Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH

Open Access

DU-Net-Cloud: a smart cloud-edge application with an attention mechanism and U-Net for remote sensing images and processing

Jiayuan Kong* and Yanjun Zhang

Abstract

In recent ages, the use of deep learning approaches to extract ground object information from remote sensing high-resolution images has attracted extensive attention in many fields. Nevertheless, due to the high similarity of features between roads, prevailing deep learning semantic segmentation networks commonly demonstrate reduced continuity in road segmentation. Besides this, the role of advanced computing technologies including cloud and edge infrastructures has also become very important due to large number of images and their storage requirements. In order to better study the road details in images related to remote sensing, this paper suggests a road extraction technique which is basically founded on Dimensional U-Net (DU-Net) network. At the deepening level of the U-Net network, a parallel attention mechanism, known as ProCBAM, is added and implemented to the feature transmission step of the classical U-Net network. Moreover, we use and implement the edge-cloud architecture to develop and construct a unique remote sensing image service system that integrates several datacenters and their related edge infrastructure. In the proposed system, the edge network is primarily used for caching and distributing the processed remote sensing images, while the remote datacenter serves as the cloud platform and is responsible for the storage and processing of original remote sensing images. The results show that the proposed cloud enabled DU-Net model has achieved good performance in road segmentation. We observed that it can achieve improved road segmentation and resolve the issue of reduced continuity of road segmentation when compared with other state-of-the-art learning networks. Moreover, our empirical evaluations suggest that the proposed system not only distributes the workload of processing tasks across the edges but also achieves data efficiency among them, which enhances image processing efficiency and reduces data transmission costs.

Keywords Deep learning, Road extraction, Cloud, Edge intelligence, Remote sensing image, U-Net, Attention mechanism

Introduction

In modern ages, with the fast and quick growth of remote sensing satellite enterprise in our country, the access to road information from remote sensing image is increasingly simple and quick. There are several uses for extracting road network data from remote sensing photos, including navigation, mapping, urban planning, and updating geographic information systems. Accurate road extraction is one of the essential technologies

*Correspondence:
Jiayuan Kong
bqt2200204045@student.cumtb.edu.cn
College of Geoscience and Surveying Engineering, China University of Mining and Technology-Beijing, Beijing, China

in the study sectors of natural disaster warning, military strike, unmanned vehicle route planning, and so forth. It is not only connected to the extraction of cars, buildings, and other ground objects. However, remote sensing picture road segmentation has its own peculiarities and challenges as compared to the general segmentation work [1–3]. After centuries of research, numerous academics have proposed and developed a multiplicity of road extraction approaches. For example, Luo et al. [4] proposed the technique of conjoining spectral features and shape topographies to extract certain characteristics of roads, and Lin et al. [5] used Angle texture features and gray least-squares matching for optimal quantitative extraction of banded roads under shadow. Chen et al. [6] proposed the method of combining Markov random field segmentation with mathematical morphology processing, and Cao et al. [7] proposed the technique of extraction for road centerline from remote sensing high-resolution images by fusing multi-scale object-level features of pixels. However, according to the experimental results, these approaches are appropriate for remote sensing images with rich information and distinct features, while for similar ground objects, they are easy to mix and produce adhesion phenomenon.

Research in computer vision is quickly advancing deep learning technologies. Road extraction issues can be better resolved by deep learning approaches, which automatically acquire the nonlinear and hierarchical properties of pictures [8]. Many scholars have made many improvements to deep learning methods [9–11], so as to progress the recognition correctness of remote sensing image roads. In 2015, Ronneberger et al. [12] suggested an enhanced U-Net grounded on the FCN [12], which achieves multi-scale image information fusion due to its encoder-decoder network structure, and takes into account low-level details while retaining high-level semantic information. Due to its strong transformation and fast training speed, at present, it is commonly used in the field of image analysis and segmentation. Yuan et al. [13] suggested a new form of loss function to effectively improve the accuracy of road segmentation. Jin et al. [14] achieved good results and outcomes by using dual U-Net network joint training and morphological post-processing. Wang et al. [15] solved the problem of overfitting well by using Batch Normalization, ELU, and Dropout in U-Net network.

In terms of road extraction. Zhong et al. [16] applied FCN network structure to extract buildings and roads from high-resolution images, comprehensively considered the influence of learning rate, input image size and other super parameters on extraction results, and determined the optimal configuration of super parameters, which significantly improved the extraction accuracy. Although, the method based on FCN achieves good results in road extraction, it is difficult to restore the output result to the resolution of the input image due to the loss of partial spatial information caused by continuous down sampling operation in FCN. To solve this problem, researchers proposed a codec structural model, which gradually recovered the target details and corresponding spatial dimensions by connecting multilevel features in the decoder part by jumping connection. Cheng et al. [17] proposed CasNet, a cascaded codec network, for the extraction of road and center line in remote sensing images. Inspired by deep residual learning and U-Net, Zhang et al. [18] proposed deep residual U-Net network for road extraction, which simplified deep network training and reduced the number of parameters through the residual structure. At the same time, a large number of jump connections in the network, in fact, promote the dissemination of information and achieve better road extraction effects and outcomes.

Although, the network based on codec structure and void convolution can obtain global context information by extracting multi-scale features and improve the accuracy of road extraction, there are still some problems. Firstly, current methods based on codec structure still have deficiencies in extracting target features, leading to incomplete consideration of context information. Secondly, the method based on void convolution also has limitations. A single convolutional layer can only extract features from some regions according to the size of its convolution kernel, and it is easy to lose small-scale targets. The classical U-Net network has produced some results, using these techniques for road segmentation frequently produces unsatisfactory outcomes. However, due to large amount of images dataset, the role of advanced computing technologies including cloud and edge infrastructures has also become very important due to large number of images and their storage requirements. We therefore suggest a DU-Net, integrated with a cloud-edge platform, in this study. The suggested method may ensure that roads are connected with each other's and can more clearly extract roads. The following are some of the fundamental and core contributions of the research conducted in this paper:

1. In order to extract road features from the row direction and column direction of the image, a new codec network and a new U-Net model are cited.
2. A new parallel attention mechanism is created in order to efficiently merge the category information from the feature map of high-level images and the location information of the road from the feature map of low-level images.

3. We use the edge-cloud architecture to develop and construct a unique remote sensing image service system that integrates several datacenters and their related edge infrastructure. The edge network is primarily used for caching and distributing the processed remote sensing images, while the remote datacenter serves as the cloud platform and is responsible for the storage and processing of original remote sensing images.

The rest of this article is structured in the following manner. In Materials and methods section, we offer a review of different materials and methods including the structure of the U-Net and various attention modules. We discuss the proposed model of the DU-Net-Cloud model in The ProCBAM attention mechanism structure section. In Results and discussion section, experimental setup along with the studied datasets are explained. The obtained outcomes are deliberated in Experiments and results section. Finally, we conclude the paper in Conclusions and future work section and offer some future directions for further research and investigation.

Materials and methods

In this section, first we discuss the basic structure of the U-Net model and offer a review of various attention modules. We also deliberate the proposed model over an edge-cloud framework to improve the processing speed.

The U-Net structure

In May 2015, Ronneberger et al. [12] developed the structure of the U-Net approach, and it was quickly and extensively implemented in various research fields including medical image analysis and segmentation. In addition, the U-Net approach has been extensively applied by numerous researchers in the research arena of image segmentation related to remote sensing and produced positive results as a fairly traditional fully convolutional network model. Figure 1 represents the U-Net network architecture and basic structure. The left and right sections make up the U-Net structure as shown Fig. 1.

The feature extraction, sometimes referred to as the down sampling, is located on the left, and the up sampling is located on the right. In the feature extraction phase, pooling and convolution calculations are, in fact, implemented to extract the deep semantic features and characteristics of the image related to remote sensing. Following two convolutions, the image is converted into a matrix with 64 additional channels, and the maximum pooling procedure is then used to cut the image's length and breadth to half of what they were initially. The final feature map is created after two 3×3 convolution processes, and the image becomes a 32x32x512 matrix after four rounds of down sampling, per the same procedure. The calculation for the up sampling portion begins with the network's base information. The down sampled feature map of the identical layer is spliced with every 2×2

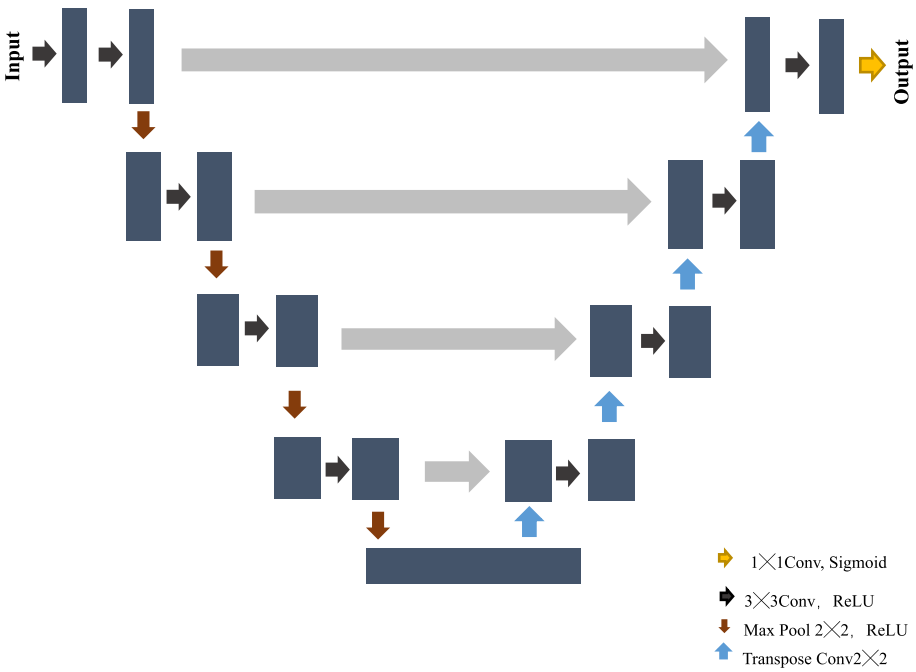

Fig. 1 The U-Net network architecture

deconvolution, and the number of channels conforming to the component of feature and characteristic extraction is fused at the equivalent scale. A 3×3 convolution operation is followed by two up samplings. Information is added to the feature map created by feature extraction in order to improve the segmentation outcome. The description of various notations and symbols used in different images are as given in the following Table 1.

The Convolutional Attention Module (CBAM)

The Convolutional Block Attention Module i.e. the CBAM (Convolutional Block Attention Module), which in fact combines the lightweight and trivial attention mechanisms of two approaches i.e. (i) channel Attention Module, and (ii) spatial Attention Module, was proposed by Sanghyun Woo et al. in 2018 [19]. It is a straightforward and efficient way to decrease the amount and quantity of various parameters and increase the method's computational effectiveness and efficiency. Moreover, it may be used to meritoriously modify the heaviness and weight of the feature map and gain additional useful feature information by combining it with any network model and incorporating it into the already-existing network architecture. Figure 2 displays the CBAM structural diagram. The Channel Attention Module (CAM) and the Spartial Attention Module (SAM), which are in fact the two separate sub-modules of the CBAM architecture, are implemented to handle channel and spatial attention, respectively. In fact, this can be added in the form of a plug-and-play module to the current network architecture while conserving parameters and processing power.

Table 1 List of notations and symbols used in the paper and images

Parameters	Description
r	reduction rate
F	Feature graph
N	Amount of channels
W_0, W_1	Layers of the MLP
Mc	Sigmoid function
V	Convolution kernel
U	Output vector of the convolution kernel
*	Convolution operation
S	Nonlinear interactions amongst the channels
Zc	Compressed feature graph
conv	Convolution layer
⊗	An activation function for convolution
⊗	Attention network
∫	Sigmoid function

The Channel Attention Module (CAM)

A channel attention diagram of the CAM is shown in Fig. 3 below. The maximum pooling (global) and the average pooling (global) are first carried out, using the feature graph, denoted by F, of the form $H \times W \times C$ as its input. Then, the results of the calculations are processed over the communal two-layer neural network, which is represented by MLP, (that comprises C/r neurons in the primary layer and relu neurons for activation, C neurons in the second layer, and measurements of $1 \times 1 \times N$ (amount of channels) to produce the supreme or maximum pooling features, characterized by F_{max} and average pooling features, characterized by F_{avg}). The weight coefficient that is given by Mc is then calculated using the activation function (Sigmoid) using the two features that were acquired after adding them together. The new characteristic which is denoted by F′ is then entered into the subsequent module of spatial attention by multiplying the weight coefficient with the original characteristic

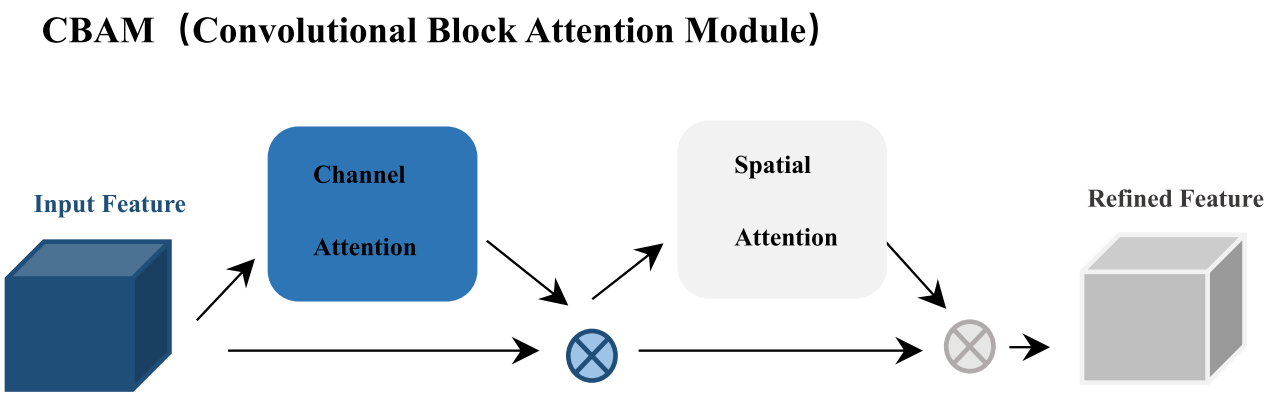

Fig. 2 Schematic diagram of the CBAM structure

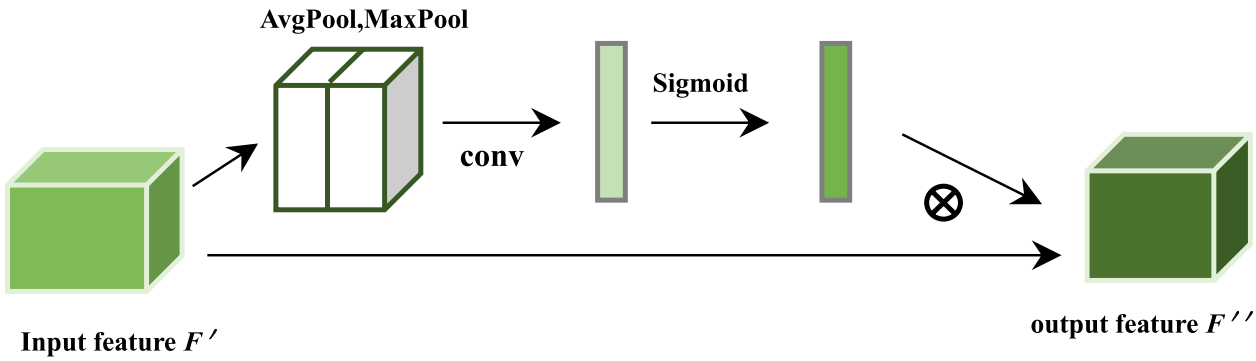

Fig. 3 Structure diagram of channel attention module [the circle shows the attention network]

given by F and applying various weights to the characteristics and feature map related to images of every channel. The way it works is as follows in Eq. 1:

and the second is combined array operation. After the convolution, the number of dimensions is decreased to one channel, and the well-known sigmoid function is

$$M_C(F) = \sigma\left(MLP\left(AvgPool(F)\right) + MLP(MaxPool(F))\right) = \sigma\left(W_1\left(W_0\left(F^c_{avg}\right)\right) + W_1\left(W_0\left(F^c_{max}\right)\right)\right) \quad W_0 \in R^{\frac{C}{r}\times C}\ W_1 \in R^{C\times\frac{C}{r}} \tag{1}$$

Where r stands for the reduction rate, *Mc* denotes the sigmoid operation, and W_0 must be triggered by relu. The MLP is made up of two layers, the first being W_0 and the second being W_1.

implemented to yield the spatial attention features and characteristics. The feature graph denoted by F″ with various spatial weights is output after the created features have been multiplied by the module's input features.

$$M_s(F) = \sigma\left(f^{7\times7}\left([AvgPool(F); MaxPool(F)]\right)\right) = \sigma\left(f^{7\times7}\left(\left[F^s_{avg}; F^s_{max}\right]\right)\right) \tag{2}$$

The Space Attention Module (SAM)

Due to the fact that as an input feature graph for this module, the extensively used CAM i.e. channel attention module's characteristic graph, which is characterized by F′ with various channel heaviness, is also used for the SAM. The first steps are: (i) maximum pooling (global), and (ii) average pooling (global) grounded on channel dimension;

In the above equation, F denotes the feature graph. This should be noted that a convolution kernel of dimension 7 × 7 is used for feature map of the spatial attention, where is a sigmoid operation, to advance and encode the areas that essentially require to be stimulated or repressed in the spatial dimension. The schematic diagram of the spatial attention module (SAM) and its operation process is given away in Fig. 4.

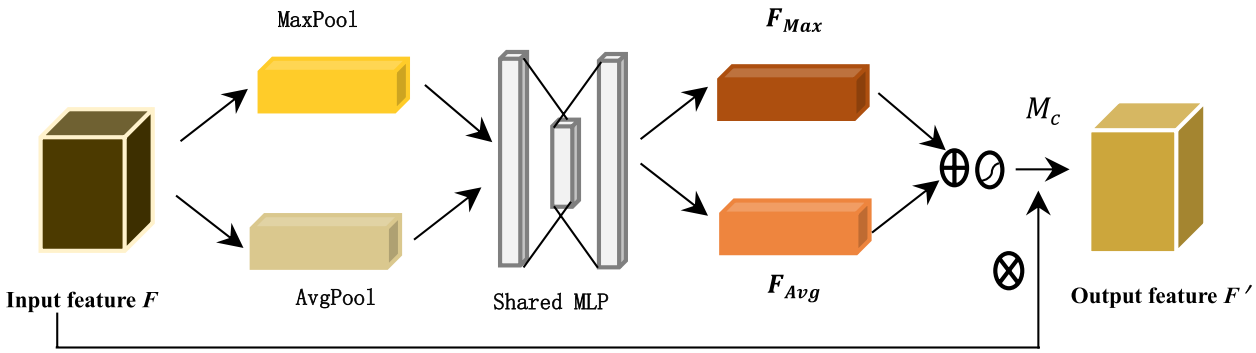

Fig. 4 The basic structure of the spatial attention module

The Squeeze-and-Excitation (SE) module

Despite the other fact, as discussed above, the Squeeze and Excitation are the major components of the SE module, which may be used with any mapping [20]. As an illustration, consider the convolution in the subsequent sentences. The convolution kernel is characterized by a vector $V = [v_1, v_2, \ldots, v_c]$, wherever the notation v_c stands for a particular convolution kernel c. The output is characterized by $U = [u_1, u_2, \ldots, u_c]$ where the each notation u_c is given by Eq. 3:

$$u_c = v_c * X = \sum_{s=1}^{C'} v_c^s * x^s \tag{3}$$

Where the 2-D convolution kernel of a s channel is represented by $\boldsymbol{v_c^s}$ and * is the convolution operation. A channel will acquire knowledge of the eigenspace associations when the spatial characteristics on the channel are input. In a similar manner, the channel feature and other characteristic relationship and its association with the spatial relationship, in fact, discovered by the convolution kernel will be combined since the convolution results of each channel are totaled. In order for the model to straightly study and acquire knowledge of the relationship between the channel feature and characteristic, the module of the SE is made to extract this crossbreed correlation and relationship.

Squeeze operation

Since convolution can only be performed locally, this is in fact a challenging task for the U and U-Net models to gather sufficient data to determine the suitable and most appropriate link amongst numerous channels. In addition, this problem is especially more severe, in particular, for trivial and shallow networks. This should be noted that the Squeeze operation uses the average pooling (global) in order to encrypt the entire spatial characteristic and feature onto the identical channel as a single feature (global). The feature graph is compressed into a $1 \times 1 \times C$ vector after compression which is mathematically expressed in Eq. 4.

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j), z \in R^C \tag{4}$$

Where z_c is a compressed feature graph which is also expressed as F_{sq}.

Excitation

The global description characteristics are retrieved after the Sequeeze procedure. The sigmoid gating mechanism is used to understand, in a superior and approved way, the nonlinear interaction amongst numerous channels. The nonlinear interactions amongst the channels is computed using the following Eq. 5.

$$S = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 ReLU(W_1 z)) \tag{5}$$

In the above equation, the notation $W_1 \in R^{\frac{C}{r} \times C}$, $W_2 \in R^{C \times \frac{C}{r}}$, characterizes the interaction. The study uses a block configuration with two (2) complete and full connection layers, where the primary full connection layer is implemented to minimize dimensivity and its preservation figure is represented by r, in an effort to simplify the model and increase generalizability. The original dimensivity will then be restored with relu activation by the whole connection layer. As illustrated in Eq. (6), the original features on U are multiplied by the activation values of individually every channel learnt (through the sigmoid activation function, value 0-1):

$$\widetilde{xc} = Fscale(u_c, s_c) = s_c \cdot u_c \tag{6}$$

The entire process may be thought of as learning the weight coefficient of each and every channel, giving the model a better chance to recognize the features of each channel and mimic the attention mechanism.

The ProCBAM attention mechanism structure

Kong et al. [21] and Zhang et al. [22] proposed the parallel attention mechanism module, also known as ProCBAM, in the remote sensing scene classification. Figure 5 depicts the structural movement and basic configuration of the ProCBAM attention technique that was suggested

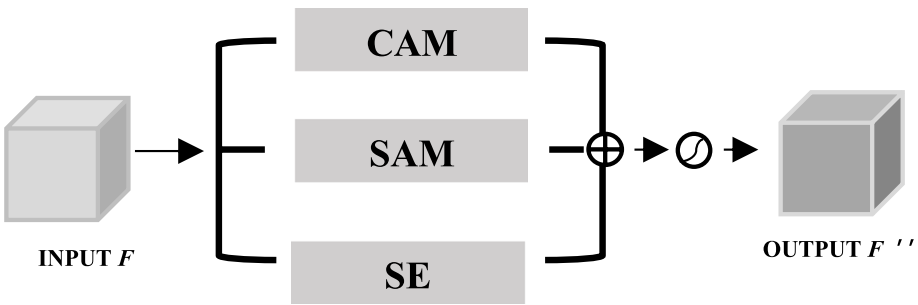

Fig. 5 The ProCBAM structure diagram

in this work. The attention mechanism and its numerous modules including the CAM (i.e. channel attention module), the SAM (i.e. spatial attention module), and the SE module are all altered into parallel structures at the same time. The computation of the later spatial attention module (SAM) will interfere with the computation and estimation of the earlier channel attention module (CAM) when both modules i.e. (i) the channel attention module (CAM), and (ii) the spatial attention module (SAM), embrace an enlightened connection. In order to better mix the two functions in the computation, the two modules are therefore merged into two divisions and the productivity properties of the two (2) modules are combined after that. In a similar manner, the SE module is developed to reform the features and characteristics of the channel and obtain the acceptable feature map after the amalgamation since the approach of adding pixels one at a time for the period of the amalgamation disturbs the association amongst numerous channels.

Network structure of the DU-Net

The following Fig. 6 shows the Dimension U-Net (DU-Net), its architecture, and its design in more detail. The ProCBAM's structure is developed from the U-Net; and it uses a multi-level subsampling structure module that, in fact, deepens the prototype to upsurge nonlinear mapping and improve fitting of the feature map. Moreover, the experimental topology has a lower (i.e. by a 7-layer) sampling module with a 3×3 ReLU convolutional layer and a 2×2 Maxpooling layer in each module. Subsequently, under a particular sampling structure and exactly after 7 times of the pooling at the coding end, the image for the input data with the magnitude and dimensions of 1024×1024 is retrieved with the dimension of 8×8, and the magnitude is approximately 1/64 of the novel and original image. Furthermore, several 7 feature graph modules with numerous levels are acquired contemporaneously. This should be noted that at-least 7 phases of up-sampling computation and estimation were also accomplished in the up-sampling section and module. To add statistics and further improve the segmentation contour texture characteristics, every layer was mixed with the characteristic map of the down-sampling coding component. In this way, the characteristics and feature image that was transmitted by the coding component is simultaneously enhanced by the multi-dimensional supervised computation in order to demonstrate and show the geometric topographies and overwhelm the contextual and related features and characteristics. This is in fact done and completed through adding a DU-Net module to the transfer section of feature image related to the coding portion of every layer.

The contour restoration of the segmentation method benefits from the shallow layer features' adequate texture characteristics for remote sensing picture features. Target categories can be distinguished using high-level semantic characteristics. In order to fully complete the

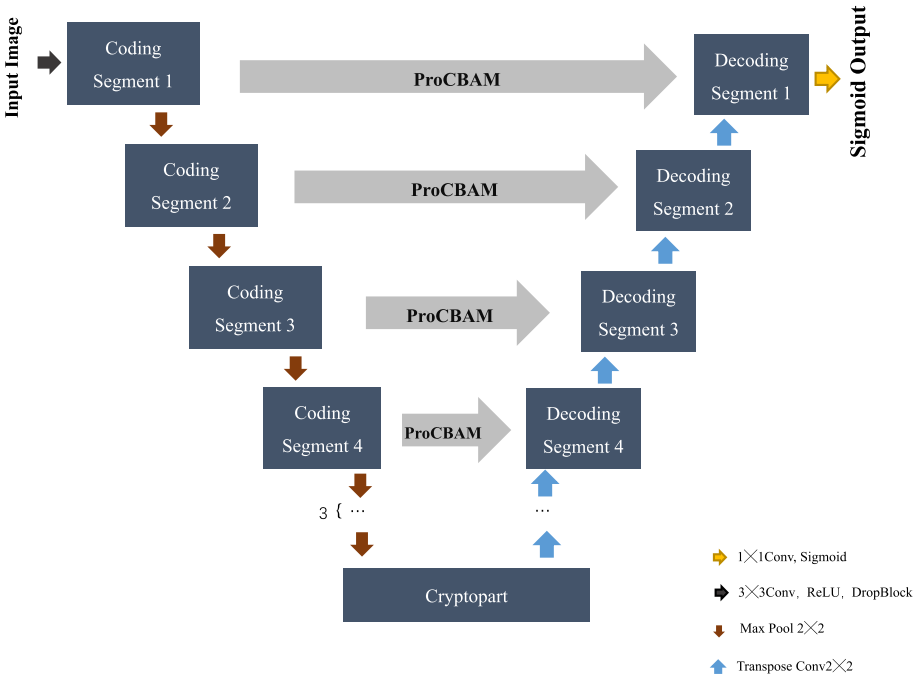

Fig. 6 The proposed DU-Net network architecture

feature information, it is important to merge the features of the two. The multi-level coding and decoding structure is used in the DU-Net network architecture and design that is anticipated in this work, and the characteristic or feature graphs of various stages in the coding portion are fully utilized to construct a deeper and improved learning network. This should be kept in mind that the characteristics and numerous features are take-out and shared in a precise manner in order to improve prediction effect.

The edge cloud architecture

The edge cloud platform is used to implement various phases of the smart application as shown in Fig. 7. In fact, the role of advanced computing technologies including cloud and edge infrastructures has also become very important due to large number of images and their storage requirements. Therefore, the remote sensing application is divided into different module in such a way that all modules communicate with each other, but, run on different locations transparently. The edge platform is responsible to preprocess the collected data and then send it for storage and training purposes to the remote cloud. Figure 8 discuss how the two important modules of the smart application i.e. channel attention module and the space attention module are implemented over the edge-cloud infrastructure so

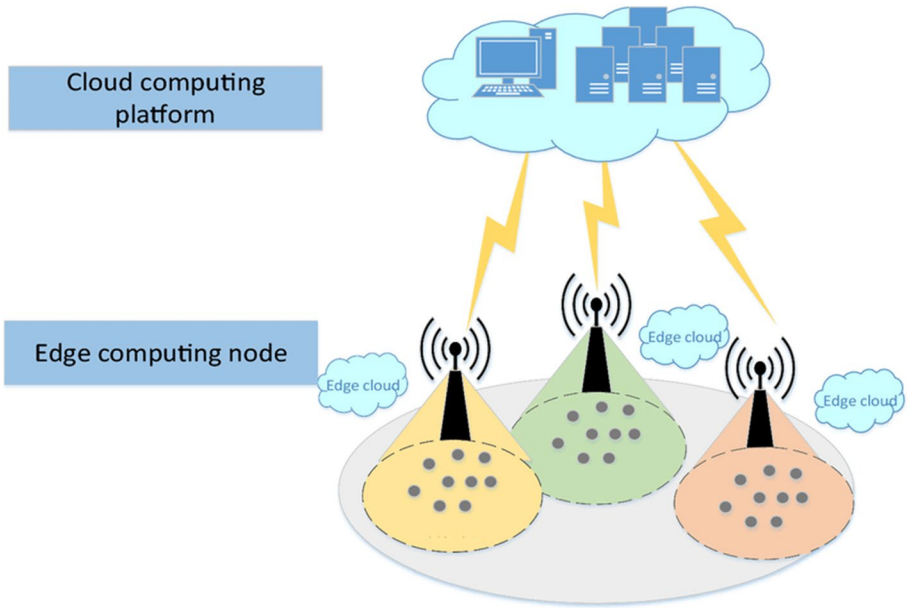

Fig. 7 The DU-Net network implementation over the edge cloud platform

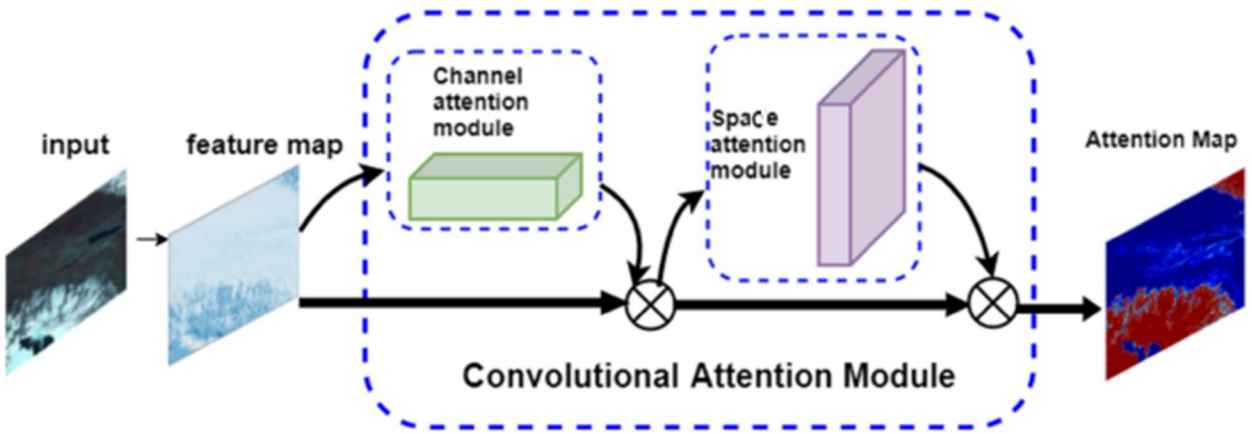

Fig. 8 Different modules and their mapping to edge cloud

that the data can be processed locally (to where they are produced or captured) and only essential data is used for training purposes. This can help in reducing the training time and the application latency. The prediction happens at the edge device while the training occurs at the cloud in a distributed AI fashion. To do so, we assume three small edge centers and one large datacenter. Each edge center has a single machine and the datacenter has 50 machines of same architecture and the characteristics.

Results and discussion

Experimental data

In this work, the road dataset of the Massachusetts province in the United States (US) was selected and refined as the experimental data. The image size was 1500 pixels $\times$1500 pixels, and the spatial resolution was 1 m. In this paper, 600 images in the dataset are nominated for training purpose, while 100 images are kept for testing purpose, and other 100 images are used for validation and computing the correctness of the model.

In deep learning, the nonexistence and shortage of the training samples will straightforwardly clue to overfitting, that is, the model will over fit the data on the training set, which will easily lead to the inaccuracy of the prediction on the validation set. The training set of 600 images used in this paper is not enough for training. Consequently, in order to rise and upsurge the quantity and volume of data which is nominated for the training stage, it is necessary to process the data image used for training. The experiment uses geometric changes to augment the dataset, such as:

- **Random cropping**: The local images at dissimilar locations and places could possibly be achieved and attained by random cropping of numerous images.
- **Flip conversion**: The process of flipping sensing images sideways the horizontal or vertical directions.
- **Contrast transformation**: Set the contrast transformation factor randomly to adjust the contrast of the image.
- **Random rotation transformation**: The process of spinning the image randomly by numerous angles and directions.

The unaltered original image is shown in Fig. 9a, right-handed spin of 90°, 180°, and 270° is shown in Fig. 9b through d, vertical and horizontal mirror inversion is shown in Fig. 9e, contrast transformation is shown in Fig. 9g, and random cropping is shown in Fig. 9h. Two thousand pictures and labels altogether were generated after the data image was enhanced using the geometric modification approach. The training set for the final road data set involves a minimum of approximately 1400 photos, the test set of up to 400 images, and the validation set of up to 200 images. At the same time, in order to explore whether the data augmentation strategy can significantly improve the quantity and volume of data available for the training stage, the data sets before and after data augmentation are respectively established for training in the experimental network.

Evaluating indicators and metrics

The outcomes of the segmentation were assessed using Recall, Precision, and F1-measure. Moreover, the metric Recall is in fact the fraction or percentage of samples that were accurately estimated and assessed to be positive to all samples that were really positive. The fraction of the percentage of successfully predicted samples to all estimated and predicted samples is known as accuracy, the F1 metric shows the harmonic average of the accuracy and the recall indicators, and in the problem of semantic segmentation, the intersection and union ratio is the intersection and union ratio between the real label and the predicted value of the class, and MIoU is the average intersection and union ratio of each class in the dataset, which is more accurate. The following are the calculating formulas for these numerous indicators:

$$P_{rec} = \frac{TP}{TP + FN} \tag{7}$$

Fig. 9 Data expansion processing results

$$P_{pre} = \frac{TP}{TP + FP} \tag{8}$$

$$F1 = 2 \times \frac{P_{pre} \times P_{rec}}{P_{pre} + P_{rec}} \tag{9}$$

$$MIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{TP}{FN + FP + TP} \tag{10}$$

In the above formulas, the notation TP is the amount of acceptably extracted road pixels, FP is the amount of erroneously extracted road pixels, and FN is the amount of all missing and un-predicted road pixels.

Experimental network settings

The unchanged collection of training examples and test data samples were selected and utilized for the assessment of various models and their comparative experimental studies to confirm the viability of the suggested DU-Net approach for road feature extraction from images related to remote sensing. Moreover, these datasets were also used to investigate the preeminence of the enriched neural network model i.e. DU-Net for road feature extraction as compared and associated with the three classical models including: (i) the U-Net network, (ii) the CAR-UNet network, and (iii) the MDAU-Net road extraction method. The experimental computer runs Windows, an operating system built using PyTorch 1.4.0, a deep learning framework. The GPU setup is GeForce GTX 1080, the graphics card is an NVIDIA 1080TI2, and the video RAM is 8G. The CPU configuration is E2650. In Table 2, experimental parameters are displayed.

Edge-cloud infrastructure settings

To run different modules of the proposed DU-Net-Cloud framework, we assume three small edge centers and one large datacenter as shown in Fig. 7. The infrastructure was modelled in the CloudSim simulator [23]. Each edge center has a single machine and the datacenter has 10 machines of same architecture and the characteristics as described in the following Table 3. The edges and datacenter are connected through a network with a link capacity of 1GB/s. We also assume that each server is virtualized and there are several types of virtual machines, as given in Table 3, running over the same servers. Subsequently, the DU-Net-Cloud application modules run inside the virtual machines. It is further believed that all services will use their supplied compute resources in a normally distributed manner. We also assume that each module of the DU-Net-Cloud runs in the place or resource that is initially allocated to it.

Table 2 Investigational parameters and their values for various models

Parameter	Value
Optimizer	Adam
Rate of learning	1×10^{-3}
Epochs	100
Size of batch	2
Loss function	Binary cross entropy
Size of block	7
Rate of dropout	0.15

Table 3 Servers characteristics for simulated edge-cloud setup [ECU = CPU speed (GHz) number of cores]

CPU model	Speed (GHz)	Cores	ECUs	Memory (GB)	Storage (TB)
Servers					
E5430	2.83	8	22.4	16	4
Virtual Machines					
t1.micro	1000 MIPS	1	1	0.613	1 GB
t2.nano	1000 MIPS	1	1	0.5	1 GB

Experiments and results

Ablation study

Influence of network structure layers on experimental results

First of all, the impact of the model deepness over various investigational findings is primarily examined using the U-Net learning network model. Note that the trial was successfully run on the Massachusetts Building dataset, and Table 4 displays how the performance of the experiment was impacted by various model depths.

The experimental findings show that the index of precision upsurges as the amount of network layers' increases, however when the amount of network layers exceeds a particular value (in this case 8), then we observed that the overfitting issue has a very substantial impact over

Table 4 Results comparison of dissimilar deep network trials on the Massachusetts dataset

Network	Layers	Recall/%	Precision/%	F1-measure/%
U-Net	4	85.26	90.74	87.03
U-Net5	5	86.92	91.64	90.21
U-Net6	6	90.73	92.87	91.62
U-Net7	7	94.38	96.39	94.27
U-Net8	8	91.53	93.33	93.62

Table 5 Massachusetts dataset for comparing network models' and experimentation outcomes

Network structure	Precision/%	Recall/%	F1-measure/%
U-Net7 + BN	96.94	94.76	95.14
U-Net7 + DropBlock	95.97	94.82	96.01
U-Net7 + BN+ DropBlock	97.12	96.78	96.68

Table 6 Massachusetts dataset accuracy evaluation table

Network	Recall/%	Precision/%	F1-measure/%	MIoU/%
U-Net	95.32	94.23	94.77	62.83
CARU-net	96.22	95.34	95.78	63.47
MDAU-Net	96.54	95.86	96.20	65.98
DU-Net	96.96	97.48	96.72	67.05

our findings. Through comparing the five network models, it can be shown that the U-Net7 approach has the greatest road extraction consequence (impacts), having the highest recall rate, accuracy rate, and F1 value, with an accuracy rate that reaches 97.36%. Therefore, it can be inferred from the experimental results that the correctness and precision of the outcomes increases with the quantity of layers in the network topology. Even though, the receptive field will expand with deeper network layers, there will be more down sampling, which means that more precise information will be lost. Additionally, as the network gets deeper, the overfitting issue will get worse, which will reduce the experimental accuracy. The experiment's best basic network model is determined to be U-Net7 based on the aforementioned factors [24].

Impacts of the DropBlock and Batch Normalization

The experiments demonstrated in this section, in fact, investigate the effects of the DropBlock and Batch Normalization techniques on the investigational outcomes grounded on the chosen U-Net7 learning network model. Furthermore, we also investigate the impacts of the DropBlock and Batch Normalization approaches over the proposed DU-Net-Cloud, and other closest rivals. This should be kept in mind that all the experimental outcomes were achieved using similar experimental parameters and setups [24, 25].

After adding the BN (batch normalization) layer and the DropBlock module to the U-Net7 network structure model, we achieved satisfactory outcomes. Table 5 displays the accuracy comparison of the experimental findings on the Massachusetts dataset. When the suggested U-Net7 network model is joined with the DropBlock and the BN approach at the identical period, under the identical and almost similar investigational circumstances and data sets, all the correctness and precision indicators have improved somewhat in comparison to when the U-Net7 network model is joined with the DropBlock and BN models. Moreover, we observed that in such situations the road extraction has also met the high accuracy requirements, with a precision of 97.12% correspondingly. This level of high precision is also indicating that the addition of both the DropBlock and BN layers can successfully resolve the gradient dispersion Boost the road identification's precision. This technique's viability and promise for extracting objects from remote sensing images are demonstrated.

Comparison of network model DU-Net with other research results

The experiment and assessment outcomes discussed in this section are based on the Massachusetts dataset, and the U-Net, CAR-Unet, MDAU-Net, and DU-Net are trained respectively. After the testing stage, the experimental results of the four networks in the dataset are compared in detail, as shown in Table 6.

Part of the visual result is shown in Fig. 10, can be seen from the results figure, U-Net, SegNet extraction result of image, the region of the a-d road fracture phenomena is evident, the existence of fracture is more prevalent, among them, the area b tiny road in the picture structural information extraction is incomplete, and the regional c road edge detection effect is poor due to the thick buildings. There are a few neighboring roads adhesion phenomena in region d, and it is difficult to identify small roads under shadow occlusion. Compared to the previous three methods, the proposed DU-Net network can effectively extract roads comprehensively, accurate segmentation on the edge of the road, and path of some small details can also have. MDAU-Net in the extraction result image, area a - d road images in small isolated points, and some road structure incomplete information extraction. The proposed DU-Net network, can effectively extract roads comprehensively, accurate segmentation on the edge of the road. Moreover, path of some small details can also has good recognition effect. We also observed that it might have some good recognition effects on some small roads and improves the adhesion of roads. The final extraction effect has a high similarity with the label image.

Model and application latencies

The training and prediction modules were separately installed on edge and cloud and the model performance was assessed in terms of training and prediction duration, as shown in Table 7 and Fig. 11. This can be seen

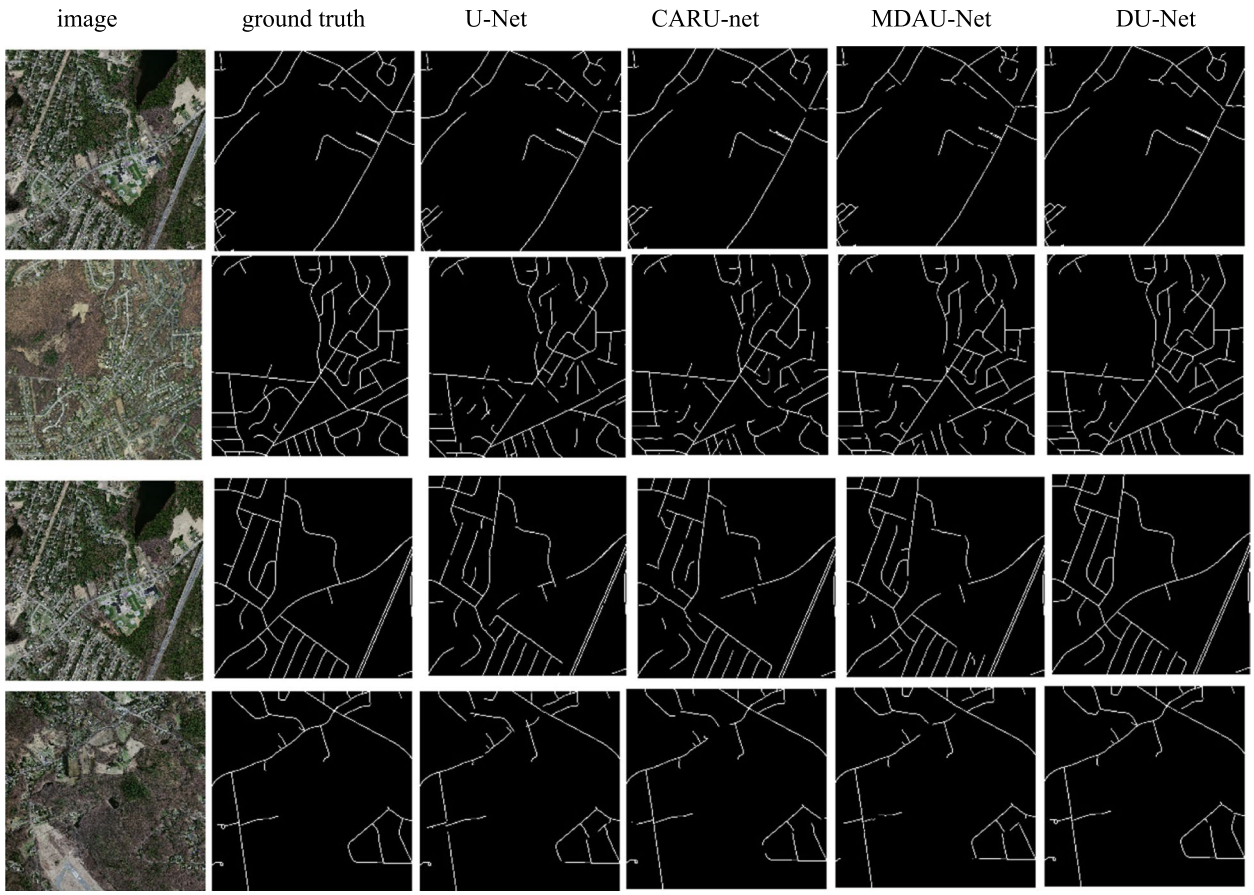

Fig. 10 Road comparison experiment extraction results

Table 7 Model performance using edge intelligence

	Training time/S	Prediction time/S
Cloud	**9943.6**	**1998.34**
Edge	**17,151.5**	**1224.4**
Edge-cloud	**9683.5**	**1026.8**

that using a combination of cloud and edge, the model performance can be significantly improved as compared to using only cloud or edge infrastructure. However, we observed variations that were related to the amount of data stored over the cloud, the model training parameters, and the network capacity. For example, the latency of the training model over a high speed network was significantly lower than the latency observed at a low capacity network and vice versa. We believe that the latency of the model along with the training durations can be considerably reduced through considering only important data for the training purpose. This can be achieved through a data aggregation scheme that works on the local edge cloud and examine the collected data for duplication and redundancy. If there are redundant data, that could be dropped at the edge level and only essential data is sent over the network to the remote cloud. Figure 12 shows the application latency when executing on the cloud, edge, or a combination of both i.e. edge-cloud. The application latency can be reduced that can be as high as approximately 11.45%.

Conclusions and future work

An improved U-Net network-based road extraction technique is presented in this paper. It has been successfully used to segment roads in remote sensing images. To increase the precision of the road segmentation method in remote sensing photos, this study combines the parallel attention mechanism module ProCBAM with the seven-layer U-Net network structure, adds DropBlock to the convolutional layer, and adds BN to the decoding path. In the meantime, trials are being done to compare it to U-Net, CARU-net, and mado-net. The findings indicate that this method has greatly improved in terms of recall

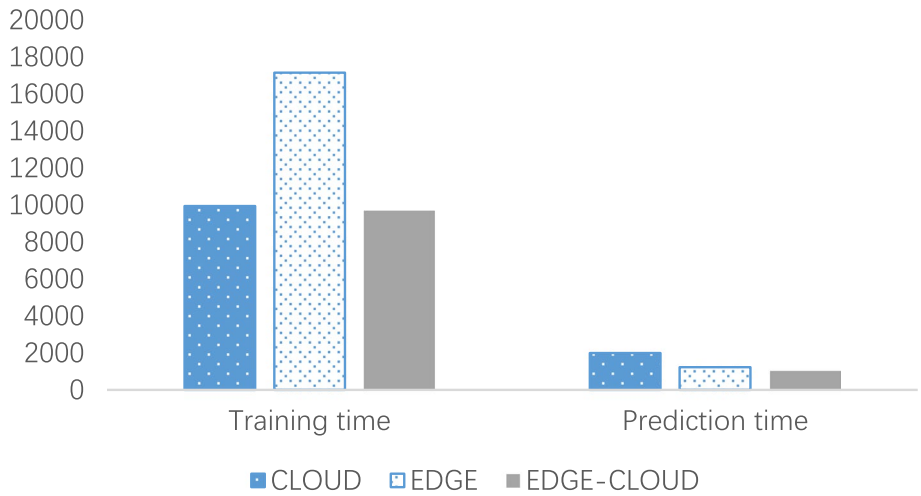

Fig. 11 Model training and prediction times using edge, cloud, and edge-cloud

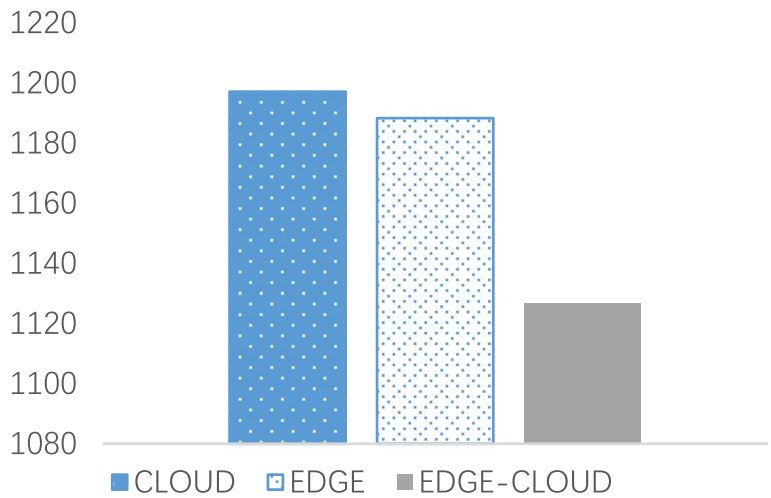

Fig. 12 Application latencies using edge, cloud, and edge-cloud

rate, accuracy rate, and F1 value. In particular, the accuracy rate reached 97.48% and the final extraction impact is more accurate for identifying and classifying road conditions. Because of this, the DU-Net network structure model suggested in this paper has significant practical implications for extracting roads from remote sensing pictures. The experimental results further demonstrate that the algorithm is capable of efficiently extracting the categories from the feature map in order to obtain the road data as well as integrating the low-level and high-level characteristics.

The accuracy enhancement provided by this method is constrained since not all neurons can be activated by the activation function used. As a result, future study should aim to improve the model's structure and find the best activation function. We will also consider the comparative study of various activation function for the proposed DU-Net-Cloud model and analyze which one is better than the other under what kind of circumstances. Also, we believe that the model behavior for a large edge-cloud platform should be studied so that it can be identified which module is running more efficiently on what types of hardware resources. In fact, this will be helpful is environments where emergent response is needed.

Acknowledgments
This research was supported by the Shanxi Graduate Education Innovation Project of China, Grant No. 2019SY126.

Authors' contributions
J.K. and Y.Z. wrote the main manuscript text and J.K. prepared figures 1-12. All authors reviewed the manuscript. The author(s) read and approved the final manuscript.

Availability of data and materials
The data was freely downloaded from the given website (https://www.cs.toronto.edu/~vmnih/data/), all accessed on 1 August 2022.

Declarations

Competing interests
The authors declare no competing interests.

Received: 9 November 2022 Accepted: 4 February 2023
Published online: 27 February 2023

References

1. Das S, Mirnalinee TT, Varghese K (2011) Use of salient features for the design of a multistage framework to extract roads from high-resolution multispectral satellite images. IEEE Trans Geosci Remote Sens 49:3906–3931
2. Lv X, Ming D, Chen YY, Wang M (2019) Very high resolution remote sensing image classification with SEEDS-CNN and scale effect analysis for superpixel CNN classification. Int J Remote Sens 40:506–531
3. Lv X, Ming D, Lu T, Zhou K, Wang M, Bao H (2018) A new method for region-based majority voting CNNs for very high resolution image classification. Remote Sens 10:1946
4. Luo QZ, Yin Q, Kuang DB (2007) Research on extracting road based on its spectral feature and shape feature. Remote Sens Technol Appl 22(3):339–344
5. Lin XG, Zhang JX, Li HT et al (2009) Semi-automatic extraction of ribbon road from high resolution remotely sensed imagery by a T-shaped template matching. Geomat Inf Sci Wuhan Univ 34(3):293–296
6. Chen LF, Wen J, Xiao HG et al (2015) Road extraction algorithm for high resolution SAR image by fusion of MRF segmentation and mathematical morphology. Chin Space Sci Technol 35(2):17–24
7. Cao YG, Wang ZP, Shen L et al (2016) Fusion of pixel-based and object-based features for road centerline extraction from high-resolution satellite imagery. Acta Geod Cartogr Sin 45(10):1231–1240
8. Zhang S, Li C, Qiu S, Gao C, Zhang F, Du Z, Liu R (2020) EMMCNN: an ETPS-based multi-scale and multi-feature method using CNN for high spatial resolution image land-cover classification. Remote Sens 12:66
9. Cheng G, Wang Y, Xu S et al (2017) Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network. IEEE Trans Geosci Remote Sens 55(6):3322–3337
10. Liu Z, Li X, Luo P et al (2015) Semantic image segmentation via deep parsing network. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), pp 1377–1385
11. Jeong M, Nam J, Ko BC (2020) Lightweight multilayer random forests for monitoring driver emotional status. IEEE Access 8:60344–60354
12. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015. Springer, Cham, pp 234–241
13. Yuan W, Xu WB, Zhou T (2021) A loss function of road segmentation in remote sensing image by deep learning. Chinese Space Sci Technol 4:1–10
14. Jin F, Wang LF, Liu Z et al (2019) Double U-net remote sensing image road extraction method. J Geomat Sci Technol 36(4):377–387
15. Wang Z, Yan HW, Lu XM et al (2020) High-resolution remote sensing image road extraction method for improving U-Net. Remote Sens Technol Appl 35(4):741–748
16. John D, Zhang C (2022) An attention-based U-Net for detecting deforestation within satellite sensor imagery. Int J Appl Earth Obs Geoinf 107:102685
17. Chen T et al (2022) A Siamese network based U-Net for change detection in high resolution remote sensing images. IEEE J Sel Top Appl Earth Obs Remote Sens 15:2357–2369
18. Zhao Q et al (2021) Semantic segmentation with attention mechanism for remote sensing images. IEEE Trans Geosci Remote Sens 60:1–13
19. Woo S, Park J, Lee JY, et al (2018) CBAM: Convolutional Block Attention Module. Proceedings of the European Conference on Computer Vision (ECCV). pp. 3-19
20. Tseng K-K et al (2021) A fast instance segmentation with one-stage multi-task deep neural network for autonomous driving. Comput Electr Eng 93:107194
21. Kong J, Gao Y, Zhang Y, et al (2021)Improved attention mechanism and residual network for remote sensing image scene classification, IEEE Access 9:134800-134808.
22. Zhang Y, Kong J, Long S et al (2022) Convolutional block attention module U-Net: a method to improve attention mechanism and U-Net for remote sensing images. J Appl Remote Sens 16(2):026516
23. Huang H, Chen Y, Wang R (2021) A lightweight network for building extraction from remote sensing images. IEEE Trans Geosci Remote Sens 60:1–12
24. Wu Y et al (2022) Edge computing driven low-light image dynamic enhancement for object detection. In: IEEE Transactions on Network Science and Engineering
25. Xu L et al (2021) HA U-Net: improved model for building extraction from high resolution remote sensing imagery. IEEE Access 9:101972–101984

Publisher's Note