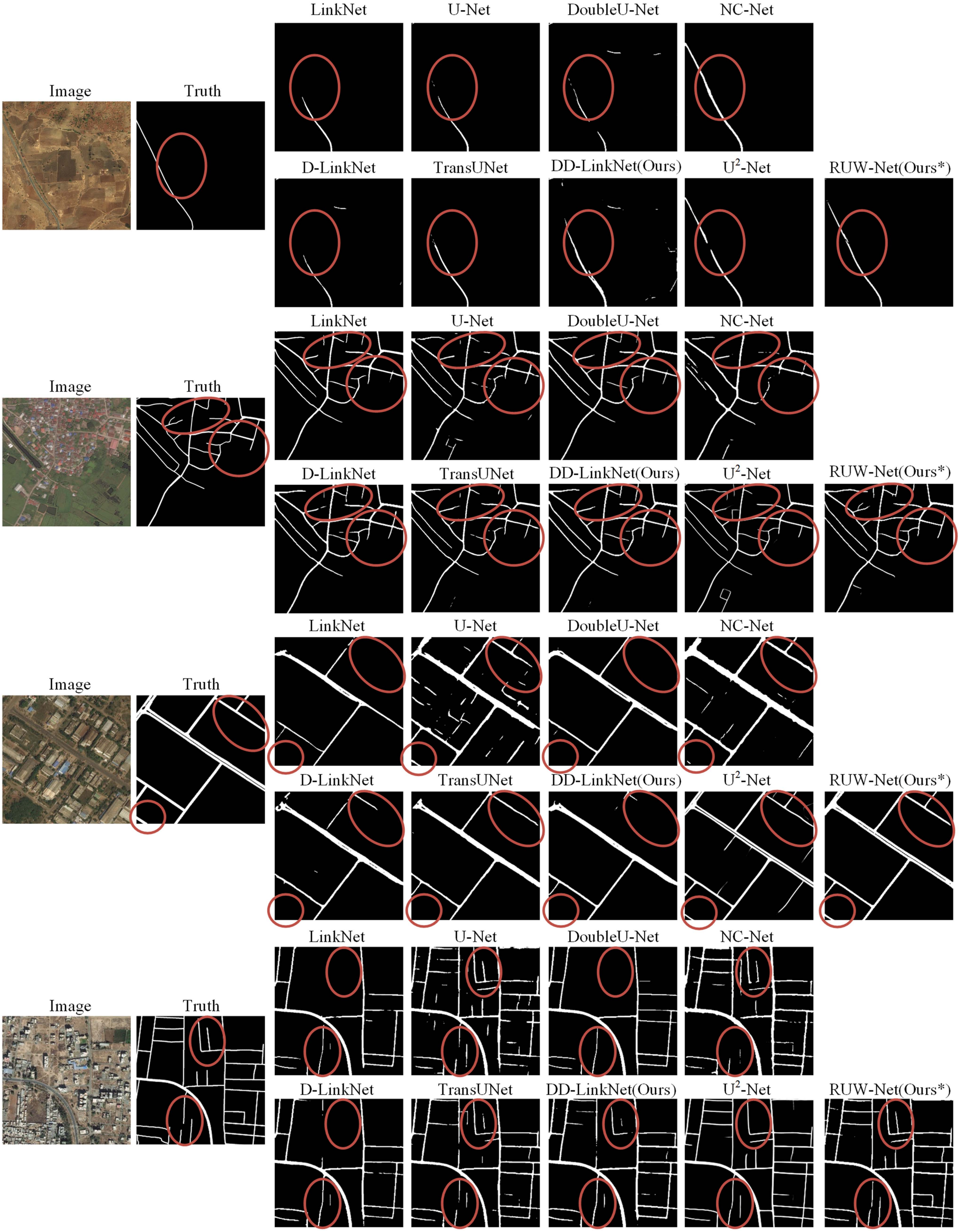

Fig. 7. Visualization of the results of different models on the DeepGlobe dataset.

TABLE I
Performance of Models on the DeepGlobe

| Model | mIoU | IoU | Recall | F1 | Precision |
|---|---|---|---|---|---|
| LinkNet | 0.713 | 0.634 | 0.591 | 0.664 | 0.760 |
| U-Net | 0.751 | 0.629 | 0.619 | 0.681 | 0.759 |
| DoubleU-Net | 0.723 | 0.649 | 0.681 | 0.659 | 0.732 |
| NC-Net | 0.742 | 0.613 | 0.714 | 0.720 | 0.727 |
| D-LinkNet | 0.757 | 0.646 | 0.697 | 0.725 | 0.756 |
| TransUNet | 0.769 | 0.662 | 0.736 | 0.739 | 0.743 |
| DD-LinkNet(Ours) | 0.767 | 0.686 | 0.732 | 0.757 | 0.784 |
| $U^2$-Net | 0.795 | 0.691 | **0.825** | 0.774 | 0.730 |
| RUW-Net(Ours*) | **0.822** | **0.708** | 0.791 | **0.796** | **0.800** |

'*' is used to indicate the final version of our model and to distinguish it from DD-LinkNet.

TABLE II
Parameters of Model Complexity

| Model | FLOPs(G) | Params(M) | F1 |
|---|---|---|---|
| LinkNet | 21.90 | 21.64 | 0.664 |
| U-Net | **5.02** | 39.50 | 0.681 |
| DoubleU-Net | 179.60 | **19.20** | 0.659 |
| NC-Net | 5.81 | 80.38 | 0.720 |
| D-LinkNet | 24.37 | 31.10 | 0.725 |
| TransUNet | 129.29 | 93.23 | 0.739 |
| DD-LinkNet(Ours) | 49.22 | 56.98 | 0.757 |
| $U^2$-Net | 150.67 | 44.01 | 0.774 |
| RUW-Net(Ours*) | 99.47 | 63.01 | **0.796** |

'*' is used to indicate the final version of our model and to distinguish it from DD-LinkNet.

results, and the underlined values in the table are the second best results.

From Table II, we can see that the RUW-Net model has a lot more computation and parameters compared to D-LinkNet, which is mainly due to the following two reasons. The first is the improvement of the encoder part of D-LinkNet, we use the RSU instead of the normal residual block for the feature extraction of the model in the downsampling stage. RSU is essentially a simple and lightweight U-shaped structure, which can effectively help the model to extract the multiscale contextual features, so we apply it in the RUW-Net model to improve the feature optimization ability. Although RSUs are relatively lightweight, applying multiple RSUs will inevitably increase the amount of model computation and number of parameters.

The second is the extension of the D-LinkNet model to the W-type. Why do we complicate the model? Because roads usually have a small percentage of pixels in complex RS images, mostly appearing as elongated strips, and the texture structure is easily affected by similar features in the neighborhood. These reasons make it difficult to extract good results by the conventional U-shaped structure [9], [11], [12]. Therefore, we take a new perspective and try to use two U-shaped structures to form a W-shape to enhance the model to extract more long-range relationships and more robust features. This process inevitably increases the complexity of the model. In our future work, we will streamline the model, reduce its complexity, and improve the practicality and ease of use while ensuring its accuracy.

The RUW-Net model has an increased complexity compared with D-LinkNet, but it achieves the highest F1 scores with medium FLOPs and Params metrics among all models. In other words, the effective improvement of the RUW-Net model on the road extraction results is mainly due to the useful dual codec structure and the DEC module designed in this article, rather than by simply expanding the number of model parameters.

*3) Robustness Experiments:* We designed an experiment to verify the robustness of our model against the interference of other surface features when extracting roads. The random mask is applied to the road entities in the test set to simulate the interference of trees, buildings, farmland, and other features in RS images. The number of occlusions is set to 10, the occluded area is set to a size of $20 \times 20$. The pixel value of the occluded area is set to the average pixel value of the original image, and the image has a $1024 \times 1024$ size. We masked the training data in the same way as the test images during the training of the RUW-Net model.

Table III shows the experimental results of the robustness of the three models, the D-LinkNet, the DD-LinkNet, and the RUW-Net on the DeepGlobe dataset. The bolded values are the overall optimal results, and the underlined values in the table are the second best results. From the results, it is easy to see that the robustness of both the DD-LinkNet, and the RUW-Net is better than that of the D-LinkNet. The average decrease in each metric is 2.5% for the D-LinkNet, and 1.6% for the DD-LinkNet with the DEC module added. The RUW-Net, which rebuilds the encoder on the basis of DD-LinkNet, saw the smallest decrease of only 1.0%. That means the extraction effect of the RUW-Net is still good. To further illustrate the robustness of the model in this article, the results are visualized in Fig. 8. It can be seen from the rectangular box marking part that the RUW-Net performs best and shows strong resistance to interference. From the observed results, we can conclude that both the DEC module proposed in this article and the dual codec structure built on top of it are conducive to improving the road extraction results of RS images.

*4) Ablation Experiments:* To further validate the effectiveness and generalization of the RUW-Net model with a dual codec structure and the DEC module proposed in this article,

TABLE III
MODEL ROBUSTNESS COMPARISON ON THE DEEPGLOBE

| Dataset | Model | mIoU | IoU | Recall | F1 | Precision |
|---|---|---|---|---|---|---|
| DeepGlobe | D-LinkNet | 0.757 | 0.646 | 0.697 | 0.725 | 0.756 |
| | D-LinkNet (mask) | 0.732 (↓2.5%) | 0.616 (↓3.0%) | 0.665 (↓3.2%) | 0.701 (↓2.4%) | 0.742 (↓1.4%) |
| | DD-LinkNet(Ours) | 0.767 | 0.686 | 0.732 | 0.757 | 0.784 |
| | DD-LinkNet(Ours) (mask) | 0.751 (↓1.6%) | 0.672 (↓1.4%) | 0.706 (↓2.6%) | 0.739 (↓1.8%) | 0.776 (↓0.8%) |
| | RUW-Net(Ours*) | 0.822 | 0.708 | 0.791 | 0.796 | 0.800 |
| | RUW-Net(Ours*) (mask) | **0.811** (↓1.1%) | **0.701** (↓0.7%) | **0.774** (↓1.7%) | **0.785** (↓1.1%) | **0.797** (↓0.3%) |

'*' is used to indicate the final version of our model and to distinguish it from DD-LinkNet.

TABLE IV
PERFORMANCE OF MODELS ON DIFFERENT DATASETS

| Dataset | Model | RSU | DEC | mIoU | IoU | Recall | F1 | Precision |
|---|---|---|---|---|---|---|---|---|
| DeepGlobe | D-LinkNet | -- | -- | 0.757 | 0.646 | 0.697 | 0.725 | 0.756 |
| DeepGlobe | DD-LinkNet(Ours) | -- | √ | 0.767 | 0.686 | 0.732 | 0.757 | 0.784 |
| DeepGlobe | RUW-Net(Ours*) | √ | √ | **0.822** | **0.708** | **0.791** | **0.796** | **0.800** |
| CHN6-CUG | D-LinkNet | -- | -- | 0.773 | 0.627 | 0.645 | 0.706 | 0.780 |
| CHN6-CUG | DD-LinkNet(Ours) | -- | √ | 0.771 | 0.649 | 0.674 | **0.742** | **0.826** |
| CHN6-CUG | RUW-Net(Ours*) | √ | √ | **0.800** | **0.676** | **0.721** | 0.740 | 0.761 |
| Massachusetts | D-LinkNet | -- | -- | 0.781 | 0.652 | 0.634 | 0.719 | 0.830 |
| Massachusetts | DD-LinkNet(Ours) | -- | √ | 0.767 | 0.658 | 0.656 | 0.741 | 0.852 |
| Massachusetts | RUW-Net(Ours*) | √ | √ | **0.799** | **0.691** | **0.681** | **0.767** | **0.877** |

'*' is used to indicate the final version of our model and to distinguish it from DD-LinkNet.

we extended the datasets and designed the ablation experiments on the DeepGlobe, CHN6-CUG, and Massachusetts datasets, respectively. Three sets of experiments were set up for each dataset, corresponding to the D-LinkNet, the DD-LinkNet, and the RUW-Net. The models were all trained for 100 epochs and evaluated on the same test set for each evaluation metric.

*Quantitative analysis:* The specific accuracy is detailed in Table IV. The bolded values are the overall optimal results, and the underlined values in the table are the second best results. As can be seen from the table, the DD-LinkNet with the DEC module has a 1.0%, 4.0%, 3.5%, 3.2%, and 2.8% improvement in mIoU, IoU, Recall, F1 scores, and Precision, respectively, compared to the D-LinkNet in the DeepGlobe dataset. The RUW-Net has the DEC module and the rebuilt encoder using RSU. Compared with the DD-LinkNet which only added the DEC module, it shows significant improvements in each metric. Specifically, The RUW-Net has an average enhancement of 3.8%. In the CHN6-CUG dataset, the RUW-Net outperforms the DD-LinkNet in mIoU, IoU, and the Recall but is lower in Precision. The reason may be that the labels of CHN6-CUG are relatively coarse-grained and there is a constraint relationship between Precision and Recall. The DD-LinkNet performs better in Recall, IoU, Precision and F1 scores compared to the D-LinkNet. The average improvement is 3.3%. In the Massachusetts dataset, the DD-LinkNet model improved on average by 1.8% in Precision, Recall, IoU, and F1 scores compared to D-LinkNet. The RUW-Net improved on average by 2.9% in all metrics compared to DD-LinkNet. The specific accuracy of the models on the three different datasets shows that the RUW-Net achieves the best overall and the DD-LinkNet is the second best.

*Qualitative analysis:* To more intuitively illustrate the effectiveness of the RUW-Net on the RS road extraction, we selected two images from each dataset for results comparison. Fig. 9 shows the visualization results of the ablation experiment. From the area circled by the ellipse in the figure, we know that the RUW-Net model performs better overall than DD-LinkNet and D-LinkNet in all three datasets. It can extract more complete road shapes and more accurate extraction results. Through the ablation experiments of RUW-Net, DD-LinkNet, and D-LinkNet, we can conclude that both the RUW-Net dual codec structure and DEC module are beneficial to the refined extraction of RS roads.

All the above-mentioned experimental results demonstrate the effectiveness and generalization of the RUW-Net model with the dual codec structure proposed in this article. The RUW-Net model can extract rich local and global contextual features in the

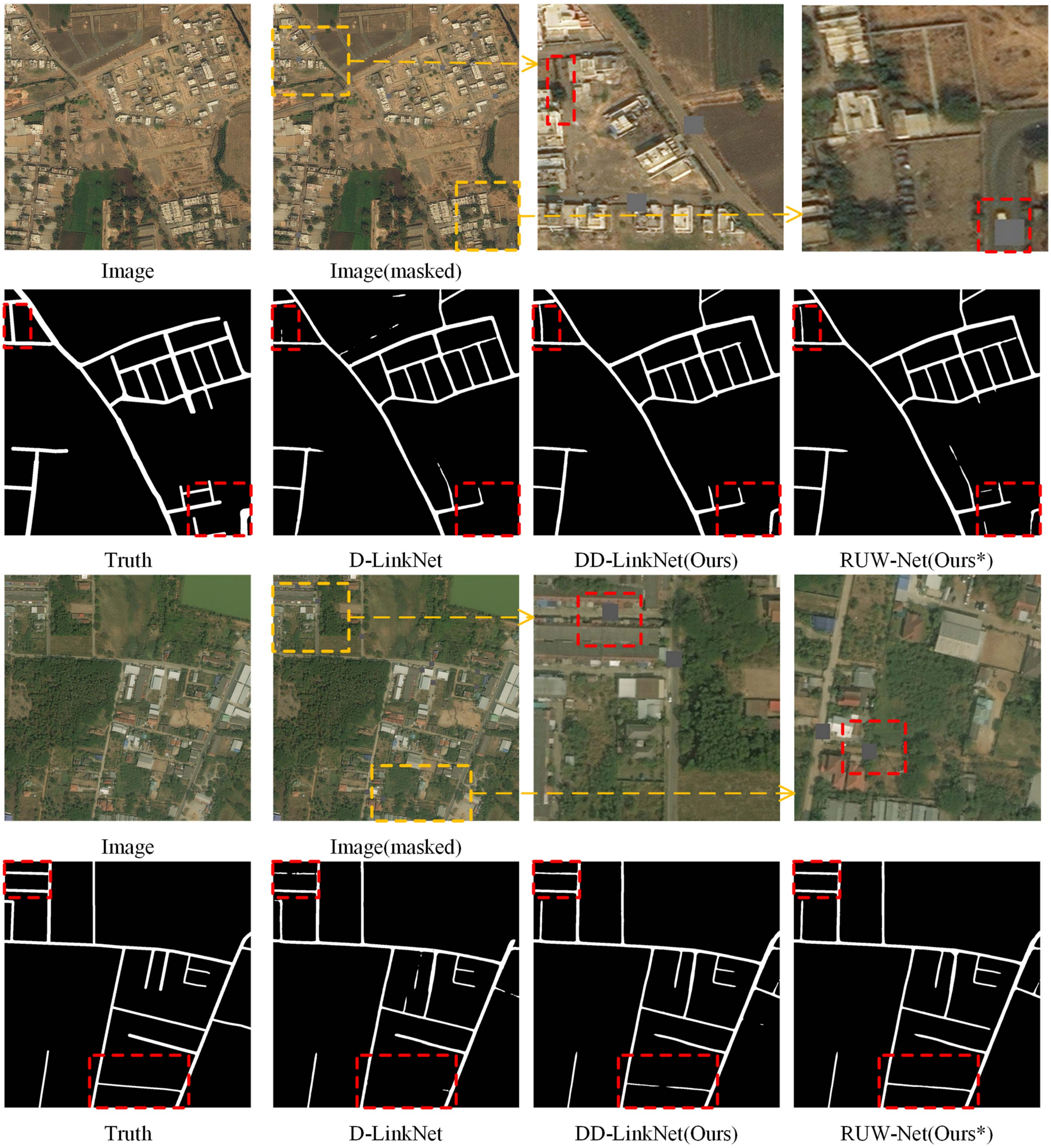

Fig. 8. Visualization of model robustness on the DeepGlobe.

encoding stage. The DEC module of the RUW-Net facilitates the feature selection process between the front decoder and the rear encoder. The dual codec structure formed by the model extracts more multiscale contextual features. After feature transfer by the DEC module, the multiscale features are fused with the features obtained from downsampling in the rear encoder region. The semantic information of road entities is enhanced in the rear decoder using the global contextual features, thus improving the road extraction results of RS images.

## V. Discussion

Extracting roads from RS images to produce high-quality road networks is actually a challenging task. Factors including occlusion of objects, texture interference, as well as the narrow and elongated nature of the roads themselves, often result in challenges such as fragmented extraction and missing content when extracting roads from RS images. Although classic models such as U-Net, LinkNet, and D-LinkNet are lightweight and capable of quickly extracting roads, they often learn primarily local features, resulting in good extraction of local details. However, due to the lack of global information, particularly global contextual information, these models struggle to meet the standards required for road extraction in practical applications. Recently, many researchers have been exploring improvements to DCNN models by focusing on the network architecture. For example, as mentioned earlier, models such as NC-Net and DoubleU-Net have been developed to enhance the model's

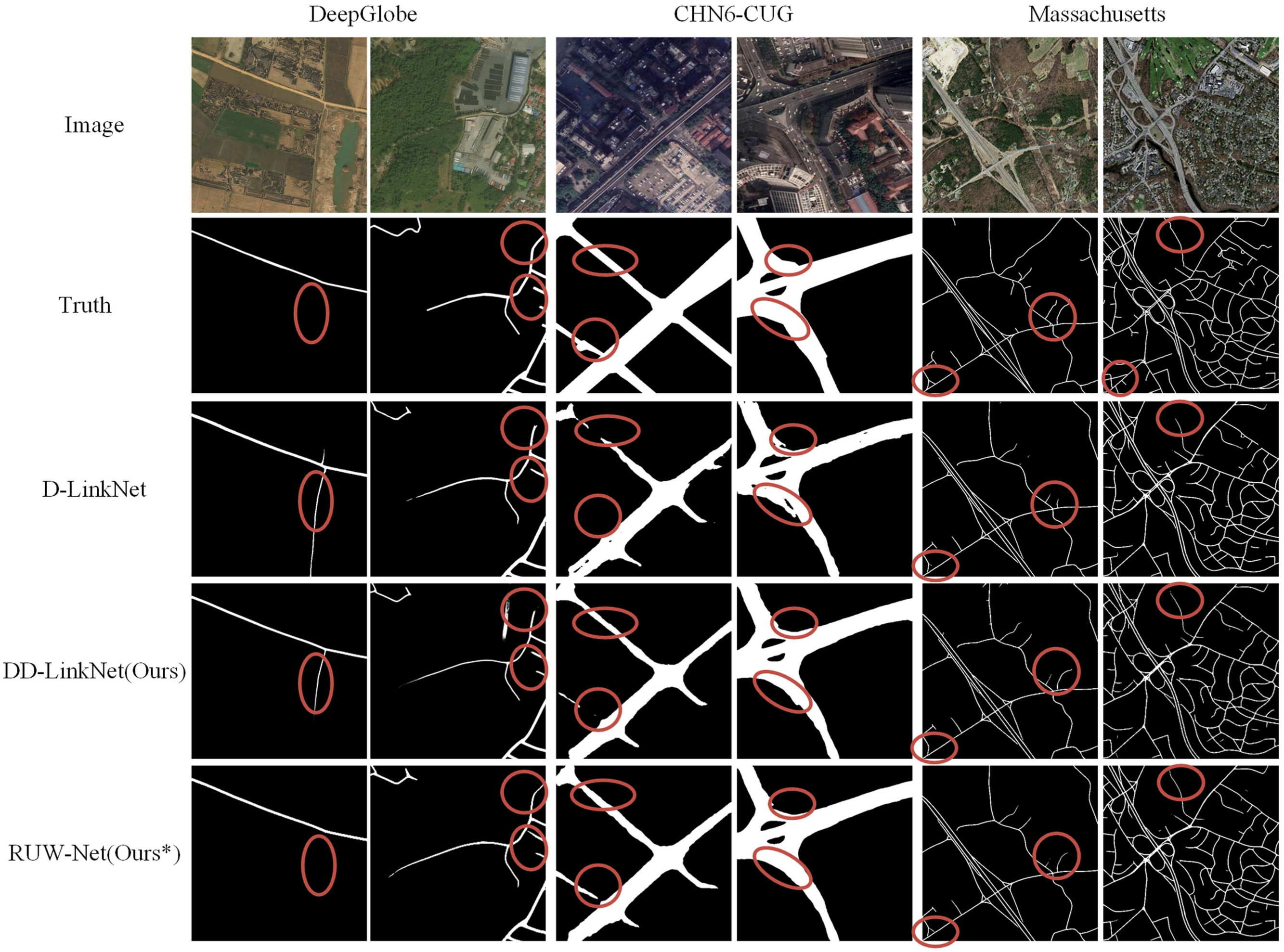

Fig. 9. Visualization of the results of models on different datasets.

perception of contextual information by increasing the number of encoders or decoders.

Similar to their work, this article also focuses on the overall network architecture by increasing the number of encoders and decoders. However, what sets our work apart is that we further adjust the network architecture. In our approach, we embed a mini-U-shaped structure within the encoder section, allowing for the repeated utilization of U-shaped structures to extract more global contextual features. In addition, we effectively utilize features by incorporating DEC modules within the W-shaped network structure. In Section IV, we conducted experiments and analysis to evaluate the effectiveness, model complexity, robustness, and generalization of our proposed model. Based on these quantitative and qualitative experimental results, it is evident that our model performs well in high-resolution road extraction from RS imagery, outperforming the other comparative models presented in this article. We believe that our model is valuable in terms of utilizing contextual information to efficiently carry out the extraction of road networks.

Certainly, our proposed model has its limitations. Enhancing the feature representation capability of a model through a nested stacked U-shaped structure inevitably increases the complexity of the model, leading to higher computational resource requirements. How to simplify the model while maintaining its performance will be a focus of our future work. In practice, algorithms often perform worse than their theoretical counterparts due to various factors that are overlooked in theoretical analysis, such as noise. The robustness of the model will be higher if it can incorporate stable statistical features and not only rely on data-driven extraction of learned features. If data-driven algorithms can be combined with heuristic algorithms, then statistical features can be utilized as well as learned features, and the cost of data labeling can be reduced at the same time. We believe that such methods will further improve the robustness and usefulness of RS road extraction.

## VI. Conclusion

This article improves the D-LinkNet model and proposes a RUW-Net semantic segmentation model with a dual codec structure for road extraction from RS images. The model alleviates the problem that road entities are easily disturbed by other surface features, resulting in inconspicuous semantic information and further leading to broken and misidentified extracted roads. With the help of our designed dual codec structure and the DEC module, the RUW-Net model can capture and fuse more multiscale contextual features. These features can enhance the semantic information during the model training stage and improve the road extraction result. A series of experiments on the three datasets show that our RUW-Net is feasible and effective. Compared with other representative methods, the RUW-Net model has more complete extraction results and higher accuracy. The RUW-Net also provides a new idea for RS image road extraction. However, the model in this article also increases the computational overhead, and the next step is to consider how to streamline the network structure under the premise of accuracy.

## References

[1] A. Van Etten and I. C. Soc, "City-scale road extraction from satellite imagery V2: Road speeds and travel times," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2020, pp. 1775–1784.

[2] R. Lian, W. Wang, N. Mustafa, and L. Huang, "Road extraction methods in high-resolution remote sensing images: A comprehensive review," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5489–5507, Sep. 2020, doi: 10.1109/jstars.2020.3023549.

[3] Z. Chen et al., "Road extraction in remote sensing data: A survey," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, Aug. 2022, Art. no. 102833, doi: 10.1016/j.jag.2022.102833.

[4] J. Dai, Y. Wang, and Y. Du, "Overview of road extraction methods from optical remote sensing images," *J. Remote Sens.*, vol. 24, no. 7, pp. 804–823, 2020.

[5] J. Du, R. Li, and F. Jin, "A modified road centerlines search method from remote sensing images," in *Proc. IEEE Int. Conf. Saf. Produce Informatization*, 2019, pp. 192–195.

[6] J. Dai, T. Zhu, and Y. Zhang, "A road extraction method for high-resolution remote sensing image," *J. Autom.*, vol. 46, no. 11, pp. 2461–2471, 2020, doi: 10.16383/j.aas.c190534.

[7] R. Alshehhi and P. Marpu, "Hierarchical graph-based segmentation for extracting road networks from high-resolution satellite images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 126, pp. 245–260, 2017, doi: 10.1016/j.isprsjprs.2017.02.008.

[8] J. Dai, Z. Miao, and L. Ge, "A road extraction method for high-resolution remote sensing images combined with path morphology," *J. Inf.*, vol. 34, no. 01, pp. 28–35, 2019.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.

[10] Z. Liu, R. Feng, L. Wang, Y. Zhong, and L. Cao, "D-resunet: Resunet and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 3927–3930.

[11] L. Zhou, C. Zhang, and M. Wu, "D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 192–196.

[12] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process.*, 2017, pp. 1–4.

[13] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Representations*, 2016, Art. no. 52152.

[14] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:abs/2102.04306*.

[15] H. Cao et al., "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Visi.*, 2022, pp. 205–218.

[16] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 4408715, doi: 10.1109/TGRS.2022.3144165.

[17] X. Liu et al., "RoadFormer: Road extraction using a Swin transformer combined with a spatial and channel separable convolution," *J. Remote Sens.*, vol. 15, no. 4, Jan. 2023, Art. no. 4, doi: 10.3390/rs15041049.

[18] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U$^2$-Net: Going deeper with nested U-structure for salient object detection," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107404, doi: 10.1016/j.patcog.2020.107404.

[19] Z. Wu, S. Zhao, and H. Li, "Remote sensing image semantic segmentation space global context information network," *J. Zhejiang Univ.*, vol. 56, no. 4, pp. 795–802, 2022. [Online]. Available: https://kns.cnki.net/kcms/detail/33.1245.T.20220408.0936.004.html

[20] M. Yang, Y. Yuan, and G. Liu, "SDUNet: Road extraction via spatial enhanced and densely connected UNet," *Pattern Recognit.*, vol. 126, Jun. 2022, Art. no. 108549, doi: 10.1016/j.patcog.2022.108549.

[21] H. Hu, J. Zuo, and Y. Lv, "Remote sensing imagery road network detection method for automated driving," pp. 310–317, vol. 35, no. 11, 2022, doi: 10.19721/j.cnki.1001-7372.2022.11.026.

[22] L. Dai, G. Zhang, and R. Zhang, "RADANet: Road augmented deformable attention network for road extraction from complex high-resolution remote-sensing Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 2023, Art. no. 5602213, doi: 10.1109/tgrs.2023.3237561.

[23] C. Li, Q. Zeng, and J. Fang, "Improved full convolution network method for rural road extraction from Gaofen 2 images," *J. Remote Sens.*, vol. 25, no. 9, pp. 1978–1988, 2021.

[24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/tpami.2017.2699184.

[25] X. Lu et al., "Cascaded multi-task road extraction network for road surface, centerline, and edge extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5621414, doi: 10.1109/tgrs.2022.3165817.

[26] Z. Xu et al., "RNGDet: Road network graph detection by transformer in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 4707612, doi: 10.1109/tgrs.2022.3186993.

[27] J. Zhang, X. Hu, Y. Wei, and L. Zhang, "Road topology extraction from satellite imagery by joint learning of nodes and their connectivity," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 5602613, doi: 10.1109/tgrs.2023.3241679.

[28] Y. Wei and S. Ji, "Scribble-based weakly supervised deep learning for road surface extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5602312, doi: 10.1109/tgrs.2021.3061213.

[29] R. Lian and L. Huang, "Weakly supervised road segmentation in high-resolution remote sensing images using point annotations," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 4501013, doi: 10.1109/tgrs.2021.3059088.

[30] P. Li et al., "Exploring label probability sequence to robustly learn deep convolutional neural networks for road extraction with noisy datasets," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2022, Art. no. 5614018, doi: 10.1109/tgrs.2021.3128539.

[31] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-Net: A deep convolutional neural network for medical image segmentation," in *Proc. IEEE Symp. Comput.-Based Med. Syst.*, 2020, pp. 558–564.

[32] Y. Wang et al., "DDU-Net: Dual-decoder-U-Net for road extraction using high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 4412612, doi: 10.1109/tgrs.2022.3197546.

[33] L. Gao et al., "Road extraction using a dual attention dilated-LinkNet based on satellite images and floating vehicle trajectory data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10428–10438, Sep. 2021, doi: 10.1109/jstars.2021.3116281.

[34] F. Zhou, R. Hang, H. Shuai, and Q. Liu, "Hierarchical context network for airborne image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 2022, Art. no. 4407612, doi: 10.1109/tgrs.2021.3133258.

[35] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proc. Comput. Vis.-ECCV 2020: 16th Eur. Conf.*, 2020, pp. 173–190.

[36] Z. Jin, B. Liu, Q. Chu, and N. Yu, "ISNet: Integrate image-level and semantic-level context for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7169–7178.

[37] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, "CGNet: A light-weight context guided network for semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 1169–1179, Dec. 2021, doi: 10.1109/tip.2020.3042065.

[38] Y. Xu, H. Chen, C. Du, and J. Li, "MSACon: Mining spatial attention-based contextual information for road extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5604317, doi: 10.1109/tgrs.2021.3073923.

[39] X. Lu, Y. Zhong, and Z. Zheng, "A novel global-aware deep network for road detection of very high resolution remote sensing imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 2579–2582.

[40] W. Chen, G. Zhou, Z. Liu, X. Li, X. Zheng, and L. Wang, "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.*, vol. 60, Jul. 2022, Art. no. 5626115, doi: 10.1109/tgrs.2022.3188908.

[41] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15979–15988.

[42] Z. Yang, D. Zhou, Y. Yang, J. Zhang, and Z. Chen, “TransRoadNet: A novel road extraction method for remote sensing images via combining high-level semantic feature and context,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, May 2022, Art. no. 6509505, doi: 10.1109/LGRS.2022.3171973.

[43] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[44] K. K. Eerapu, B. Ashwath, S. Lal, F. Dellacqua, and A. V. N. Dhan, “Dense refinement residual network for road extraction from aerial imagery data,” *IEEE Access*, vol. 7, pp. 151764–151782, 2019, doi: 10.1109/access.2019.2928882.

[45] Z. Liu, M. Wang, F. Wang, and X. Ji, “A residual attention and local context-aware network for road extraction from high-resolution remote sensing imagery,” *Remote Sens.*, vol. 13, no. 24, Dec. 2021, Art. no. 4958, doi: 10.3390/rs13244958.

[46] Q. Wu, S. Wang, and B. Wang, “Spatial information-aware semantic segmentation model for road extraction from high-resolution remote sensing images,” *J. Remote Sens.*, vol. 26, no. 9, pp. 1872–1885, 2022.

[47] Q. Zhu et al., “A global context-aware and batch-independent network for road extraction from VHR satellite imagery,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 175, pp. 353–365, May 2021, doi: 10.1016/j.isprsjprs.2021.03.016.

**Jingyu Yang** (Member, IEEE) received the Ph.D. degree in intelligent transportation and information systems engineering from Lanzhou Jiaotong University, Lanzhou, China, in 2019.

He is currently the backbone research member of the innovation team of “Research on Information Engineering and Control Technology of Plateau Transportation” of the Ministry of Education and the key research member of the virtual simulation experimental teaching team of Gansu Province Railway Transportation Information and Control Course. His research interests include remote sensing monitoring, pattern recognition, image processing algorithms, and intelligent computing.

**Zongliang Gu**, biography and photograph not available at the time of publication.

**Ting Wu**, biography and photograph not available at the time of publication.

**Yousef Ameen Esmail Ahmed**, biography and photograph not available at the time of publication.