ISPRS Journal of Photogrammetry and Remote Sensing 175 (2021) 353–365

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Review Article

A Global Context-aware and Batch-independent Network for road extraction from VHR satellite imagery

Qiqi Zhu [a,c], Yanan Zhang [a], Lizeng Wang [a], Yanfei Zhong [b], Qingfeng Guan [a,c,*], Xiaoyan Lu [b], Liangpei Zhang [b], Deren Li [b]

[a] School of Geography and Information Engineering, China University of Geosciences, Wuhan 430078, Hubei Province, China
[b] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430074, Hubei Province, China
[c] National Engineering Research Center of GIS, China University of Geosciences, Wuhan 430078, Hubei Province, China

ARTICLE INFO

Keywords:
Road extraction
Deep learning
Global Context-aware block
FRN layer
Roads dataset

ABSTRACT

Road extraction is to automatically label the pixels of roads in satellite imagery with specific semantic categories based on the extraction of the topographical meaningful features. For governments, timely and accurate road mapping is crucial to plan infrastructure development and mobilize relief around the world. Recent advances in deep learning have shown their dominance on road extraction from very high-resolution (VHR) satellite imagery. However, previous road extraction based on deep learning mainly stacked the multiple convolution operators and failed to predict the contextual spatial relationship correctly. Besides, the precision of cross-domain road extraction is limited by an insufficient amount of labeled data and the transferability of the model. To remedy these issues, a Global Context-aware and Batch-independent Network (GCB-Net) is proposed, which is a novel road extraction framework extract complete and continuous road networks. In GCB-Net, the Global Context-Aware (GCA) block is added to the encoder-decoder structure to effectively integrate global context features. The Filter Response Normalization (FRN) layer is used to enhance the original basic network, which eliminates the batch dependency to accelerate learning and further improve the robustness of the model. Experimental results on two diverse road extraction data sets demonstrated that the proposed method outperformed the state-of-the-art methods both quantity and quality. Moreover, to test the robust generalizability of the proposed method, the proposed CHN6-CUG Roads Dataset was used for spatial transfer evaluation, and GCB-Net achieved significantly higher transferability than other methods.

1. Introduction

A large amount of high-resolution satellite imagery can be obtained, providing an important data source for automatic road extraction. Compared with low-resolution satellite imagery, high-resolution satellite imagery has finer spectral and texture features, making it possible to extract more precise roads (Zhu et al., 2018). Road extraction is considered to be an important tool for urban planning and decision-making. It has been applied in many fields, such as urban planning (Wu et al., 2019; Huang et al., 2018), automated driving services (Claussmann et al., 2019), humanitarian aid (Bonafilia et al., 2019), and geographical information upgrading (Levin and Duke, 2012; C. Zhang et al., 2019).

The road extraction method can be divided into the pixel-based method, object-oriented method, and deep learning method. The pixel-based method mainly takes advantage of the different waveband characteristics, which can extract the rural roads in the satellite images with the simple background (Sghaier and Lepage, 2015; Liu et al., 2017; Zhang et al., 2017; Dai et al., 2019). The object-based method identifies the road object as a whole, which has good noise resistance and applicability (Xie and Weng, 2016; Zang et al., 2016; L. Chen et al., 2018; Maboudi et al., 2018). However, due to the richer spectral features and more scaled geometric features in VHR satellite imagery, the traditional road extraction methods lack consideration of the diverse characteristic of complex roads. And complex radians and curvature of roads make it difficult for traditional methods to extract complete, continuous road intersections.

In recent years, the development of deep learning has greatly boosted the progress of road extraction (Mnih and Hinton, 2010; Cheng et al., 2017; Yang et al., 2019; Yuan et al., 2020). The Fully Convolutional

* Corresponding author.
E-mail addresses: zhuqq@cug.edu.cn (Q. Zhu), guanqf@cug.edu.cn (Q. Guan).

https://doi.org/10.1016/j.isprsjprs.2021.03.016
Received 23 October 2020; Received in revised form 24 February 2021; Accepted 18 March 2021
Available online 5 April 2021

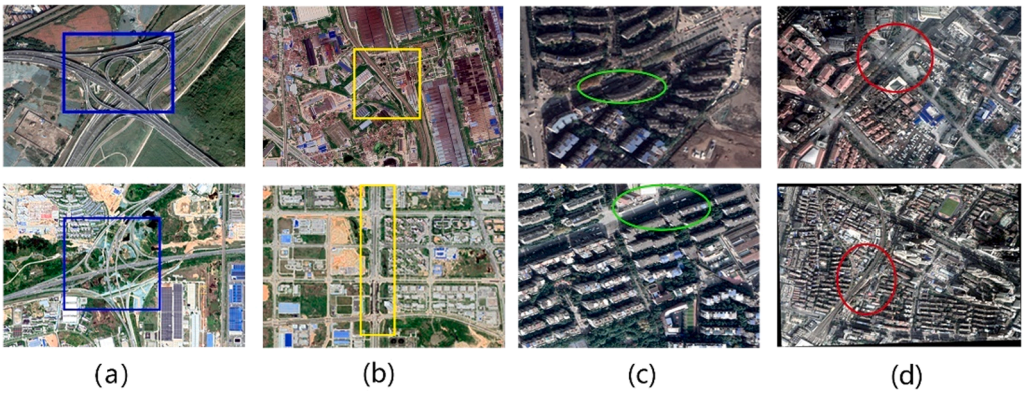

Fig. 1. Challenges in road extraction of VHR satellite imagery: (a) complex intersection; (b) roads with different spectral; (c) the shadows of buildings and trees; (d) the similarity of road texture with other materials.

Neural Network (FCN) (Long et al., 2015) stacks multiple convolution layers and pooling layers to gradually enlarge the receptive fields of the network, which is more conducive to the extraction of road information (Buslaev et al., 2018; Han et al., 2018). U-Net architecture (Ronneberger et al., 2015) has been employed to effectively extract road information, which concatenates the different levels of the feature maps (Zhang et al., 2018; Lu et al., 2019; Sun et al., 2018; Diakogiannis et al., 2020). With the encoder-decoder architecture, D-LinkNet (Zhou et al., 2018) obtains the coarse localization of road extraction through the high-level features and refines the boundary through the low-level features containing the spatial structure details (Chen et al., 2020).

There are several problems in previous works for road extraction: (1) As shown in Fig. 1, since the shadows of buildings and trees, diverse imaging conditions, and the similarity of road spectral with other features, the extracted road networks often produce fragmented sections. This inhibits the accurate estimation of road extraction in VHR satellite imagery; (2) In most of the previous works, the features of the entire Spatio-temporal space of input image (L.C. Chen et al., 2018; Zhu et al., 2020) were ignored, which is conducive to interpreting the relationship among multiple road regions and generating more complete road results; (3) Previous works usually use the batch normalization (BN) (Ioffe and Szegedy, 2015) layer to accelerate model training and improve performance. However, the BN layer relies on a sufficiently large batch, which results in inconsistencies in training and testing (Singh and Shrivastava, 2019); (4) The road is very different all over the world, and most available datasets for road extraction are heavily biased towards particular regions (Bonafilia et al., 2019). It is difficult to obtain high enough quality results only by transferring the knowledge learned from these available datasets to other regions directly.

To solve the mentioned issues, a Global Context-aware and Batch-independent Network (GCB-Net) is proposed for VHR satellite imagery road extraction. GCB-Net uses an encoder-decoder structure for capturing sharper road object boundaries by gradually recovering the spatial information, and the Global Context-Aware (GCA) block is added between the two stages of encoder blocks. To aggregate multi-scale contextual information, the feature maps obtained by the encoder are passed through the dilated convolution.

From the above, the main contributions of this paper are as follows:

(1) GCB-Net is proposed to address the difficulty of VHR satellite imagery road extraction. Specifically, the framework is an end-to-end pixel-wise deep convolutional neural network (CNN) architecture, which integrates the low-level detail information, high-level semantic information, and global context information in an interweaved way.
(2) To improve the feature representation, the GCA block is constructed to capture road-specific contextual information. GCA block is appended on multi-level feature maps of the encoder. This improves the completeness of the generated road map.
(3) By using Filter Response Normalization (FRN) as a normalization method, the training results of the proposed network are no longer affected by the batch size. With the simplified optimization allowing the model to train better, the proposed network achieved better robustness for diverse road scenes.
(4) A new largescale data set is constructed, i.e., CHN6-CUG Roads Dataset, for VHR satellite imagery road extraction. Due to geographic differences and low coverage, the models are trained on public road datasets to perform poorly in China. CHN6-CUG Roads Dataset is established to provide a better data resource for studying the road of China. Besides, to assess the transferability of the proposed model, a set of representative methods are evaluated on the proposed dataset.

Three different road datasets, i.e., DeepGlobe Roads Dataset, SpaceNet Roads Dataset, and CHN6-CUG Roads Dataset, are comprehensively evaluated to confirm the effectiveness of GCB-Net.

The rest of this paper is organized as follows. Section 2 reviews the related works road extraction methods based on deep learning. Section 3 describes two public data sets for road extraction and the details of the CHN6-CUG Roads Dataset. In Section 4, the proposed GCB-Net for VHR satellite imagery road extraction is introduced. The experimental results and analysis are reported in Section 5. Finally, discussions and conclusions are presented in Sections 6 and 7, respectively.

2. Related work

Based on per-pixel road segmentation, some road frameworks (Máttyus et al., 2017; Bastani et al., 2018; Batra et al., 2019; Zhou et al., 2020) incorporated topological information and contextual priors to improve robustness for road extraction. To further enhance the ability of feature expression, some road extraction works using encoder-decoder structure models, such as U-Net (Ronneberger et al., 2015), Segnet (Badrinarayanan et al., 2017), Linknet (Chaurasia and Culurciello, 2017), DeepLabv3 (Y. Chen et al., 2018), which integrate features in multiple layers of CNN to exploit the multi-scale information at different semantic levels. Attempting to enrich the encoder-decoder networks, some works employed refinement strategy to incorporate the multi-scale contextual information. For example, Zhou et al. (2018) used skip connections and added dilated convolution layers in the center part. This structure realizes multi-scale road information recovery by adjusting the receptive fields of feature points. To improve the poor effect of PSPNet (Singh and Krishnan, 2020) in multi-scale road extraction, Gao et al. (2018) customized a pyramid pooling structure and designed a tailored pooling pyramid module (TPPM) for strip roads. Inspired by the U-Net and atrous spatial pyramid pooling (ASPP) (Chen

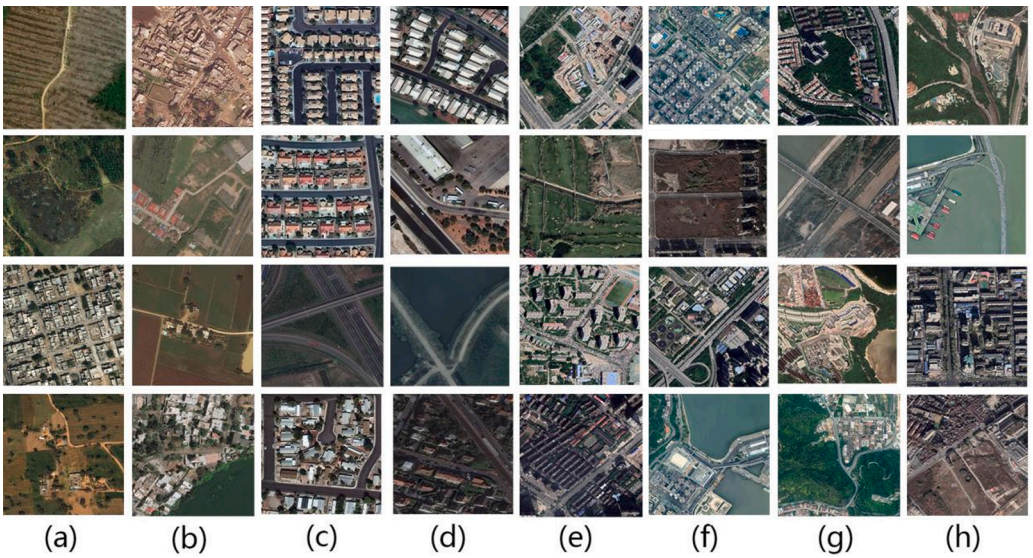

Fig. 2. Overview of the data set from (a)-(b) DeepGlobe Roads Dataset; (c)-(d) SpaceNet Roads Dataset; (e)-(h) CHN6-CUG Roads Dataset.

et al., 2017) approach, H. He et al. (2019) introduced an ASPP-integrated U-Net to extract multi-scale features of road targets. Tan et al. (2020) added two modules to the decoder, the scale fusion module and the scale sensitive module, which to effectively utilize the different levels of convolutional layers to perceive the road. To capture the road-specific contextual information, Tao et al. (2019) proposed a spatial information inference structure (SIIS).

In recent years, some works based on Generative Adversarial Network (GAN) model have also provided high-quality road extraction results (Y. Zhang et al., 2019; X. Zhang et al., 2019; Shamsolmoali et al., 2020). Y. Zhang et al. (2019) proposed a Multi-supervised Generative Adversarial Network (MsGAN) road extraction framework, which enables not only road area detection but also road topology reconstruction. X. Zhang et al. (2019) proposed an improved generative adversarial network-based road extraction method for aerial images. This does not require a large training dataset and still has a good performance. Shamsolmoali et al. (2020) fused the feature pyramid network into a generative adversarial network for road segmentation, so that improved the performance of adversarial learning for generating realistic synthetic remote sensing images. The GAN-based method is a generative model as well as a semi-supervised and unsupervised learning model. It can learn deep representations without the need for large amounts of labeled data. They tended to select training samples with consistent local structures (Y. Zhang et al., 2019; X. Zhang et al., 2019).

Although the above methods capture road targets at different scales, for narrow and sparse roads, there is a lack of comprehensive

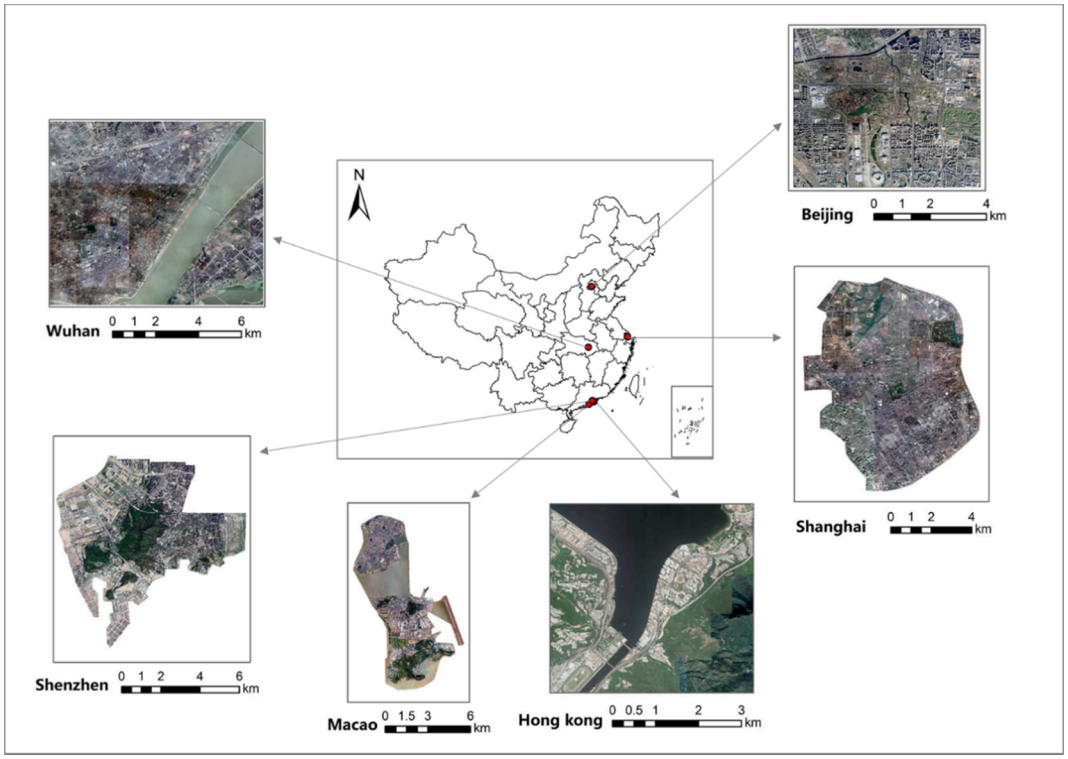

Fig. 3. Overview of the study areas in CHN6-CUG Roads Dataset.

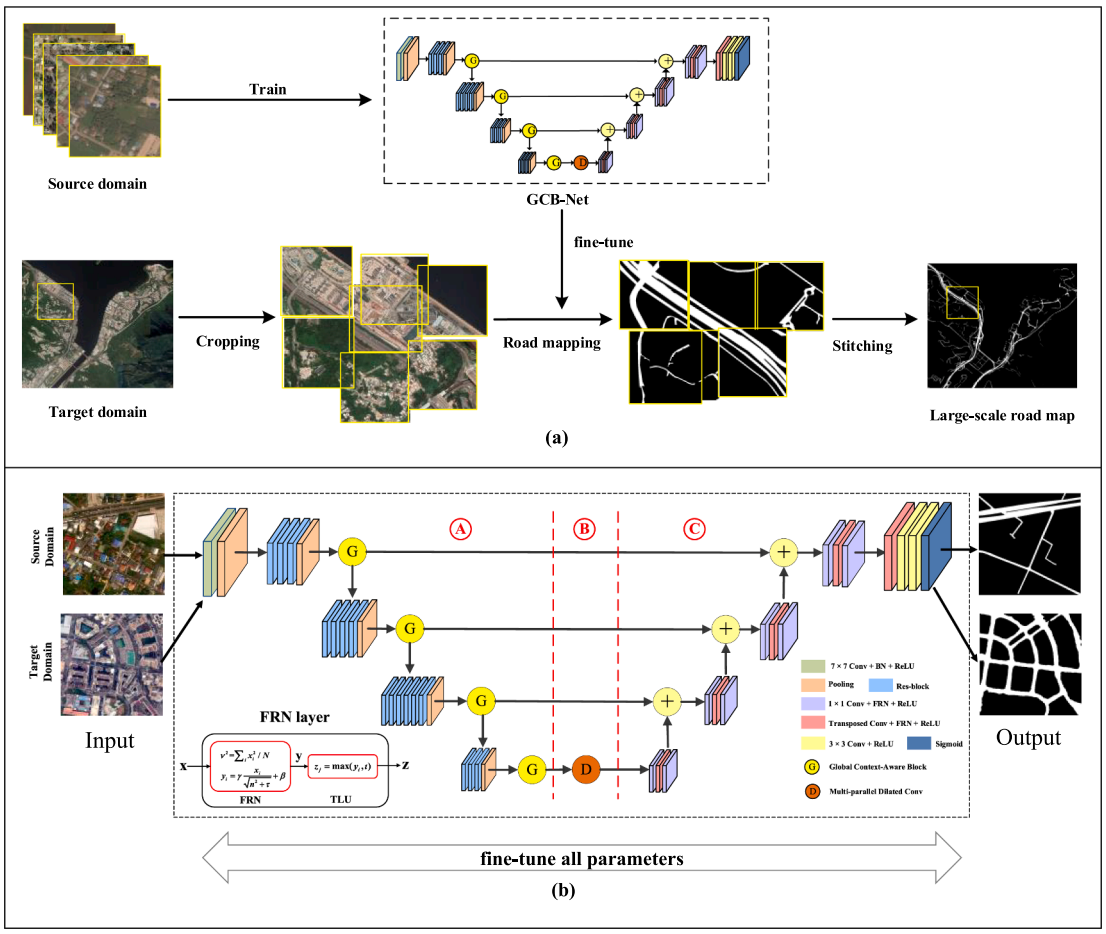

Fig. 4. Overview of the proposed framework for road extraction. (a) Is the process of obtaining large-scale road maps. (b) Is the detailed design of GCB-Net.

consideration of the relation and correlation of each position, which is also important for road extraction. Motivated by the recent success of attention networks (such as Hu et al., 2018; Wang et al., 2018; Hu et al., 2019; J. He et al., 2019), several recent works (Wang et al., 2019; Xie et al., 2019) combined attention block to overcome the limitation of local convolution operations for road extraction. Wang et al. (2019) introduced a road extraction network named Non-Local LinkNet (NL-LinkNet) with non-local blocks in order to model long-range dependencies. Xie et al. (2019) embedded an attention block between the encoder and decoder, which was designed to preserve the dependencies of long-distance spatial information and different feature channels. However, the global relation modeling of previous methods is usually limited, which cannot represent the characteristics of roads under different diverse conditions.

3. Study areas and data

3.1. Public roads datasets for road extraction

DeepGlobe Roads Dataset (Demir et al., 2018): As shown in Fig. 2, this data set contains data with pixel-level annotations from Thailand, India, and Indonesia. The ground resolution of each image is 50 cm/pixel, and the pixel resolution is 1024 × 1024. Following Batra et al. (2019), the original 6226 training images are split into 4976 images for training and 1250 for testing. To augment the training set, crops of the size 512 × 512 with an overlapping area of 256 pixels are created. Finally, the new DeepGlobe Roads Dataset has 42,255 training images and 6116 testing images.

SpaceNet Roads Dataset (Van Etten et al., 2018): As shown in Fig. 2, this data set provides imagery roads centerline annotations from Vegas, Paris, Shanghai, Khartoum, which consists of 2780 images and centerline maps. The ground resolution is 30 cm/pixel, and the size of each image is 1300 × 1300. Following Batra et al. (2019), the Gaussian distance transformation along the centerline of the road is used to generate a road pixel-level mask. The data set is divided into 2213 images for training and 567 for testing. To increase the training dataset, crops of 650 × 650 with an overlapping area of 215 pixels are created, thus providing 35,392 images training images and 2264 testing images.

3.2. CHN6-CUG roads dataset: A new data set for road extraction

To advance the state of the arts in road extraction, CHN6-CUG Roads Dataset is constructed, which is a manually labeled pixel-level VHR satellite imagery. As shown in Fig. 3, this data set is a new largescale satellite image data set of representative cities in China. The images are selected from six cities with different levels of urbanization, i.e., the Chaoyang area of Beijing, the Yangpu District of Shanghai, Wuhan city center, the Nanshan area of Shenzhen, the Shatin area of Hong Kong, and Macao.

This data set is acquired from Google Earth (Google Inc.). All the images are labeled by image interpretation experts, and some samples are shown in Fig. 2. According to the surface coverage, the labeled road includes the pavement covered by track and the pavement without track cover. And according to the physical perspective of geographical factors, labeled roads include railways, highways, urban roads, and rural roads. Each image has a size of 1024 × 1024 pixels and a resolution of 50 cm/pixel. To be consistent with the DeepGlobe data set, the images are scaled from 1024 × 1024 to 512 × 512. Finally, 4511 labeled images

with the size of 512 × 512 are split into 3608 images for training and 903 for testing.

3.3. Why is CHN6-CUG roads dataset for road extraction?

Compared with existing road extraction data sets (such as DeepGlobe Roads Dataset and SpaceNet Roads Dataset), CHN6-CUG Roads Dataset has the following properties.

(1) Relatively large data sets for China: Public roads datasets and benchmarks such as Massachusetts, DeepGlobe, SpaceNet, and corresponding challenges have led to more diverse algorithms for public benchmarks to evaluate (Demir et al., 2018). Deep learning is a data-driven technology, and its generalization performance is limited by the diversity of training data. Considering the low coverage of public road datasets in China, as well as the complexity and difference of China's roads, the model trained on these public data sets performs well in the training set, while it performs poorly in China and other regions. Therefore, the application of CHN6-CUG Roads Dataset can make up for the situation that the public roads datasets cover fewer areas in China, and promote the establishment of road extraction systems in China and other areas.

(2) Higher heterogeneous: The diversity of data sets is the key to improve the generalization performance of models. There are various types of roads in the CHN6-CUG Roads Dataset, which almost cover roads under various complex scenes, as shown in Fig. 3, including urban road, forest road, country road, factories and mines road, built-up road, railway crossing, overpass, viaduct, and intersection. Therefore, the images in CHN6-CUG Roads Dataset are highly heterogeneous and distinctive from each other in road extraction. This is a good representation of the roads in China and other countries and can be used as a benchmark to evaluate the availability and generalization of the deep learning model.

(3) More geographically robust: The higher the intraclass diversity in the dataset, the stronger the generalization ability of the model. This helps the model to recognize more road scenes. CHN6-CUG Roads Dataset covers cities in different geographical regions of China, these cities have different urbanization levels, city sizes, development degrees, urban structures, history, and culture. The geographical differences make the roads have different spectral features, texture features, and geometric features, thus, it is helpful to improve the robustness and transferability of the deep learning model.

4. Methodology

The proposed GCB-Net is designed to enhance the performance of road extraction, which includes three main parts. First, the rich road semantic information is extracted by encoder modules. By incorporating the GCA block, the encoder module is enriched in the encoder-decoder networks. Continuously, outputs of the encoder features are fed into the multi-parallel dilated convolution layers for generating multi-scale contextual information. Finally, the feature maps are recovered to final road segmentation maps by the effective decoder modules. And FRN is used as a normalization method, which is aimed at eliminating the batch dependency and improving the robustness of the road extraction model. Fig. 4 shows the architecture of the proposed framework.

4.1. Encoder with global context-aware block

4.1.1. Relation-augmented road features generation by encoder modules

The road extraction network takes the RGB images as input. The encoder starts with an initial block which performs convolution with a kernel size of 7 × 7 and a stride size of 2 on the input image. In the later portion of the encoder, there are four repeated groups of convolutional layers, each of which consists of residual blocks (He et al., 2016). The spatial max-pooling is used for down-sampling, with a 7 × 7 kernel and a stride of 2. Then the feature maps are fed into the GCA Block in four groups, which gather and distribute long-range road features. There are four max-pooling operations in the down-sampling, and the size of the subsequent feature maps is 16 × 16. At the beginning of each down-sampling group, the number of feature channels doubles. Thus, there are 512 channels at the end. The GCB-Net is initialized by the parameters of ResNet34 pre-trained on ImageNet (Deng et al., 2009) dataset, which is designed to accelerate the convergence rate of gradient descent and retain more details and efficiently improve model performance.

Due to multiple down-sampling operations in the encoder, some spatial information will be lost. And using only the encoder under-sampling output is difficult to restore the spatial information. In this paper, an aggressive yet efficient operation is adopted to fuse features by addition operation, which includes the high-level features after up-sampling and the low-level features after GCA blocks of each encoder layer. By doing this, the spatial information lost during down-sampling is recovered, and the relation-augmented feature from different levels are integrated into the decoder.

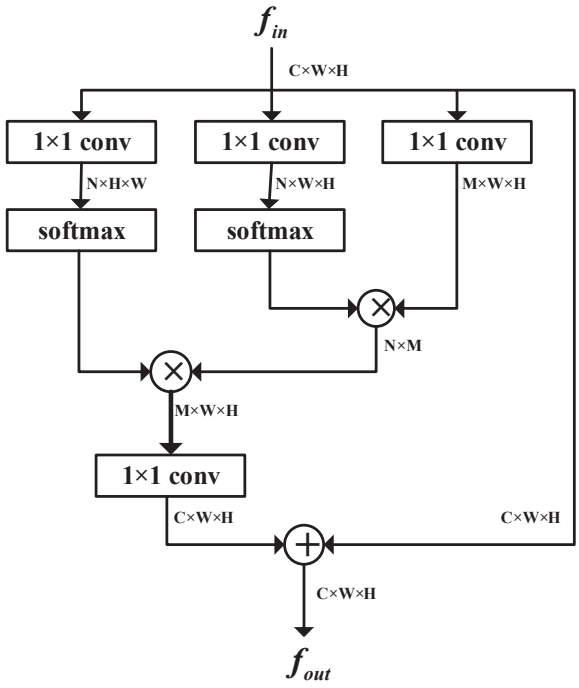

Fig. 5. The overall flowchart of the Global Context-aware Block.

4.1.2. Global contextual information incorporating

For the natural properties of roads, such as natural connectivity, long span, and the variation of topological information. The simple convolutional operator focuses on local neighborhoods and may fail to sensitively capture the global relationship among the entire spatial and temporal space. The attention modules and correlation operators are built in a wide range of recognition tasks, via aggregating query-specific global context to each query position. Based on this observation, the design of the GCA block was inspired by Double Attention Networks (Chen, Y. et al., 2018), Non-local Neural Networks (Wang et al., 2018), and Compact Generalized Non-local Networks (Yue et al., 2018).

As shown in Fig. 5, input feature maps $f_{in}^{t} \in R^{C\times W\times H}(t = 1,2,3,4)$are firstly fed two convolutional layers with 1 × 1 filters to generate two feature maps K^{t} and V^{t}, respectively, where $K^{t} \in R^{N\times W\times H}$ and $V^{t} \in R^{M\times W\times H}$. $M$and$N$are the number of channels, less than C for

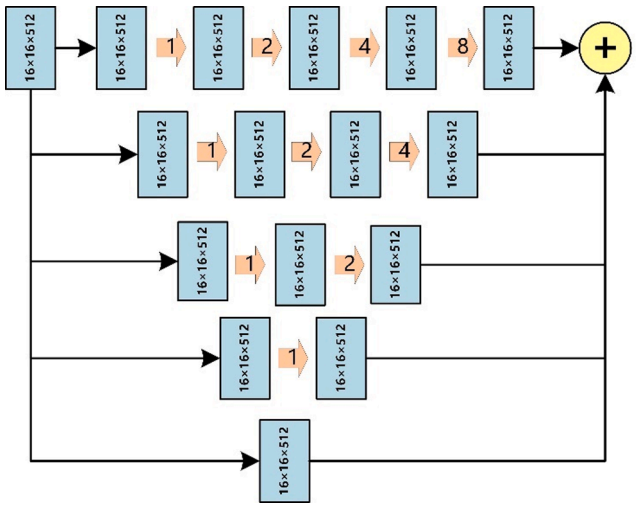

Fig. 6. The overall flowchart of the multi-parallel dilated convolution layers.

dimension reduction. Then, a softmax layer is applied on V^t over the channel dimension to gather local feature weights, and rewriting V^t as $\overline{V}^t \in R^{H\times W\times M}$. After obtaining feature maps$K^t$and$\overline{V}^t$, a valid attention weighting vector$A^t \in R^{N\times M}$is further generated via bilinear pooling. The affinity operation is defined as follows:

$$A^t = K^t \otimes \overline{V}^t = K^t \otimes softmax(V^t)^T \tag{1}$$

where $\otimes$ denotes element-wise multiplication, and t is the stage index.

Another convolutional layer with 1×1 filters is applied on f_{in} to generate $Q^t \in R^{N\times W\times H}$ for distributing an adaptive bag of visual primitives. To normalize Q^t into the one with unit sum, the softmax function is applied to give better convergence. At each location in the select feature maps, we can obtain a vector subset of feature vectors from A^t with soft attention:

$$F^t = A^t \otimes softmax(Q^t) = \left(K^t \otimes softmax(V^t)^T\right) \otimes softmax(Q^t) \tag{2}$$

$$f_{out} = a \cdot F^t + f_{in} \tag{3}$$

where a is a custom optimized parameter, the output result F^t is given by conducting two matrix multiplications with necessary reshape and transpose operations. As the weights are iteratively updated, this block can gradually filter out useless information and finally focus on useful contextual information, thus correct the prediction of this pixel.

GCA blocks enhance the description of pixel features by modeling pixel-to-pixel relationships. This inevitably increases the complexity of the model. However, compared to other attentional blocks, such as the Non-local block (Wang et al., 2018), this block can better overcome the contradiction between performance and complexity. The Non-local block is the first attention mechanism that captures the long-term dependence of deep neural networks with multiplication. Optimized for Non-local block, the GCA block performs a low-rank reconstruction of the pixel features. The GCA block can achieve a large receptive field with fewer layers that is close to that brought by more layers, which is applicable to lightweight networks. In practice, compared with Non-local blocks, GCA blocks reduce the computational complexity by about 10G FLOPs and Memory footprint by 3000 MB. In summary, the GCA block achieves a better balance of accuracy and computational cost, thus improving the overall performance of road extraction.

4.2. Multi-scale contextual information aggregating

The feature map obtained by the encoder is passed through the dilated convolution, which takes into account the expansion of the receptive field, to strengthen the spatial regions and feature channels.

As shown in Fig. 6, dilated convolutions between the encoder and decoder are inserted to support exponentially expanding receptive fields. Let each group of layers is denoted by F^{ℓ}, and i^{th} is the layer in the group ℓ by F_i^{ℓ}. The dilated convolutions defined as:

$$(F_i^{\ell} *_l k_i^{\ell})(p) = \sum_{s+lt=p} F_i^{\ell}(s) k_i^{\ell}(t) \tag{4}$$

where k_i^{ℓ} is the discrete filter with layer F_i^{ℓ}, $*l$ is an $l-$dilated convolution, and the domain of $p$is the feature map in F_i^{ℓ}. Following (Zhou et al., 2018), the dilation rates are 1, 2, 4, 8. So the features are extracted by different dilatation rates, which are processed in separate branches and then fused to produce the final result.

4.3. Filter Response Normalization (FRN) used in decoder

4.3.1. Final road segmentation maps generation by decoder modules

The up-sampling path performs symmetric operations. Each of the four groups has three convolutions, including two 1×1 convolution layers and one transposed convolution layer, each followed by an FRN layer and a ReLU. Where four transposed layers up-sampling the feature maps to the size of 64×64, 128×128, 256×256, and 512×512 respectively, the ReLU activation function is employed to alleviate the problem of disappearing gradient, and the FRN layer is used to remove the scaling effect of both the filter weights and pre-activations. In order to recover more segmentation details, such as texture, boundary, and spatial information, the high-level features with the low-level features are fused by addition. Specifically, in each group, the decoder features are first up-sampled by a factor of 4 through transposed convolution layers and then added to the corresponding low-level features from the encoder with the same spatial resolution. Then the feature maps followed by another up-sampling through the 4×4 transposed convolution layers, it returns to the original input image size, i.e., 512×512 pixels, and the number of channels is 32. Finally, a sigmoid classifier is used to classify the road target and background.

4.3.2. Filter Response Normalization (FRN)

Normalization of training data can accelerate the learning and make the training of deep neural network structure possible. Under the limited hardware environment, due to the need for more computation, the training batch size of the semantic segmentation network is usually smaller than that of the image classification network (Zhang and Wang, 2019). When trained with small batch sizes, Batch Normalization (BN) (Ioffe and Szegedy, 2015) layers degrade significantly, leading to poor training results. Several methods have been proposed to solve this shortcoming, such as Batch Renormalization (BR), Layer Normalization (LN), and Group Normalization (GN), etc. However, these methods still have some limitations, such as requiring considerable GPU infrastructure or performing worse than BN for large mini-batches. Therefore, a more stable batch independent regularization method is needed to obtain the performance gains.

Filter Response Normalization (FRN) (Zhao et al., 2017) is a newly proposed normalization method that eliminates the dependence on other batch samples or channels of the same sample. These characteristics enable the network to obtain better results with stronger baselines, which makes FRN outperforms other normalization methods in road extraction.

After a convolution operation, the filter response of each layer is a four-dimensional vector denoted as $[B, W, H, C]$, where B is the mini-batch size, W , H are the width and height size of the feature maps respectively, C is the channel size. The FRN layer can be written as follows:

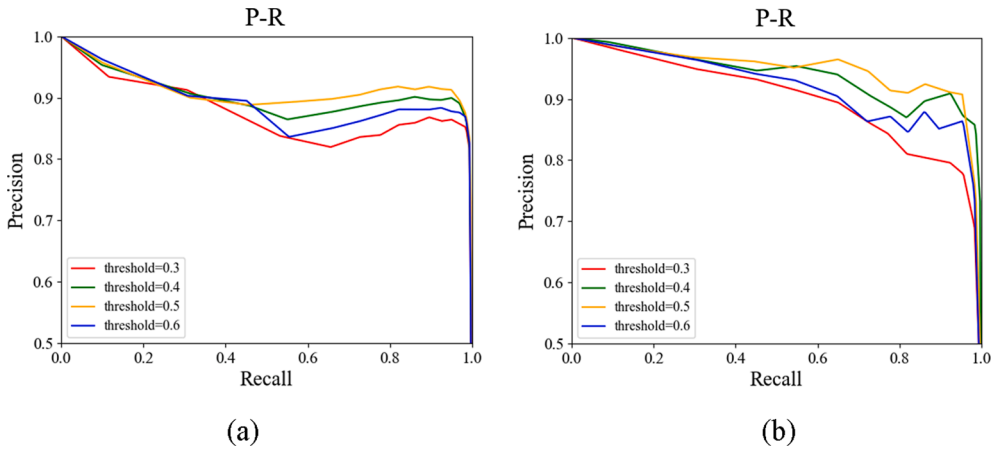

Fig. 7. The precision–recall curves on (a) DeepGlobe Roads Dataset and (b) SpaceNet Roads Dataset.

$$\nu^2 = \sum_i x_i^2 / N \qquad \widehat{x}_i = \frac{x_i}{\sqrt{\nu^2 + \varepsilon}} \tag{5}$$

where x is the feature map in a neural network layer with the index i, $N = W \times H$, and ε is a small positive constant to prevent division by zero errors.

After normalization, affine transformation is carried out to undo the effect of the network on normalization:

$$y_i = \gamma \widehat{x}_i + \beta = \gamma \frac{x_i}{\sqrt{\nu^2 + \varepsilon}} + \beta \tag{6}$$

where γ and β are learned parameters. Finally, FRN augments ReLU through a learned threshold τ to yield Thresholded Linear Unit (TLU) defined as:

$$z_i = max(y_i, \tau) \tag{7}$$

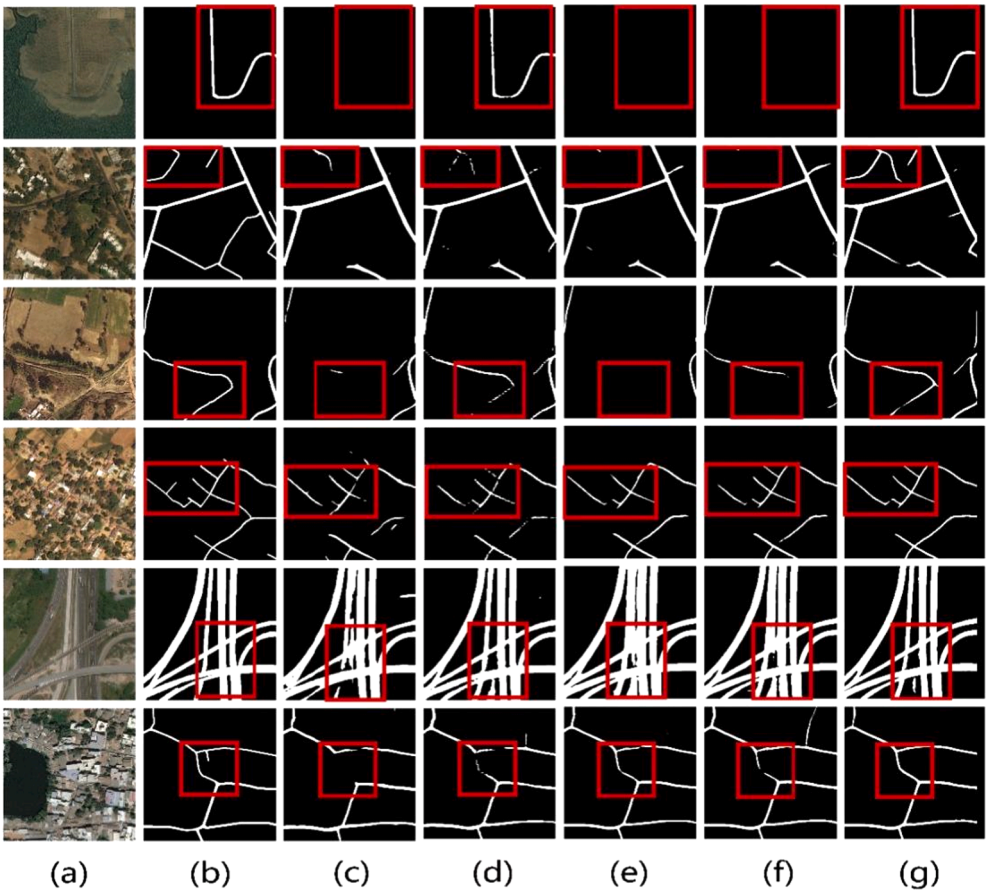

Fig. 8. Road extraction results using the DeepGlobe Roads Dataset. (a) Satellite imagery. (b) Ground truth. (c) U-Net. (d) Deeplabv3. (e) D-LinkNet. (f) HsgNet. (g) Ours.

5. Experiments

5.1. Implementation details

To improve the generalization of the network, data augmentation was employed, including random flip ambitious color jittering, image shifting, scaling. Pytorch (Paszke et al., 2017) was used to implement our model. Adam (Kingma and Ba, 2014) was adopted as an optimizer with the batch size of 4, and Binary Cross Entropy (BCE) + dice coefficient loss as the loss function. The learning rate was initially set to be 2e-4 and reduced by a factor of 5 for 3 times while observing the training loss decreasing slowly. All models were trained on 4 NVIDIA RTX2080 GPUs.

5.2. Evaluation metrics

The purpose of the road extraction from VHR satellite imagery, which has often been regarded as a semantic segmentation problem. For pixel-based road extraction, the number of true positive pixels forms the intersecting area, and the sum of true positive (TP), false positive (FP), and false negative (FN) pixels forms the union. The pixel-wise intersection over union (IoU) score given in Eq. (8) below:

$$IOU = TP/(TP+FP+FN) \tag{8}$$

where *TP* means the number of pixels correctly extracted as road, *FP* means the number of other object pixels extracted as road, *FN* means the number of road pixels extracted as other objects.

F1 score is given in Eq. (11) is primarily used to compare the performance of extraction networks. It is the weighted average of Precision given in Eq. (9) and Recall given in Eq. (10).

$$Precision = TP/(TP+FP) \tag{9}$$

$$Recall = TP/(TP+FN) \tag{10}$$

$$F1 = 2 \times (Precision \times Recall)/(Precision + Recall) \tag{11}$$

5.3. Results

5.3.1. Experiment using the DeepGlobe Roads Dataset

In this paper, the proposed GCB-Net was compared with four road extraction methods based on semantic segmentation, including U-Net (Ronneberger et al., 2015), Deeplabv3 (Chen et al., 2017), D-LinkNet (Zhou et al., 2018), and HsgNet (Xie et al., 2019). U-Net has become one of the most popular baselines for road extraction, which transmitting the encoder-generated feature maps to the corresponding decoder. Deeplabv3 adopts atrous convolution to capture multi-scale context. D-LinkNet is the winner of the CVPR DeepGlobe Road Extraction Challenge in 2018. This uses the LinkNet architecture and adds dilated convolution in the central part. HsgNet takes LinkNet as basic architecture and inserts an attention block to preserve global context semantic information. The classifier provides the road/non-road probabilities of the map pixels, and the thresholds are needed in specific predictions to determine the class of pixels. Therefore, the precision–recall curves are used to evaluate our model. Fig. 7 shows the precision–recall curves for the four thresholds we evaluated on the two datasets. The recall value is increasing. The more convex the precision–recall curve is to the upper right, the better the effect is. As shown in Fig. 7, the best results are achieved for both data sets at a threshold value of 0.5.

As shown in Fig. 8, in the first row, the spectral features of the road and the background are very similar, while the second row shows the areas are shaded by woods. In both cases, it is almost impossible to distinguish between roads and non-roads using only a simple combination based on semantic and appearance information (Fig. 8c, d, and e), whereas Fig. 8 shows extraction results from the proposed GCB-Net can extract complete road information. In other challenging road scenes, such as rural narrow road without clear boundary (the third row of Fig. 8), complex road intersection (the fourth and sixth rows of Fig. 8), shadows of buildings and trees (the fifth row of Fig. 8), as well as urban roads with rich spectral and textural features (the seventh row of Fig. 8), GCB-Net achieved good visual performance.

Table 1
Comparison of different road network extraction methods on public road data set. The best results on each data set are highlighted in boldface.

Model Name	DeepGlobe Roads Dataset		SpaceNet Roads Dataset	
	road IoU (%)	F1 (%)	road IoU (%)	F1 (%)
U-Net	65.76	77.75	61.97	69.81
Deeplabv3	66.73	78.08	64.76	72.50
D-LinkNet	68.83	79.80	65.07	72.33
HsgNet	69.29	80.34	66.07	73.23
Ours	**70.80**	**81.54**	**69.05**	**76.33**

As shown in Table 1, the proposed GCB-Net achieved the largest IoU of 70.80% with an F1 score of 81.54%. Compared with other methods, including U-Net, Deeplabv3, D-LinkNet, and HsgNet, the proposed method achieves the best performance in all indicators. In detail, when compared with U-Net, the GCB-Net achieves a large increase in IoU and F1 score of 5.04% and 3.79%, respectively. Deeplabv3 outperforms U-Net by 0.97% in IoU, indicating that the multi-scale context aggregation is effective. As shown in Table 1, compared with Deeplabv3, D-LinkNet and GCB-Net employing multi-parallel dilated convolution layers further improves the performance by 2.10% and 4.07%, respectively. Besides, compared with D-LinkNet, the use of global context information brings increments of 0.46% and 1.74% in IoU with respect to HsgNet and GCB-Net.

5.3.2. Experiment using the SpaceNet Roads Dataset

To further demonstrate the generality of the proposed GCB-Net, the proposed method was also compared with U-Net, Deeplabv3, D-LinkNet, and HsgNet on the SpaceNet Roads Dataset. The visual assessment of these methods is shown in Fig. 9. In these experimental areas, compared with the CNNs extract information from the local area only, the advantages of the proposed network that model global context relationships are illustrated, especially in complex road scenarios. As shown in rows 1 and 3 in Fig. 9, the road is not well distinguished from the surrounding features in the results of the comparison model, while the proposed model maintains a clearer boundary. In rows 2 and 4 in Fig. 9, GCB-Net gets the most accurate intersection information. These results indicate that GCB-Net can maintain more complete road skeletons and clearer road boundaries.

The accurate results on the SpaceNet Roads Dataset are shown in Table 1. The proposed GCB-Net achieved the best performance, with road IoU and F1 scores improving by 7.08% and 6.52%, respectively. Compared with Deeplabv3, D-LinkNet, and HsgNet, GCB-Net achieves a significant improvement of 4.29%, 3.98%, and 2.98% in road IoU, respectively. These suggest the importance of capturing long-range spatial relations for detecting road objects.

5.3.3. Experiment using the CHN6-CUG Roads Dataset

Due to the changes of sensors used or the difference of road radiation, the model segmentation effect is often poor when using images of non-training areas. To test the generalization of the proposed model, the DeepGlobe Roads Dataset was used as the source domain and the CHN6-CUG Roads Dataset was used as the target domain. The fine-tuned pre-trained models (trained on source dataset) were transferred to the target data set for direct testing. Fig. 10 shows the qualitative results of the proposed GCB-Net and other networks for DeepGlobe Roads Dataset to CHN6-CUG Roads Dataset. As shown in Fig. 10 in detail, the roads in CHN6-CUG Roads Dataset poses some unique features like completely blocked roads, extremely narrow paths, heavily trafficked and elevated roads. These features make it more difficult for the road extraction

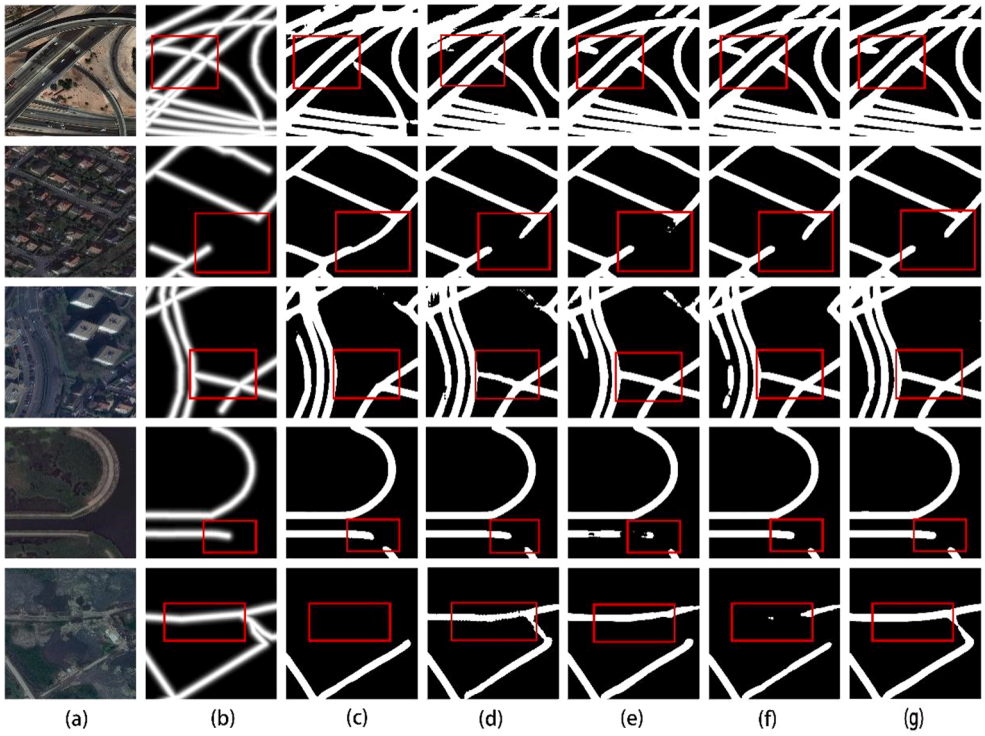

Fig. 9. Road extraction results using the SpaceNet Roads Dataset. (a) Satellite imagery. (b) Ground truth. (c) U-Net. (d) Deeplabv3. (e) D-LinkNet. (f) HsgNet. (g) Ours.

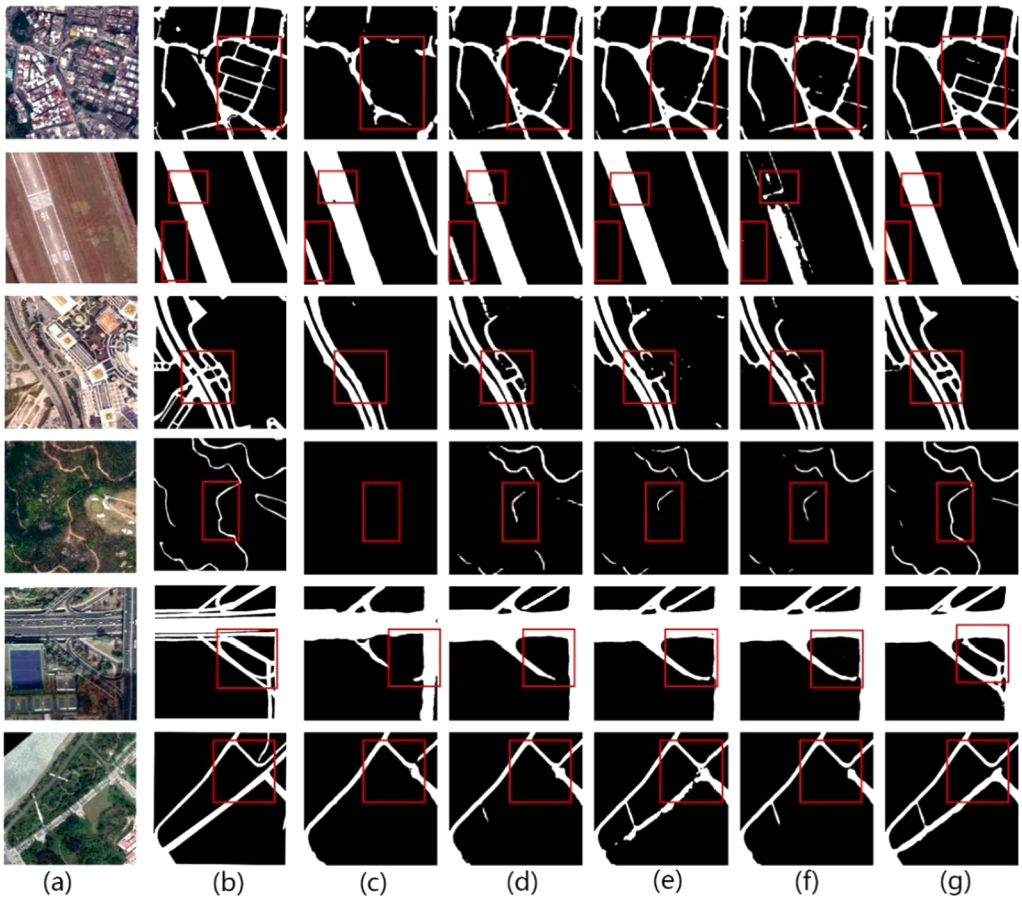

Fig. 10. Road extraction results using the CHN6-CUG Roads Dataset. (a) Satellite imagery. (b) Ground truth. (c) U-Net. (d) Deeplabv3. (e) D-LinkNet. (f) HsgNet. (g) Ours.

Table 2
Comparison of different road network extraction methods on CHN6-CUG Roads Dataset. The best results on each data set are highlighted in boldface.

Model Name	China Roads Dataset	
	road IoU (%)	F1 (%)
U-Net	47.68	61.13
Deeplabv3	54.50	66.96
D-LinkNet	55.74	68.59
HsgNet	57.68	70.35
Ours	**60.44**	**72.70**

models to adapt to such a data set. The proposed GCB-Net shows superior performance compared to other state-of-the-art models. This indicates that the features learned by this module have strong robustness and can be effectively transferred to the new baseline for road extraction.

As shown in the quantitative comparison listed in Table 2, GCB-Net produced the best results of the comparison method, reaching road IoU and F1 scores value of 60.44% and 72.70%, respectively. In terms of road IoU, the proposed GCB-Net exceeds 12.76%, 5.94%, 4.7%, and 2.76% respectively compared with U-Net, Deeplabv3, D-LinkNet, and HsgNet. GCB-Net achieves better performance even when the target domain lack labeled data. This demonstrates that it is effective to long-range feature interdependencies for domain adaptation.

Fig. 11 shows the road mapping results for every study area in CHN6-CUG Roads Dataset. The validation zones in Wuhan and Shanghai with relatively complex road scenes also achieved satisfactory performance. The typical subsets in Fig. 11 (such as Fig. 11b in Beijing, Fig. 11b in Shanghai and Wuhan as well as Fig. 11b and Fig. 11c in Macao) revealed

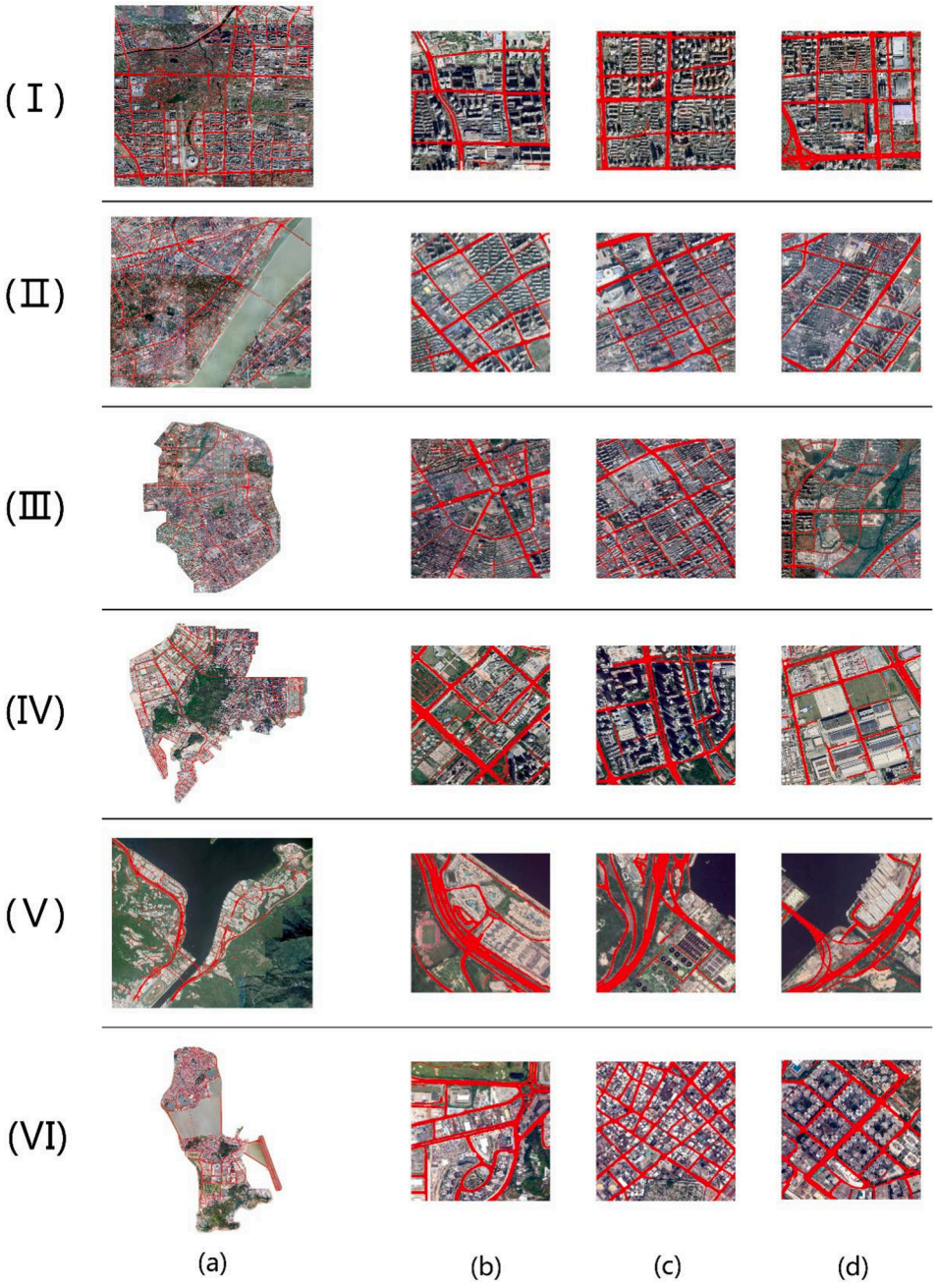

Fig. 11. Road mapping results for (I) Beijing, (II) Wuhan, (III) Shanghai, (IV) Shenzhen, (V) Hong Kong and (VI) Macao study areas: (a) road maps overlaid with validation areas. (b)-(d) the road mapping results from typical subsets of each study areas.

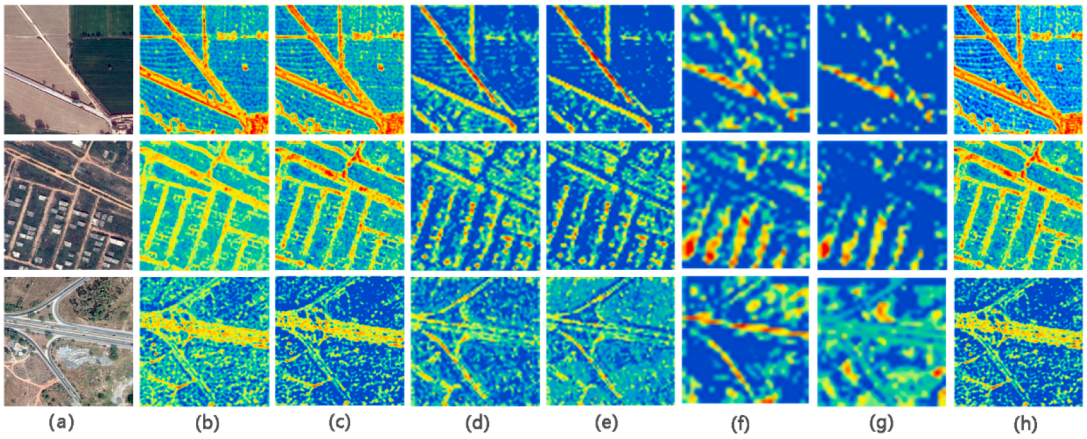

Fig. 12. The visualization of different level features: (a) input images, (b) before adding the first GCA block, (c) after adding the first GCA block, (d) before adding the second GCA block, (e) after adding the second GCA block, (f) before adding the third GCA block, (g) after adding the third GCA block, (h) after the last decoder block.

the proposed method performed well in extracting roads of various urban scenes. The validation zones in Wuhan and Macao both show well extraction results, but are very different from each other in the geographical environment. Due to the high heterogeneity of roads in Wuhan, the baseline network trained on the DeepGlobe Roads Dataset is difficult to be segmented. The geometric features of the roads in Macao are clearer, and the model transferred from the source domain can achieve a better segmentation effect.

6. Discussion

6.1. Feature representations for roads

In this section, taking the DeepGlobe Roads Dataset as an example, the feature representations of the proposed GCB-Net are discussed in the aspects of robustness and connectivity. As shown in Fig. 8, different brightness levels represent the sizes of activation values. By modeling the global context relationship, GCB-Net can achieve accurate interpretation on diverse road scenes. To better understand how the GCB-Net model extracts the road, the feature representations of the four different stages are shown in Fig. 12. After adding the GCA block, the learned road feature boundary is clearer and the redundant information is reduced. As the feature level increases, semantic information becomes more and more abstract, while road boundary becomes more and more blurred. As shown in Fig. 12, Overall, it demonstrates that the proposed GCB-Net has strong robustness and reasoning ability.

6.2. Evaluation of model transferability performance

Large-scale road-mapping requires the model to be quickly constructed and well transferred. To validate the spatial generalization capability of the proposed method, the fine-tuned pre-trained models gained from training areas were applied to the road mapping in China. As presented by Fig. 2, the road type of the source domain (DeepGlobe Roads Dataset) and the target domain (CHN6-CUG Roads Dataset) is not quite consistent. There are also great differences between China's internal roads in terms of geometric features, spectral features, context features, etc.

Qualitative results are presented in Fig. 11. Although road scenes from CHN6-CUG Roads Dataset are more easily misidentified, GCB-Net obtains more accurate results. Compared with other methods, no matter in the case of little difference in spectral characteristics between road and background, or the complex road intersection, etc., GCB-Net can still show better connectivity and noise resistance. This demonstrates that GCB-Net can effectively solve the problem of visual ambiguity by exploiting global relations. Table 2 shows the accuracy evaluation index of the study area. GCB-Net achieves the best performance of 60.44. This precision can meet the precision requirement of rapid model construction in practical application. Both qualitative and quantitative evaluation confirmed that the proposed method has strong generalization ability in the spatial transfer.

Table 3
Ablation study with different components combinations on DeepGlobe Roads Dataset.

Baseline	FRN	MDC	GCA	road IoU	F1
√				61.35	76.04
√	√			65.73	78.18
√	√	√		66.43	78.55
√	√		√	69.56	80.84
√	√	√	√	70.80	81.54

6.3. Ablation study

In this section, the ablation study was conducted to verify the effectiveness of each key components designed in the proposed model. The ablation analysis was performed on the DeepGlobe Roads dataset. The Linknet-34 (Ronneberger et al., 2015) was adopted as the baseline model, and then each module was added progressively. The results improved by 4.38% after FRN replaced the BN in the baseline. This demonstrates the effectiveness of the FRN layer design for road extraction. From Table 3, the addition of the Multi-parallel Dilated Convolution (MDC) module improves the baseline from 65.73 to 66.43 in terms of road IoU. This implies that the MDC module improves the occurrence of the wrong recognition and road connectivity problems. The road IoU is largely improved by 8.21% compared with the basic model after adding the Global Context-aware (GCA) module and replacing the BN in baseline with FRN. Finally, the combination of the MDC module and the GCA module has already achieved the best performance. This demonstrates the necessity of each component for the proposed model to obtain the best road extraction results.

7. Conclusions

This study proposed a Global Context-aware and Batch-independent Network (GCB-Net) to achieve road extraction. In the GCB-Net, the encoder part combined Global Context-Aware (GCA) blocks to enhance the global spatial relationship, the multi-parallel dilated convolution was applied to extract multi-scale road features, and by using FRN to

makes the model achieved better performance. The proposed method was first evaluated our model on two public datasets (DeepGlobe Roads Dataset and SpaceNet Roads Dataset). The results showed the advantages of the proposed method in the topological connectivity of the road surface. To drive the state-of-the-art models, CHN6-CUG Roads Dataset is constructed, which provided the research community with a road benchmark that covers a much wider range of China. Fine-tune method was used to transfer the knowledge learned from the DeepGlobe Roads Dataset to the CHN6-CUG Roads Dataset. The experimental results confirmed that the proposed framework performs better in handling complex roads compared with the other state-of-the-art approaches. Future work will aim to combine GAN-based methods for semi-supervised road extraction.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the editor, associate editor, and anonymous reviewers for their helpful comments and advice. This work was supported in part by the National Natural Science Foundation of China under Grant No. 41901306, in part by the Open Research Fund of National Earth Observation Data Center (NODAOP2020006), and in part by the Fundamental Research Funds for the Central Universities, China University of Geosciences, Wuhan, China under Grant No. G1323519214.

References

Bonafilia, D., Gill, J., Basu, S., Yang, D., 2019. Building High Resolution Maps for Humanitarian Aid and Development with Weakly-and Semi-Supervised Learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9.

Buslaev, A., Seferbekov, S., Iglovikov, V., Shvets, A., 2018. Fully convolutional network for automatic road extraction from satellite imagery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 207–210.

Bastani, F., He, S., Abbar, S., Alizadeh, M., Balakrishnan, H., Chawla, S., Madden, S., DeWitt, D., 2018. Roadtracer: Automatic extraction of road networks from aerial images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4720–4728.

Batra, A., Singh, S., Pang, G., Basu, S., Jawahar, C.V., Paluri, M., 2019. Improved road connectivity by joint learning of orientation and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10385–10393.

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

Claussmann, L., Revilloud, M., Gruyer, D., Glaser, S., 2019. A review of motion planning for highway autonomous driving. IEEE Trans. Intell. Transp. Syst. 21 (5), 1826–1848.

Chen, L., Zhu, Q., Xie, X., Hu, H., Zeng, H., 2018. Road extraction from VHR remote-sensing imagery via object segmentation constrained by Gabor features. ISPRS Int. J. Geo-Inf. 7 (9), 362.

Cheng, G., Wang, Y., Xu, S., Wang, H., Xiang, S., Pan, C., 2017. Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network. IEEE Trans. Geosci. Remote Sens. 55 (6), 3322–3337.

Chen, Z., Xu, Q., Cong, R., Huang, Q., 2020. Global Context-aware progressive aggregation network for salient object detection, arXiv preprint arXiv: 2003.00651.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp. 801–818.

Chaurasia, A., Culurciello, E., 2017. Linknet: Exploiting encoder representations for efficient semantic segmentation. In: Proceedings of the IEEE Visual Communications and Image Processing (VCIP), pp. 1–4.

Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation, arXiv preprint arXiv: 1706.05587.

Chen, Y., Kalantidis, Y., Li, J., Yan, S., Feng, J., 2018. Aˆ 2-nets: Double attention networks. In: Advances in neural information processing systems, pp. 352–361.

Dai, J., Zhu, T., Zhang, Y., Ma, R., Li, W., 2019. Lane-Level Road Extraction from High-Resolution Optical Satellite Images. Remote Sens. 11 (22), 2672.

Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C., 2020. Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data. ISPRS J. Photogramm. Remote Sens. 162, 94–114.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raska, R., 2018. Deepglobe 2018: A challenge to parse the earth through satellite images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 172–17209.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255.

Gao, X., Sun, X., Zhang, Y., Yan, M., Xu, G., Sun, H., Jiao, J., Fu, K., 2018. An end-to-end neural network for road extraction from remote sensing imagery by multiple feature pyramid network. IEEE Access 6, 39401–39414.

Huang, B., Zhao, B., Song, Y., 2018. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. Remote Sens. Environ. 214, 73–86.

Han, X., Lu, J., Zhao, C., Li, H., 2018. Fully Convolutional Neural Networks for Road Detection with Multiple Cues Integration. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pp. 1–9.

He, H., Yang, D., Wang, S., Wang, S., Li, Y., 2019. Road extraction by using atrous spatial pyramid pooling integrated encoder-decoder network and structural similarity loss. Remote Sens. 11 (9), 1015.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141.

Hu, H., Zhang, Z., Xie, Z., Lin, S., 2019. Local relation networks for image recognition. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3464–3473.

He, J., Deng, Z., Zhou, L., Wang, Y., Qiao, Y., 2019. Adaptive pyramid context network for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7519–7528.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

Levin, N., Duke, Y., 2012. High spatial resolution night-time light images for demographic and socio-economic studies. Remote Sens. Environ. 119, 1–10.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440.

Liu, J., Qin, Q., Li, J., Li, Y., 2017. Rural road extraction from high-resolution remote sensing images based on geometric feature inference. ISPRS Int. J. Geo-Inf. 6 (10), 314.

Lu, X., Zhong, Y., Zheng, Z., Liu, Y., Zhao, J., Ma, A., Yang, J., 2019. Multi-scale and multi-task deep learning framework for automatic road extraction. IEEE Trans. Geosci. Remote Sens. 57 (11), 9362–9377.

Maboudi, M., Amini, J., Malihi, S., Hahn, M., 2018. Integrating fuzzy object based image analysis and ant colony optimization for road extraction from remotely sensed images. ISPRS J. Photogramm. Remote Sens. 138, 151–163.

Mnih, V., Hinton, G.E., 2010. Learning to detect roads in high-resolution aerial images. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 210–223.

Máttyus, G., Luo, W., Urtasun, R., 2017. Deeproadmapper: Extracting road topology from aerial images. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3438–3446.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z, Desmaison, A, Antiga, L, Lerer, A., 2017. Automatic differentiation in pytorch.

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 (MICCAI), vol. 9351. Springer, pp. 234–241.

Sghaier, M.O., Lepage, R., 2015. Road extraction from very high resolution remote sensing optical images based on texture analysis and beamlet transform. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 9 (5), 1946–1958.

Sun, T., Chen, Z., Yang, W., Wang, Y., 2018. Stacked U-Nets with multi-output for road extraction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 202–206.

Singh, S., Shrivastava, A., 2019. Evalnorm: Estimating batch normalization statistics for evaluation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3633–3641.

Singh, S., Krishnan, S., 2020. Filter response normalization layer: Eliminating batch dependence in the training of deep neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 11237–11246.

Shamsolmoali, P., Zareapoor, M., Zhou, H., Wang, R., Yang, J., 2020. Road segmentation for remote sensing images using adversarial spatial pyramid networks. IEEE Trans. Geosci. Remote Sens.

Tan, X., Xiao, Z., Wan, Q., Shao, W., 2020. Scale Sensitive Neural Network for Road Segmentation in High-Resolution Remote Sensing Images. IEEE Geosci. Remote Sens. Lett. 1–5.

Tao, C., Qi, J., Li, Y., Wang, H., Li, H., 2019. Spatial information inference net: Road extraction using road-specific contextual information. ISPRS J. Photogramm. Remote Sens. 158, 155–166.

Van Etten, A., Lindenbaum, D., Bacastow, T. M., 2018. Spacenet: A remote sensing dataset and challenge series, arXiv preprint arXiv: 1807.01232.
Wu, S., Du, C., Chen, H., Xu, Y., Guo, N., Jing, N., 2019. Road extraction from very high resolution images using weakly labeled OpenStreetMap centerline. ISPRS Int. J. Geo-Inf. 8 (11), 478.
Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7794–7803.
Wang, Y., Seo, J., Jeon, T., 2019. NL-LinkNet: Toward Lighter but More Accurate Road Extraction with Non-Local Operations, arXiv preprint arXiv:1908.08223.
Xie, Y., Weng, Q., 2016. Updating urban extents with nighttime light imagery by using an object-based thresholding method. Remote Sens. Environ. 187, 1–13.
Xie, Y., Miao, F., Zhou, K., Peng, J., 2019. HsgNet: A Road Extraction Network Based on Global Perception of High-Order Spatial Information. ISPRS Int. J. Geo-Inf. 8 (12), 571.
Yang, X., Li, X., Ye, Y., Lau, R.Y., Zhang, X., Huang, X., 2019. Road detection and centerline extraction via deep recurrent convolutional neural network u-net. IEEE Trans. Geosci. Remote Sens. 57 (9), 7209–7220.
Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., Gao, J., Zhang, L., 2020. Deep learning in environmental remote sensing: Achievements and challenges. Remote Sens. Environ. 241, 111716.
Yue, K., Sun, M., Yuan, Y., Zhou, F., Ding, E., Xu, F., 2018. Compact generalized non-local network. In: Proceedings of the Advances in Neural Information Processing Systems (NIPS), pp. 6510–6519.
Zhang, X., Xiao, P., Feng, X., Yuan, M., 2017. Separate segmentation of multi-temporal high-resolution remote sensing images for object-based change detection in urban area. Remote Sens. Environ. 201, 243–255.
Zang, Y., Wang, C., Cao, L., Yu, Y., Li, J., 2016. Road network extraction via aperiodic directional structure measurement. IEEE Trans. Geosci. Remote Sens. 54 (6), 3322–3335.
Zhang, C., Wei, S., Ji, S., Lu, M., 2019. Detecting large-scale urban land cover changes from very high resolution remote sensing images using cnn-based classification. ISPRS Int. J. Geo-Inf. 8 (4), 189.
Zhang, Z., Liu, Q., Wang, Y., 2018. Road extraction by deep residual u-net. IEEE Geosci. Remote Sens. Lett. 15 (5), 749–753.
Zhou, L., Zhang, C., Wu, M., 2018. D-LinkNet: LinkNet With Pre-trained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 182–186.
Zhang, Y., Xiong, Z., Zang, Y., Wang, C., Li, J., Li, X., 2019. Topology-aware road network extraction via multi-supervised generative adversarial networks. Remote Sens. 11 (9), 1017.
Zhang, X., Han, X., Li, C., Tang, X., Zhou, H., Jiao, L., 2019. Aerial image road extraction based on an improved generative adversarial network. Remote Sens. 11 (8), 930.
Zhou, M., Sui, H., Chen, S., Wang, J., Chen, X., 2020. BT-RoadNet: A boundary and topologically-aware neural network for road extraction from high-resolution remote sensing imagery. ISPRS Int. J. Geo-Inf. 168, 288–306.
Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2881–2890.
Zhang, Z., Wang, Y., 2019. JointNet: A common neural network for road and building extraction. Remote Sens. 11 (6), 696.
Zhu, Q., Zhong, Y., Zhang, L., Li, D., 2018. Adaptive deep sparse semantic modeling framework for high spatial resolution image scene classification. IEEE Trans. Geosci. Remote Sens. 56 (10), 6180–6195.
Zhu, Q., Li, Z., Zhang, Y., Guan, Q., 2020. Building Extraction from High Spatial Resolution Remote Sensing Images via Multiscale-Aware and Segmentation-Prior Conditional Random Fields. Remote Sens. 12 (23), 3983.