IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 February 2023, accepted 28 February 2023, date of publication 6 March 2023, date of current version 9 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3253289

RESEARCH ARTICLE

AGD-Linknet: A Road Semantic Segmentation Model for High Resolution Remote Sensing Images Integrating Attention Mechanism, Gated Decoding Block and Dilated Convolution

YINAN JIANG, CHAOLIANG ZHONG, AND BOTAO ZHANG
School of Automation, Hangzhou Dianzi University, Hangzhou, Zhejiang 310018, China

Corresponding author: Chaoliang Zhong (zhongchaoliang@hdu.edu.cn)

This work was supported in part by the Natural Science Foundation of Zhejiang Province China under Grant LY21F030009, and in part by the Key Research and Development Program of Zhejiang Province under Grant 2019C04018.

ABSTRACT Road information is an important geographic information. Road information extracted from remote sensing images has been widely used in map, traffic, navigation and many other fields. However, the autonomous extraction of road information from high resolution remote sensing images has some problems such as incoherence, incompleteness and poor connectivity, therefore, a semantic segmentation model for roads in high resolution remote sensing images, called AGD-Linknet, is proposed, which integrates attention mechanisms, gated decoder block, and dilated convolution. This model mainly consists of three parts. Firstly, the stem block is used as the initial convolution layer of the model to reduce the information loss in the convolution stage; Secondly, the series-parallel combined dilated convolution and coordinate attention block into the center of the network, which enlarges the receptive field of the network and improves the feature extraction ability of spatial domain and channel domain information; Finally, in the decoder part, gated convolution is introduced to improve the extraction of road edge. Compared with U-Net, Linknet and D-Linknet on the DeepGlobe dataset, the proposed AGD-Linknet has improved the pixel accuracy, mean intersection over union and F1-Score index of road recognition by 1.41%-11.52%, 0.0077-0.1473, 0.0057-0.1292, and has certain effectiveness and feasibility in many scenarios in rural areas, urban, and suburbs. And can be apply to the tasks of road recognition and extraction in high-resolution remote sensing.

INDEX TERMS Feature extraction, remote sensing, road extraction, convolutional neural networks, attention mechanisms, dilated convolution, gated convolution.

I. INTRODUCTION

Remote sensing data show the characteristics of high spatial, high spectral and high temporal resolution [1]. Compared with the low-resolution remote sensing image data, the high-resolution remote sensing image has more abundant ground information, spatial structure and geometric texture features. Road information, as an important part of surface image data, has been widely used in urban planning, military operations, disaster relief and many other aspects, and has received extensive attention from researchers [2].

The methods of road extraction from high resolution remote sensing images can be divided into two categories, namely, the traditional road extraction method and the modern road extraction method based on deep learning. The traditional road extraction method is an experience-based extraction method, which uses the shallow features of the image (such as grayscale, edge, color, texture and geometric shape, etc.). The modern road extraction method uses deep neural network to implicitly extract deep abstract features

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson.

from the original image, and uses the features to extract road information autonomously.

The traditional road extraction methods include grayscale segmentation method [3], [4], template matching method [5], morphological algorithm [6], [7] and extraction algorithm based on edge feature [8], [9]. For example, Jie et al. [3] proposed an improved watershed algorithm based on threshold marking, which can use gray segmentation method to extract road information from remote sensing images. Lin et al. [5] used template matching algorithm to extract road information. Wang [6] proposed an image denoising method based on path morphology to extract road networks. Faming et al. [8] proposed an automatic road extraction method based on Canny edge detection operator applied to roads in mining areas.

Although the traditional methods can extract the road information in specific images to a certain extent, it is difficult to extract effectively in complex and changeable environments. For example, vehicles, sidewalk signs and building shadows in urban environment all affect the extraction results of the traditional road extraction methods. Especially, with the improvement of remote sensing data resolution, the ground object information, spatial structure and geometric texture features in the image are more abundant, and the road extraction effect of traditional methods is also decreasing. In a word, the traditional road extraction method has the characteristics of weak feature expression ability, poor robustness, small adaptation range, and dependence on human subjective judgment, which is difficult to meet the practical application requirements in complex and changing environments.

In view of the shortcomings of traditional road extraction methods, the road extraction method based on deep learning has been widely concerned and studied because of its advantages of high extraction efficiency, good effect and wide range of adaptation [10], [11], [12], [13], [14], [15], [16], [17], [18]. Ronneberger et al. [10] proposed the U-net network based on FCN [11], which is a U-shaped network structure that uses transposed convolution as its upsampling module and connects features from the encoder part of the network to the decoder part to fuse information from shallow and deep features. Such fusion helps the network to recover more spatial information during the upsampling process, which is very important for fine-grained segmentation. Kaiming He et al. [16] proposed the residual network (ResNet) to solve the problem that the deepening of the depth of the neural network may lead to network degradation in training and gradient disappearance. The Linknet proposed by Chaurasia and Culurciello [12] combines the above two networks, which is a variation of the U-type network, changing the way of combining the deep and shallow features of the U-net from ''concatenate'' to ''skip connection''. This makes the network optimization convergence simpler and can better reduce the memory consumption and improve the computational efficiency. Zhang et al. [18]proposed a semantic segmentation neural network by combining residual learning and U-Net to extract road information from high-resolution remote sensing images. Xin et al. proposed DenseUNet [17], consisting of dense connection units and skip connections which can help deal with the problem of tree and shadow occlusion. In the above-mentioned network structure, the initial input layer often uses two successive downsampling to convert the input information into fixed-size image information before training, which leads to a large amount of information loss. STDN [13] has tried to solve this problem by using the Stem block and achieved good results.

The network using the U-shaped structure often faces the problem of insufficient receptive field in the central part, and the segmentation effect of small objects is not good. Therefore, it is necessary to combine multi-scale information to improve the network structure. Dilated convolution has been proposed to improve the receptive field of objects. For example, DeepLabV3 [14] uses the ASPP (Atrous spatial pyramid pooling) module and uses convolution kernels with different dilation rates to obtain receptive fields of different sizes. Zhou et al. [15] proposed a semantic segmentation network D-Linknet for high-resolution satellite images. In the central area of Linknet34, a series-parallel combination of dilated convolutions was added, which solved the lack of Linknet central receptive field so that the continuity of the extracted road has been improved. Shamsolmoal et al. [19] incorporate a feature pyramid (FP)network into generative adversarial networks to minimize the difference between the source and target domains,in order to generate a remote sensing road image extraction diagram.

Although the above methods have a good effect on the fusion of remote sensing image features, the traditional convolution operation has its limitations, although it can fuse the spatial and channel information in the local receptive field, it often ignores the dependence between the various dimensions on the whole. Therefore, some soft attention mechanisms such as SENet [20], Convolutional Block Attention Module [21], and SKNet [22] are proposed to extract key features in channel domain, spatial domain, and mixed channel and spatial domain. And all of these modules almost have plug-and-play characteristics and can be embedded in any existing network to improve the speed, quality of results, and generalization capability. Shamsolmoali et al. [19] use a spatial attention module insert into the SSD layers in order to extract the generated features of the Lightweight RConv layers. Jin et al. [23] proposes an innovative Cascaded Attention DenseUNet (CADUNet) semantic segmentation model by embedding two attention modules, such as global attention and core attention modules, in the DenseUNet framework to extract road information. Wu et al. [24] added a channel attention module in the center of the network. This greatly improves the details of the extraction results. Wang et al. [25] introduce the dilated convolution attention module(DCAM) between the encoder and decoders to increase the receptive field.

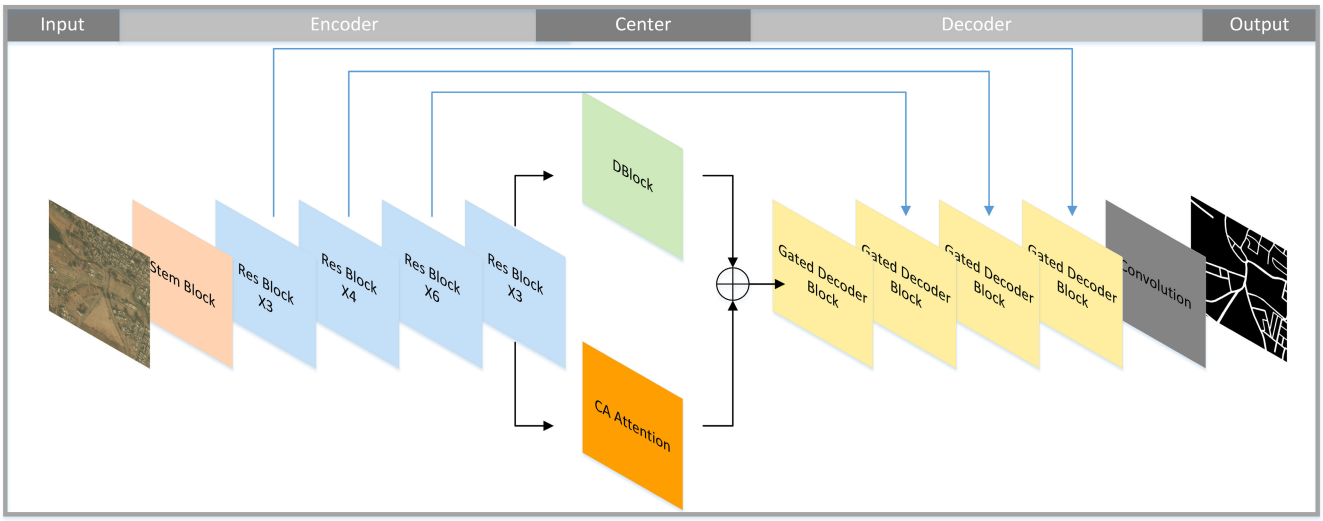

FIGURE 1. The network structure of AGD-Linknet model is mainly composed of three parts, which are encoder, central processing part and decoder.

However, the above road extraction network based on deep learning has some shortcomings in road integrity, edge straightness and road network connectivity of road information extracted in complex scenes. Therefore, under the premise of ensuring the appropriate depth of the model, a network structure that can make full use of high-level semantic information, effectively integrate channel domain and spatial domain information, minimize the loss of information and filter out effective information is needed to solve the problem of poor road extraction effect in high-resolution remote sensing images. Excellent image processing networks such as ResNet, Linknet, D-Linknet, Coordinate Attention [26], STDN, and Gated Convolution, respectively adopt residual connection, U-shaped network structure, skip connection, DBlock, attention mechanism, Stem block, gated convolution, and other operations have achieved good results in the field of image segmentation. The hybrid methods (more than one method) usually give better results and accuracy [27]. Inspired by them, this paper designs a road information semantic segmentation model which combines attention mechanism, gated convolution decoder and dilated convolution, called AGD-Linknet model, which means LinkNet With Attention model, Gated convention decoder, and Dilated Convolution.

The main contributions of this article are as follows:

1) We proposed a novel remote sensing road extraction model that obtains stronger semantic characteristics and improves the accuracy of remote sensing road extraction by integrating attention mechanisms, gated decoder block, and dilated convolution.
2) The mechanism of parallel processing of the dialated convolution and attention mechanism in the central part of AGD-Linknet encoder and decoder is very novel, combining the multi-scale features obtained by the null convolution with the features extracted by the attention mechanism in two perpendicular directions, and finally concat them to obtain a combined feature.
3) Through the analysis of the experimental comparative comparative analysis of the ablation experiments and the experimental comparative analysis in many scenarios in rural areas, cities, and suburbs, we discussed the direction of improvement and the value of proposing a new structure. Experiments show that the proposed method has distinct advantages over existing methods we have discussed on a benchmark dataset.

The rest of the paper is organized as follows. Section II describes the overall network structure of AGD-Linknet, as well as the role of each of its component modules. Section III introduces the evaluation metrics of the experimental results. In Section IV, presents the details of the experimental setup, including the data set, hardware environment, and the selection of the loss function optimizer. Experiments are presented in Section V. Finally, the conclusion is made in Section VI.

II. AGD-LINKNET NETWORK STRUCTURE

The specific structure of the high-resolution remote sensing road semantic segmentation model (AGD-Linknet) based on the attention mechanism, gated convolution decoder, and dilated convolution is shown in Figure 1. The model is an end-to-end Encoder-Decoder structure. It takes high-resolution remote sensing satellite images as input, and the output is a black and white mask image of road extraction, and roads are represented in white.

The main structure of the network is inspired by Linknet, the encoder, decoder, and its connected central part are adjusted. First, use the pre-trained ResNet-34 to replace the ResNet-18 of Linknet and replace its initial convolutional layer with the Stem block. Second, we introduce the series-parallel combination dilated convolution and add Coordinate Attention between the encoder and decoder. Finally, the

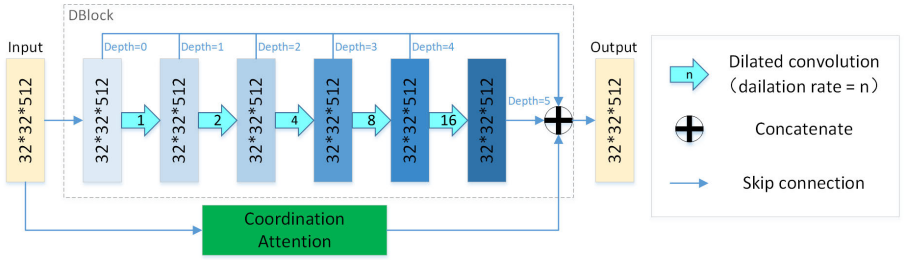

FIGURE 2. The structure of center part, which consist with DBlock and Coordinate attention block.

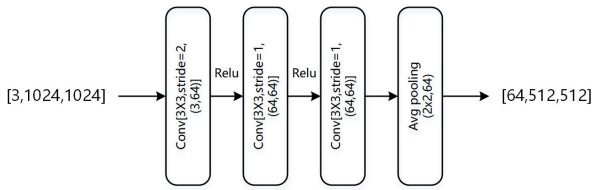

FIGURE 3. The structure of stem block.

decoder module with embedded gated decoder blocks is used for the decoder part. Finally, the classification of pixels in the image is completed, and the road segmentation map is obtained.

A. ENCODER:Resnet34 WITH STEM BLOCK

The encoder of a convolutional neural network usually starts from the initial convolutional module, which first inputs the image for convolution with a convolutional kernel of size 7×7 and a stride of 2, followed by a max pooling with a stride size of 2, and the number of output channels of the initial block is 64. Due to the two consecutive downsampling operations in the structure of the initial convolutional module, a large amount of information is lost, which makes it difficult to recover the required edge target information when the decoding stage is reduced. Therefore, Inspired by STDN replaced the initial block with the Stem module which consists of three consecutive 3×3 convolutional layers and one 2×2 average pooling layer. The stride of the first convolutional layer is 2 and the others are 1. And the output channels for all three convolutional layers are 64. Therefore, the use of the Stem block (shown in Figure 3) can reduce the information loss in the initial block, and the use of two more nonlinear activation functions strengthens the nonlinearity of the network and deepens the network structure, which can effectively network learning performance.

The subsequent encoder part uses the residual network ResNet-34. This structure can avoid the problem of model degradation due to the increase in network depth, and the ResNet-34 encoder composed of multiple residual structures can increase the complexity of the network without increasing the burden of calculation and strengthen the ability of the network model to extract semantic features. Pretrained the initialized model can greatly accelerate the convergence of the network and alleviate the optimization problems that the network may have. Therefore, the pre-trained ResNet-34 model on ImageNet [28] is used as the encoder of the model.

B. CENTER:DBlock WITH COORDINATE ATTENTION BLOCKS

The coordinate attention blocks and the dilated convolution blocks are introduced between the encoder and the decoder(the center part which is shown in Figure2). These two modules adopt a parallel structure, which combines the high receptive field capability of the dilated convolution module and the long-connected feature information extraction capability of the coordinate attention blocks. Thereby, the network's ability to recognize road information in remote sensing images is improved.

1) DBlock

Due to the problem of limited receptive field in the central region of the LinkNet network structure, it cannot capture the semantic information in a large area, especially when it encounters the information with a large spatial span like the remote sensing image ground information, which may produce the segmentation discontinuity. In order to extract more information, referring to the multi-scale fusion method in the Deeplab, a series-parallel dilated convolution blocks(DBlock) is added to the central area of LinkNet, the structure of which is shown in Figure 4. The purpose of this is to enlarge the receptive field without reducing the resolution of the feature maps and retaining spatial detail information as much as possible. At the same time, the advantages of LinkNet's skip connection and residual coding module are fully maintained.

DBlock can be divided into six branched paths with different depths and receptive field by the combination of dilation convolutions with different dilation rates. The receptive fields of the input feature map of the network are 63, 31, 15, 7, 3, 1 and the depths are 5, 4, 3, 2, 1, 0 (0 means it is the identity map). This structure allows the fusion of features of different depths and different breadths and keeps the spatial resolution of the features unchanged while augmenting the receptive field of the central part of the network, which ensures the breadth of the receptive field of the feature map and combines the multidimensional semantic information, and the feature scale is unchanged without losing the spatial relative information.

2) COORDINATE ATTENTION BLOCKS

The attention module uses Coordinate Attention module, the structure of which is shown in Figure 5. Unlike the SE module, which only considers channel information, this module combines channel information and spatial information together, which can greatly increase the scope of attention. Since road information has long connectivity, this module replaces the two-bit global average pooling operation with a pooling operation in two one-dimensional directions, so that the information in two independent directions, horizontal and vertical, can be integrated together very effectively, which is very convenient for the extraction of road information. The method also reduces the computational overhead to a great extent.

Specifically, two pairs of one-dimensional directional features are first extracted for a given input, and each channel is first encoded along two one-dimensional directions of horizontal and vertical coordinates using pooling kernels of size (H,1) or (1,W), respectively. Thus, the output of the height h at c-th channel can be formulated as

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i \le W} x_c(h, i) \tag{1}$$

Similarly, the output of the width w at c-th channel can be written as

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j \le H} x_c(j, w) \tag{2}$$

After that, we concatenate two extracted aggregated feature maps and then send them to a shared 1×1 convenlutional transformation function F, yielding

$$y = \sigma\left(F_1\left[z^h, z^w\right]\right) \tag{3}$$

where $[\cdot, \cdot]$ is the concatenation operation along the spatial dimension,$\sigma$is the non-linear activation function, and $y$is the intermediate feature mapping that encodes the spatial information in the horizontal and vertical directions. And then the number of channels at this point is controlled by 1×1 convolutional pair, and the information is reduced to the size of $C/r \times H \times W$, and then it is decomposed into $y^h \in R^{C/r \times H}$ and $y^w \in R^{C/r \times W}$ along the spatial dimension. Afterwards, the 1×1 convolutional transform is performed separately to reduce it to the same tensor with C channels to obtain.

$$g^h = \delta(F_h(y^h)), \tag{4}$$

$$g^w = \delta(F_w(y^w)) \tag{5}$$

where δ is the Sigmoid activation function, and finally the g^h and g^w are expanded and then multiplied to obtain the corresponding attention parameter Out.

$$Out = g^h(i) \times g^w(j) \tag{6}$$

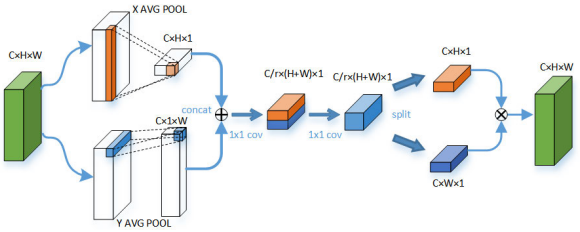

FIGURE 4. The structure of coordinate attention.

C. DECODER:GATED DECODER BLOCK

With the extraction of deep information by convolutional neural network, in order to obtain deeper information features, the convolutional neural network uses pooling and convolution, which will reduce the spatial resolution of the feature map, but the semantic segmentation task needs to restore the resolution of the image to the same as the original input image. Therefore, a decoder is needed to restore the size of the feature map.

Transposed convolution [29] is often used in decoders for upsampling, but the filled pixels produced by transposed convolutional operation often deviate from the true annotation, especially in the edge part which is more blurred. In order to improve the system to correspond to the road edge information, this paper injects gated convolution into the decoder module. Gated convolution for each channel and spatial location can learn to generate a dynamic feature selection mechanism acting on the feature map, which helps to improve the extraction of edge texture information.

First, gated convolution doubles the number of channels of the input feature map by convolution, and divides it into a learning part and a feature map. Secondly, the learning part is activated by the Sigmoid function to obtain the specific position threshold weights. Finally, the feature map with the ELU activation function is dotted with the position threshold weights generated by the learning part to obtain the result. This can be expressed as equation 7.

$$O_{x,y} = ELU(F_{x,y}) \odot Sigmoid(G_{x,y}) \tag{7}$$

where $O_{x,y}$ is the output result map, $F_{x,y}$ is the feature map, $Sigmoid(G_{x,y})$ is the gated threshold weight.

Therefore, Gated Decoder Block is designed in this paper, and the structure is shown in Figure 6. The decoder uses a 1×1 convolution layer to reduce the overall computation and improve the characterization ability of the network; subsequently, a transposed convolutional upsampling is used to recover the pixels; two consecutive gated convolution layers are used to enhance the extraction of road edge information by the network.

III. EVALUATION METRICS

The pixel accuracy (PA), average intersection ratio (mIoU) and F1-Score are used as evaluation indicators to evaluate the output results of the model to demonstrate the effectiveness of the model

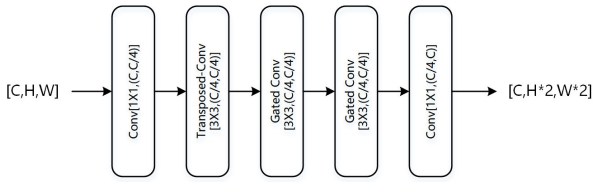

FIGURE 5. The structure of gated decoder block.

- Pixel accuracy. PA is the ratio of correctly recognized pixels to the total pixels, which can be used to illustrate the segmentation accuracy of the road category, and the larger value of PA means a better road extraction capability. This can be expressed as equation 8.

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

where TP(True Positive) is the number of pixels, which are correctly predicted as road pixels.
FP(False Positive) is the number of pixels, which are incorrectly predicted as road pixels.
FN(False Negative) is the number of non-road pixels, which are incorrectly predicted as non-road pixels.
TN(True Negative) is the number of non-road pixels, which are correctly predicted as non-road pixels.

- The Mean Intersection over Union. [30] mIoU is the average of the ratios of the intersection and union of the prediction results of all categories and the true value. The larger value of the mIoU, the better the prediction effect.

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{TP}{TP + FP + FN} \tag{9}$$

- The F1-Score metric is represented by Recall and Precision, and the larger the F1 result, the more similar the predicted image is to the real label provided.
Precision indicates the number of actual positive samples among those predicted to be positive.

$$Precisin(P) = \frac{TP}{TP + FP} \tag{10}$$

Recall indicates the proportion of samples that are actually positive that are judged to be positive.

$$Recall(R) = \frac{TP}{TP + FN} \tag{11}$$

The F1-Score indicator is the result of a combination of the above indicators. That is

$$F1 - Score = 2 \times \frac{P \times R}{P + R} \tag{12}$$

IV. EXPERIMENT

A. DATASET

The experimental dataset is derived from the DeepGlobe Road Extraction dataset, which is used to extract road information in high-resolution remote sensing images. The size of all images in the dataset is 1024 × 1024 , the images originate from Thailand, India, and Indonesia, and the image scenes include various scenes such as urban, rural, and suburbs. The dataset contains 6226 pairs of training images and image labels, which are randomly divided into test sets, validation sets and training sets according to the ratio of 1:1:8, with 624 sets of test images in the test set and validation sets, 4978 sets of training images in the training set, as the Table 1 shown.

TABLE 1. The detail of dataset.

Spatial Resolution	50 cm
Resolution	1024 × 1024
the number of validation sets	624
the number of test sets	624
the number of traing sets	4978

B. IMPLEMENTATION DETAILS

This experiment uses the PyTorch deep learning framework for Windows to build the neural network. Where the environment versions are torchvision=0.90, torch=1.8.0, and Python version 3.7. The hardware configuration of the experimental platform is Central processing Unit Intel i7-11700 and NVIDIA GeForce RTX3060 Graphical Processing Unit (with a RAM of 12GB). Due to hardware limitation, the batch size is set to 8 for each network in this experiment.

C. LOSS FUNCTION AND OPTIMIZER

In order to obtain the best model weights, this algorithm uses the binary cross entropy and the dice coefficient loss function as the loss function, and the change of the loss function can be used to determine the specific optimization direction of the model parameters.

The Binary Cross Entropy Loss. $loss_{BCE}$ is a variance calculation using cross-entropy. Due to the logarithmic function used, the loss does not increase linearly but exponentially as the penalty to the model increases. This has the advantage that the model tends to bring the predicted output closer to the true label, It is defined as

$$loss_{BCE} = -w_i \left[y_i \log x_i + (1 - y_i) log(1 - x_i) \right] \tag{13}$$

where y_i is the predicted probability of being road of pixel i, while x_i is the ground truth label.

The Dice loss.$loss_{Dice}$ is a widely used metric in the computer vision community to calculate the similarity between two images as the following equation shown

$$loss_{Dice} = \frac{2\,|X \cap Y|}{|X| + |Y|} \tag{14}$$

The loss function used in this algorithm combines the cross-entropy loss function as well as the dice coefficient loss function, with the specific loss L shown in the following equation

$$Loss = w_1 \cdot loss_{BCE} + w_2 \cdot loss_{dice} \tag{15}$$

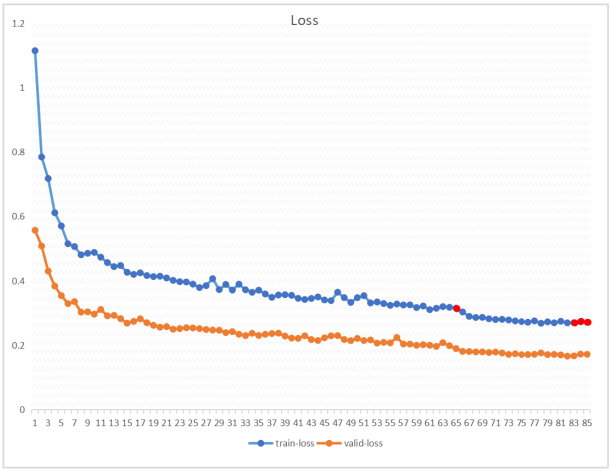

FIGURE 6. The loss curves for the train and validation set, the learning rate is updated at the red dot.

The Adam Optimizer [31] is chosen as an optimization algorithm that performs first-order gradient computation on a randomized target. It outperforms both the RMSProp optimizer and SGD optimizer, two powerful optimization algorithms, in terms of model convergence speed and convergence effect. The initial learning rate is set to 2e-4, and whenever the loss curve tends to flatten, the learning rate is reduced to one-fifth of the previous rate. When the learning rate is less than $5*10^{-7}$, the model will exit training. Figure 6 shows the loss function curves of the training and validation sets. From this figure, we can see that the training loss of the model is no longer converging at the 65th, 82nd, 83rd, and 84th epochs(as the red dot shown in Figure 6), and the learning rate is reduced at this time to make the loss value in the training process decreasing, and finally convergence is achieved at the 85th epoch, while the loss value in the validation set, the overall trend and the trend of the loss value in the training set are basically the same, indicating that the network is not overfitted. It proves the reasonableness of the optimizer, the loss function and the hyperparameters set in this paper.

V. ANALYSIS OF RESULTS AND PERFORMANCE COMPARISON

In order to verify the effectiveness and feasibility of the AGD-Linknet model, this section first analyzes the influence of each module of the AGD-Linknet model on the overall recognition effect through ablation study. Then, the AGD-Linknet model is experimentally analyzed through the overall data scenario and three common scenarios of suburbs, urban and rural areas, and compared and analyzed with Unet, Linknet and D-Linknet. Finally, the application of AGD-Linknet model in practical scenarios is analyzed.

A. ABLATION STUDY AND ANALYSIS OF AGD-LINKNET

To verify the effect of each module on the overall model, ablation study were conducted on Linknet, DBlock, attention mechanism, and Stem Block, and we chose Linknet with Resnet34 as the encoder as the baseline model and Deep-Globe Road Extraction dataset as the experimental dataset. The results are shown in Table 2 and Figure 7, and the results are compared to prove that all settings in the network are optimal.

TABLE 2. Results of ablation study in road extraction task.

	Method	PA(%)	Growing(%)
1	Linknet34	87.30%	-
2	Linknet34+DBlock(D-Linknet)	87.80%	0.5
3	Linknet34+DBlock+CA	88.39%	1.09
4	Linknet34+Stem+DBlock+CA	88.65%	1.35
5	**AGD-Linknet(Ours)**	**89.21%**	**1.81**

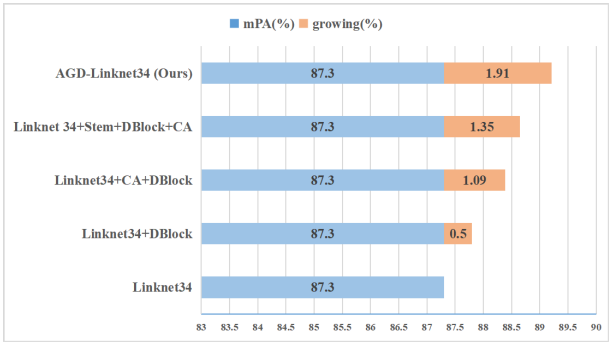

FIGURE 7. Ablation study of accuracy improvement method.

1. Adding the series-parallel combined dilated convolution (DBlock). Compared with the benchmark model Linknet34, the pixel accuracy improves by 0.5 compared to the benchmark model after adding DBlock, and the average pixel accuracy can reach 87.80%.

2. Combining coordinate attention block. After adding the series-parallel combined dilated convolution to the central region of Linknet34 and integrating it with the coordinate attention block, the pixel accuracy increases relative to the benchmark model, and the mean pixel accuracy is 88.39%.

3. Stem block. The DBlock in the central region of Linknet and the coordinate attention block are retained, and the initial 7×7 convolution is replaced by the stem block, and the experiments are conducted again. The results show that the accuracy increases by 1.35 with respect to the benchmark model, and the average pixel accuracy is 88.65%.

4. Adding the Gated Decoder. The Stem Block of the initial convolutional layer and the DBlock of the center region, as well as the attention mechanism, are retained, and the decoder is changed to incorporate the Gated Decoder, and the results show that the accuracy increases by 1.81 relative to the benchmark model, and the average pixel accuracy is 89.21%.

B. ROAD EXTRACTION EXPERIMENT IN SUBURBS SCENES

In this section, the suburbs image input network in the Deep Globe Road Extraction dataset is selected for comparison.

TABLE 3. Segmentation results of urban, suburbs and rural on Deep Globe Road Extraction dataset.

Evaluation metrics	Situation	Unet	Linknet	D-Linknet	AGD-Linknet
PA	Urban	71.25%	87.63%	88.17%	**88.98%**
	Suburbs	77.27%	86.89%	87.42%	**87.51%**
	Rural	81.73%	88.51%	88.84%	**90.19%**
mIoU	Urban	0.6115	0.7911	0.8023	**0.8056**
	Suburbs	0.6719	0.7969	**0.8081**	0.8071
	Rural	0.7088	0.8246	0.8279	**0.8397**
F1-score	Urban	0.7135	0.8725	0.8803	**0.8827**
	Suburbs	0.7532	0.8669	**0.8775**	0.8757
	Rural	0.7879	0.8909	0.8879	**0.9014**

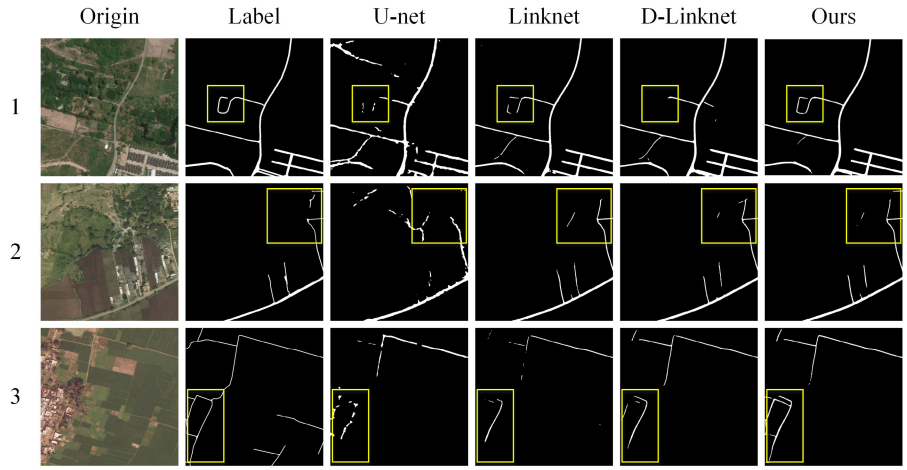

FIGURE 8. Experimental results of road extraction in suburbs scenes.

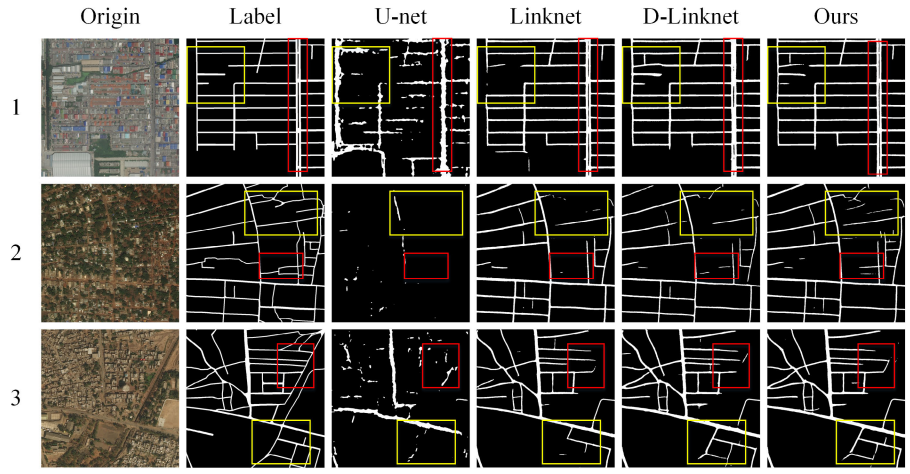

FIGURE 9. Experimental results of road extraction in urban scenes.

The results can be obtained from the Table3. It can be seen that in the suburbs scenario, D-Linknet is slightly better than AGD-Linknet in mIoU and F1-Score, but AGD-Linknet is higher in PA. Overall, the difference between AGD-Linknet and D-Linknet in the suburbs scenario is not much, at the range within ± 0.002.

A specific analysis of the suburbs image is carried out, as shown in Figure 8. Since there are fewer houses in the suburbs, the light and shadow information is relatively stable, and the road information is dominated by the main road, which is relatively clear and the road can be easy to identify. Therefore, it can be seen that the Unet network has poor connectivity and more edge burrs, However, the extraction effects of models other than the Unet network are equivalent. However, if we look closely at the detailed information in the yellow box, we can see that AGD-Linknet has better

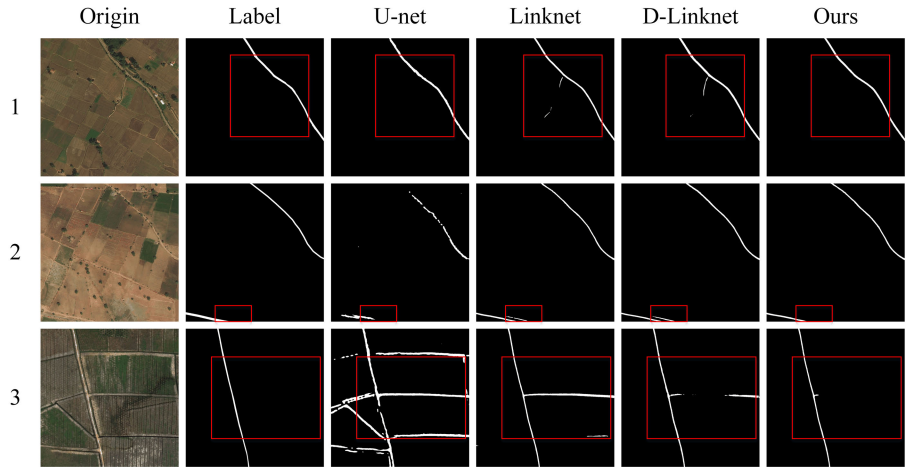

FIGURE 10. Experimental results of road extraction in rural scenes.

extraction effect on the details due to the addition of the Stem module and attention module.

C. ROAD EXTRACTION EXPERIMENT IN URBAN SCENES

In this section, the image of urban scene input network in the Deep Globe Road Extraction dataset is selected for comparison. The results can be obtained from Table 3. The performance of AGD-Linknet in cities with more complex ambient light and shadow information is better than Unet, Linknet and D-Linknet in all three evaluation indicators. Compared with D-Linknet in the urban scene, PA increased by 0.81%, mIou increased by 0.0033, and F1-Score increased by 0.0024.

Urban road images are the most complex(as shown in Figure 9), composed of various buildings, small areas of greenery, and artificial lakes. The roads are crisscrossed, and the road conditions are also more complex. Shadows and green occlusions have a greater impact on the road extraction effect, and extraction requirements of the edge information are higher. And when the buildings and roads are more intricate urban roads are extracted, it can be seen that the connectivity of the road images extracted by the U-net due to loss of the residual module and the extraction accuracy are not as good as other networks. When observing the internal details of the yellow box, and comparing Linknet, Dlinknet and AGD-Linknet, it can be found that the latter two have added the Dblock to increase the receptive field of the central part of the network, so that the network can identify and extract more subtle roads. And it works better. Looking at the internal details of the red box, it can be found that AGD-Linknet better integrates the information of the space and channel domain due to the addition of the coordinate attention block, so it has a stronger extraction ability on roads blocked by greenery and shadows of tall buildings and roads in communities with similar colors. It can also better notice the areas that the other three networks have not paid attention to so that the extraction results have stronger connectivity and clearer edge information. Looking at the details inside the blue box, we can see that the AGD-Linknet with the addition of the gated convolutional decoder module has clearer processing of the edge information in the middle of the two-way lanes of the overpass.

D. ROAD EXTRACTION EXPERIMENT IN RURAL SCENES

In this section, the suburbs image input network in the Deep Globe Road Extraction dataset is selected for comparison. The results can be obtained from Table3. Compared with D-Linknet in the rural scene, PA increased by 1.35%, mIou increased by 0.0077, and F1-Score increased by 0.0135.

As for the images of rural areas, as shown in Figure 10, the remote sensing images of rural areas often only consist of some main roads and most fields. Due to the influence of various information such as light and shadow, camera parameters and environmental factors, the dirt roads between the farmland and the main roads in some of the images show a very high similarity in the remote sensing images and are very easy to be confused, and the dirt roads between rural fields are very easy to be recognized as main roads by the network model, and our network also has better anti-interference ability for the dirt roads.

E. ROAD EXTRACTION EXPERIMENT IN OVERALL SCENES

The test set images of the Deep Globe Road Extraction dataset in this section include urban, rural, suburbs, seaside, tropical rainforest and other scenes. Input this test set image into Unet, Linknet34, DLinknet34, and the AGD-Linknet of this article, and compare them pixel accuracy, mean intersection ratio and F1-Score compare their performance.

As shown in Table 4, compared with the U-net and Linknet models, D-Linknet can improve the performance of the network by adding the series-parallel combined dilated convolution in the middle which can effectively enhance the range of the receptive field and concentrating the global information. The AGD-Linknet network, due to the

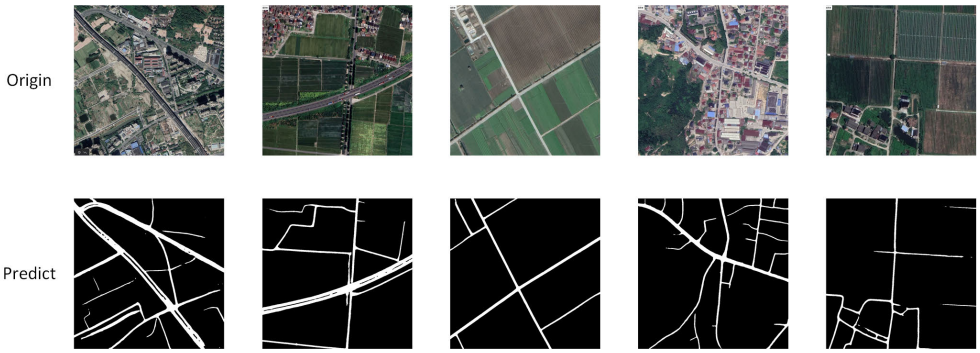

FIGURE 11. The extraction effect of actual scene.

TABLE 4. Comparative Analysis between AGD-Linknet and Traditional Algorithm.

Model	PA	mIoU	F1-Score
Unet	77.69%	0.6729	0.7572
Linknet34	87.30%	0.8040	0.8758
D-Linknet34	87.80%	0.8125	0.8807
AGD-Linknet(Ours)	**89.21%**	**0.8202**	**0.8864**

combination of the Stem module, the DBlock module and the Coordinate Attention block in the central part, and the gated convolutional decoder module, improves the PA by 1.41%-11.52%, mIoU by 0.0077-0.1473, and the F1-Score by 0.0057-0.1292. In all three evaluation metrics, it is higher than the other three classical algorithms discussed.

F. APPLICATION OF AGD-LINKNET MODEL IN ACTUAL SCENES

This section intercepts some remote sensing images of the actual scenes in Google Earth from the network, and inputs them into the trained network, the results are shown in the Figure 11. It can be seen that the network proposed in this paper not only has a good extraction effect in the training Deep Globe Road Extraction data set, but also has a good extraction effect in other remote sensing image road information, which proves the universality of the network proposed in this paper.

VI. CONCLUSION

In this paper, the AGD-Linknet network based on attention mechanisms, gated convolution, and dilated convolution is proposed to extract road information from remote sensing images. The network takes the Linknet34 structure as the basic framework, and the stem module is introduced to reduce the information loss in the convolution stage;the series-parallel combined dilated convolution is used to adjust receptive fields of feature points and the Coordinate Attention is used to enhance the weight extraction of the target region by combining the channel and the spatial information; finally, the Gated Decoder Block is embedded in the decoder, which can effectively discriminate the edge information. The AGD-Linknet network is tested by using the Deep Globe Road Extraction dataset, and the ablation experiments on the proposed structure demonstrate the effectiveness and feasibility of the improved network structure. The AGD-Linknet network in rural areas, suburbs, urban roads and the overall scene were analyzed, and it was found that the extraction effect of AGD-Linknet was good, the anti-interference ability was strong, and the extraction ability of small branches was strong. Compared with other classical algorithms such as U-Net, Linknet and D-Linknet on the DeepGlobe dataset, our proposal has improved the pixel accuracy, mean intersection over union and F1-Score index of road recognition by 1.41%-11.52%, 0.0077-0.1473, 0.0057-0.1292. And the road integrity, edge straightness and road network connectivity are also improved.

In the future, we will continue to study the expansion and optimization of road extraction data sets and the neural network architecture of road extraction. An image enhancement algorithm based on an unsupervised network is adopted to enhance the data set so as to improve the definition of the remote sensing image; We will continue to optimize the network structure, such as changing the encoder part to ResNext [32] or introducing Transformer structure [33] to improve the accuracy of semantic segmentation, in order to extract more clear road information.

REFERENCES

[1] X. Haodong, "Review and prospect of road feature extraction from high resolution remote sensing images," *Geomatics Spatial Inf. Technol.*, vol. 36, no. 8, pp. 202–206, 2013.
[2] L. Chen, J. Jia, and H. Wang, "An overview of applying high resolution remote sensing to natural resources survey," *Remote Sens. Natural Resour.*, vol. 31, no. 1, pp. 1–7, 2019.
[3] J. Li, K. Feng, L. Zhu, and H. Yun, "Road extraction from remote sensing image by watershed algorithm based on threshold marker," *J. Changchun Univ.*, vol. 29, no. 6, pp. 10–14, 2019.
[4] H. Mu, Y. Zhang, H. Li, Y. Guo, and Y. Zhuang, "Road extraction base on Zernike algorithm on SAR image," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 1274–1277.
[5] X. Lin, J. Zhang, Z. Liu, J. Shen, and M. Duan, "Semi-automatic extraction of road networks by least squares interlaced template matching in urban areas," *Int. J. Remote Sens.*, vol. 32, no. 17, pp. 4943–4959, 2011.
[6] H. Y. G. Wang, "Road extration of high-resilution remote sensing image," *J. Geomatics*, vol. 45, no. 3, pp. 34–38, 2020.
[7] H. Ma, X. Cheng, X. Wang, and J. Yuan, "Road information extraction from high resolution remote sensing images based on threshold segmentation and mathematical morphology," in *Proc. 6th Int. Congr. Image Signal Process. (CISP)*, vol. 2, Dec. 2013, pp. 626–630.

[8] F. Zeng, B. Yang, D. Wu, P. Tang, and H. Zhang, "Extraction of roads in mining area based on Canny edge detection operator," *Remote Sens. Land. Res.*, vol. 25, no. 4, pp. 72–78, Nov. 2013.
[9] B. Sirmacek and C. Unsalan, "Road network extraction using edge detection and spatial voting," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3113–3116.
[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[12] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.
[13] P. Zhou, B. Ni, C. Geng, J. Hu, and Y. Xu, "Scale-transferrable object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 528–537.
[14] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[15] 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/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 182–186.
[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[17] J. Xin, X. Zhang, Z. Zhang, and W. Fang, "Road extraction of high-resolution remote sensing images derived from DenseUNet," *Remote Sens.*, vol. 11, no. 21, p. 2499, 2019.
[18] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.
[19] P. Shamsolmoali, M. Zareapoor, H. Zhou, R. Wang, and J. Yang, "Road segmentation for remote sensing images using adversarial spatial pyramid networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 4673–4688, Jun. 2020.
[20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[21] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[22] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.
[23] J. Li, Y. Liu, Y. Zhang, and Y. Zhang, "Cascaded attention DenseUNet (CADUNet) for road extraction from very-high-resolution images," *ISPRS Int. J. Geo-Inf.*, vol. 10, no. 5, p. 329, May 2021.
[24] M. Wu, C. Zhang, J. Liu, L. Zhou, and X. Li, "Towards accurate high resolution satellite image semantic segmentation," *IEEE Access*, vol. 7, pp. 55609–55619, 2019.
[25] Y. Wang, Y. Peng, W. Li, G. C. Alexandropoulos, J. Yu, D. Ge, and W. Xiang, "DDU-Net: Dual-decoder-U-Net for road extraction using high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4412612.
[26] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13713–13722.
[27] N. Khan, I. Khaleel, and E. Daghighi, "Improved feature selection method for features reduction in intrusion detection systems," *Mesopotamian J. Cybersecur.*, vol. 2021, pp. 9–15, Jan. 2021.
[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[29] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2018–2025.
[30] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," 2017, *arXiv:1704.06857*.
[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[32] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.
[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.