ORIGINAL RESEARCH PAPER

Road extraction from high resolution remote sensing image via a deep residual and pyramid pooling network

Yibo Han  |  Pu Han  |  Manlei Jia

Nanyang Institute of Big Data Research, Nanyang Institute of Technology, Nanyang, Henan, China

Correspondence
Pu Han, Nanyang Institute of Big Data Research, Nanyang Institute of Technology, Nanyang, Henan, China.
Email: hanpu@nyist.edu.cn

Funding information
Research on Key Technologies of Service Deployment, Discovery and Orchestration Optimization, Grant/Award Number: 202102210355

Abstract
The road extraction from high resolution remote sensing image is of great importance in a variety of applications. Recently, the abundant deep convolutional neural networks are proposed for road extraction task. However, the existing approaches lack suitable strategy to utilize multiple views road features for road extraction, which fails to extract road with smooth appearance and accurate boundary under complex scenes. To address this problem, the authors propose a novel deep residual and pyramid pooling network (DRPPNet) for extracting road regions from high resolution remote sensing image. The DRPPNet consists of three parts: deep residual network (DResNet), pyramid pooling module (PPM) and deep decoder (DD). Specially, the DResNet uses several residual blocks to extract deep road features from input images, which can enhance learning ability of DRPPNet and avoid gradient vanish. Then, PPM is proposed to fuse road features from multiple views and it aims to address disadvantage of single view feature. Finally, the DD is used to recover size of feature maps to input size. Extensive experiments on two challenging road datasets demonstrate that proposed method outperforms the state-of-the-art methods greatly on performance of road extraction task.

1 | INTRODUCTION

The road is the backbone and essential infrastructure in many domains, such as urban planning, geographic information system updating and traffic navigation [1–7]. Recently, with the rapid development of satellites, high resolution remote sensing image can be easily accessed and contains road information in more detail, which provides a promising way to extract road. However, manually labeling roads in high resolution remote sensing image will cost a lot of time and effort [8, 9]. To overcome this problem, many automatic methods, such as SVM [10, 11] and CRF [12], have been proposed to extract road from high resolution remote sensing image. Compared with manual extraction method, automatic road extraction methods are more efficient and economical.

With the development of deep learning, convolutional neural networks (CNNs) not only have been widely and successfully applied in semantic segmentation [13–15], but also play a key role in road extraction [16, 17]. For example, [18–20] first used CNN to extract road from high resolution remote sensing image. Their method has achieved great road extraction performance. Following this work, [21, 22] assumed that the relationship between multiple ground objects can improve performance of road extraction. Thus, they proposed a three channels CNN that considers the relationship among road, building and background to improve road performance. Furthermore, [23] proposed a junction-aware water flow model to extract road region from high resolution remote sensing image, and [24] introduced road skeletons as powerful information to enhance linear structure of road segments. These methods obtain satisfied performance and demonstrate effectiveness of CNNs in road extraction. However, their methods limit shallow network and there is still a huge space for improving road extraction performance.

In recent works, deep convolutional neural networks (DCNNs) have been proposed to further improve performance of road extraction. The first deep convolutional neural network is Vgg19 [25] that is a larger network and has the ability to learn appearance of more numerous and complicated objects. The behind philosophy of deep convolutional neural network is that deeper architectures can promote hierarchical feature
representation of visual data[26], which, in turn, improves extraction performance. Furthermore, [27] combined objects obtained from source high resolution remote sensing image segmentation for automatic road extraction. [28] combined conditional random field (CRF) and deep convolutional neural network to extract road regions from object-level. [29] used adaptive graph cuts with multiple features, such as spectral, spatial, and gradient for road extraction. These methods have favorable anti-noise properties and are widely applicable in road extraction. However, these kind of methods can easily confuse road features and adjacent objects that have similar shapes with road. Therefore, the road extraction results are not satisfactory when remote sensing image is extremely complex. More recently, the residual network has become the most important network in deep learning and has been used in road extraction task. The core of idea of residual network is to add shortcut connections that bypass two or more stacked convolution layers by performing identity mapping. [30] proposed deep residual learning framework to facilitate training. [31] proposed a novel network Unet that removes skip connection in fully convolutional network. The Unet concatenates feature maps of low-level detail information and high-level semantic information for semantic segmentation. Following this work, [32] combined strengths of residual learning and Unet for road extraction task.

It is well known that accurate road extraction relies on enriched and abstract road features. The authors find that the major issue of current DCNN-based road extraction methods is lack of suitable strategy to utilize multiple views road features. In reality, road is usually covered by many kinds of ground objects, like vehicles, trees, and shadows, which causes huge difficulties in extracting road features, such as shape and texture. In this situation, the existing DCNN-based road extraction methods fail to obtain satisfactory road extraction performance. In contrast, multiple views road features can provide road structure information in more detail, which is beneficial for model to extract road network from high resolution remote sensing image, especially for road covered by other objects.

To this end, the authors propose a novel end-to-end deep residual and pyramid network (DRPPNet) in this paper, which can extract deep road features from multiple views for road extraction task. Specially, DRPPNet is composed of deep residual network (DResNet), pyramid pooling module (PPM) and deep decoder (DD). The DResNet is build based on several residual blocks and it aims to extract deep road features from high resolution remote sensing image. Then, the pyramid pooling module (PPM) receives feature maps of DResNet and fuses road features from different views by using four maxpooling operations. Finally, the DD is used to recover feature map to same size as input image. By embedding PPM, DRPPNet has strong ability to obtain multiple views road features, which is important for achieving smooth and coherent road extraction results. Moreover, since residual unit is used to extract road features, DRPPNet can avoid gradient vanish effectively and obtain strong learning ability. The experiments for verifying performance of proposed method are conducted on two challenging road datasets, Cheng-roads and Mnih-Roads. The contribution lies in following three aspects:

1. By combining residual unit and pyramid pooling, the authors propose a deep residual and pyramid network for road extraction task, which can achieve more consistent road extraction results under complex scenes.
2. Benefit to PPM, DRPPNet can obtain multiple views road features that are beneficial for DRPPNet to extract road network covered by other objects.
3. The DRPPNet is evaluated with two challenging road datasets and obtains better performance than other state-of-the-art road extraction methods.

The rest of this paper is organized as follows. Section 2 introduces background and related works. Section 3 presents technical details. Section 4 presents experiments and evaluation. Section 5 concludes this paper.

2 BACKGROUND AND RELATED WORKS

The methods of road extraction from remote sensing image have been a hot research in recent years. In this section, the authors discuss innovated works related to our method.

The road extraction methods mainly depend on pixel-level or superpixel-level classification. [33] performed road extraction in two steps. In first steps, the support vector machine (SVM) was employed to classify image into different categories. Then, the road region was segmented into homogeneous objects using region growing technique. [34] first segmented input image using a traditional k-means clustering. Then the fuzzy logic classifier was used to automatically extract road region. To some extent, this method can distinguish road area from parking point. [35] presented an automatic road extraction method from remote sensing image using locally excitatory globally inhibitory oscillator network. Then, [36] proposed a multistage framework for road extraction from remote sensing image. Their work is based on probabilistic SVM and salient features. [37] proposed a convolutional neural network-based algorithm to learn features for road extraction.

Many researchers have provided various strategies to describe spatial features[38–41]. [42] proposed recurrent neural networks (RNN) based on conditional random field and convolutional neural network. The RNN is an end-to-end network and achieves great success in learning high level features. However, the fixed-size receptive field in RNN limits the performance in obtaining fine segmentation maps. [43] first proposed a deconvolution network that was demonstrated to be a great network in obtaining detailed structures. The core of deconvolution network is to recover resolution feature map same size as input maps. After that, [44] proposed a SegNet that is an encoder-decoder architecture. The decoder enlarged feature maps according to indices from corresponding encoder layer. Recently, a patch-based deep neural network method was proposed by [19] to extract road network from high resolution remote sensing image. [45] proposed a high order CRF model to extract road network. The road prior was obtained by high order cliques that connects sets of super-pixels along straight line segments in their method.
3 | METHODOLOGY

In this study, the authors design a spatial deep residual and pyramid pooling network (DRPPNet) to better model road features from different views. The DRPPNet can explicitly explore different views road features by pyramid pooling module. In this following, the authors first describe the basic structure of proposed DRPPNet. Then, the learning algorithm is introduced to show how to train DRPPNet with an end-to-end strategy.

3.1 Deep residual and pyramid pooling network

The DRPPNet is an encoder–decoder network, and Figure 1 shows architecture of DRPPNet. As seen from Figure 1, the input to DRPPNet is RGB channels of high resolution remote sensing image, whereas the output is road map of corresponding remote sensing image. DRPPNet is composed of three modules, deep residual network (DResNet), pyramid pooling module (SPPM) and deep decoder (DD). In the following, the authors describe design in detail.

3.1.1 Deep residual module

For DResNet module, it aims to obtain general feature about road in remote sensing image. The core of DResNet is residual unit assuming that multiple nonlinear layers can asymptotically approximate residual function. The specific formula of residual unit can be written as

\[ \begin{aligned}
  y_j &= h(x_j) + F(x_j, W_j) \\
  x_{j+1} &= f(y_j)
\end{aligned} \tag{1} \]

where \( x_j \) and \( x_{j+1} \) are input and output of \( j^{\text{th}} \) residual unit. \( F \) is residual function, \( f(y_j) \) is activation function, and \( h(x_j) \) is an identity mapping function, a typical one is \( h(x_j) = x_j \). The add layers of residual unit are constructed as identity mappings. In this way, the DResNet should have training error no greater than its shallower counterpart. In addition, with the help of residual unit, when the identity mappings achieve optimal, the solvers may simply drive weights of multiple nonlinear layers toward zero to approach identity mappings.

Table 1 visualizes structure of DResNet that is inspired by Vgg networks[25]. Specifically, for Conv1, the kernel size is 7x7, the stride is 1 and the padding size is 4. To keep the size of output features as same with input one, we reshape convolutional results to 128 x 128. For example, let \( m \) is feature map after convolutional operation, we obtain output feature map by \( m = m(0 : 128, 0 : 128) \). There are 4-level residual unit in proposed DResUnet and the convolutional layers mostly have 3 x 3 filters. In each residual unit, instead of using pooling operation to downsample feature map size, a stride of 2 is applied to first

| Convolutional block | Kernel size | Stride | Pad | Output size |
|---------------------|-------------|--------|-----|-------------|
| Input Image         | -           | -      | -   | 128 x 128 x 3 |
| Conv1               | 7 x 7, 64   | 1      | 4   | 128 x 128 x 64 |
| conv2_x             | \[1 x 1 64\] | 1      | 0   | 64 x 64 x 256 |
|                     | \[3 x 3 64\] x 3 | 2/1 | 1   | 64 x 64 x 256 |
|                     | \[1 x 1 256\] | 1      | 0   | 64 x 64 x 256 |
| conv3_x             | \[1 x 1 128\] | 1      | 0   | 32 x 32 x 512 |
|                     | \[3 x 3 128\] x 4 | 2/1 | 1   | 32 x 32 x 512 |
|                     | \[1 x 1 512\] | 1      | 0   | 32 x 32 x 512 |
| conv4_x             | \[1 x 1 256\] | 1      | 0   | 16 x 16 x 1024 |
|                     | \[3 x 3 256\] x 6 | 2/1 | 1   | 16 x 16 x 1024 |
|                     | \[1 x 1 1024\] | 1      | 0   | 16 x 16 x 1024 |
| conv5_x             | \[1 x 1 512\] | 1      | 0   | 8 x 8 x 2048 |
|                     | \[3 x 3 512\] x 3 | 2/1 | 1   | 8 x 8 x 2048 |
|                     | \[1 x 1 2048\] | 1      | 0   | 8 x 8 x 2048 |
convolution block to reduce feature map by half. The number of residual units is gradually increasing, such as 64 in first residual unit, 128 in second residual unit and so on. The aim is to obtain more general and abstract context feature about road region.

The structure of DResUnet follows two simple design rules: (1) for same output feature size, the layers have same number of filters; and (2) if the feature map size is halved, the number of filters is doubled so as to preserve time complexity per layer. Benefit from DResUnet, the proposed DRPPNet has two great advantages in extracting road from high resolution remote sensing image. First, since the residual unit is introduced, the DRPPNet will ease to train. Second, the skip connections within a residual unit and between low levels and high levels of network will facilitate information propagation without degradation, which make it possible to design a neural network with much fewer parameters, however, could achieve comparable ever better performance on road extraction from the remote sensing image. The experimental results show that satisfactory performance can be obtained by our method.

3.1.2 Pyramid pooling module

Designing an effective method for extracting road is difficult due to complex morphology of road features in terms of shape and scale in remote sensing image. Therefore, a pyramid pooling module (PPM) is introduced, which encodes spatial context at multiple views, thereby making full use of feature extraction in a complex background. The overall architecture of PPM is shown in Figure 2.

The PPM module fuses road features from different views by using four maxpooling operations. The size of four maxpooling operations are different range from 1 × 1 to 6 × 6, which aims to obtain road feature from different views. The large size of maxpooling operations is global pooling highlighted in extracting road features from coarse level. By contrast, the small maxpooling operations is used to obtain fine feature of road. The four maxpooling operations are arranged in pyramid and can operate input feature maps into different regions. As a result, the output of different maxpooling operations in PPM has different sizes. In order to keep weight of general features, the 1 × 1 convolution layer is added into after maxpooling operations to reduce dimension of feature map. For example, if the size of input feature map is $N$, the output of 1 × 1 convolution layer will be $\frac{N}{4}$. Next, the upsample technology, such as bilinear interpolation, is applied to enlarge feature maps to same size as original input features. Finally, the authors use concatenated operation to fuse all feature maps from different maxpooling operations. In this way, the output of PPM can collect road features at a high semantic level from different views.

It is noted that the structure of PPM is not fixed and it can change with the size of feature maps that is produced by DResNet. In addition, the strides of maxpooling operations also is set by width road in remote sensing image. In here, the authors adapt four levels with size of 1 × 1, 2 × 2, 3 × 3 and 6 × 6 respectively for PPM. The experiment shows that our structure of PPM is reasonable for extracting road from remote sensing image.

3.1.3 Deep-decoder module

In proposed DRPPNet method, the pooling layer reduces size of input image. To upsample back to image resolution, the authors introduce a deep decoder (DD) as third part of DRPPNet. The DD consists of three unsampling layers and five convolutional layers. The role of decoder network is to recover feature map to same size as input image. The detail of decoder network is described as Table 2.

The DD module receives feature map produced by PPM as input and then one convolutional layer is applied. Following that, the upsampling operations is performed and the resulting output is enlarged by a factor of 2. The similar operations are performed twice for enlarging size of feature maps in
Decoder network. The last two convolutional layers are used to adjust size of feature map to input map. The output of DD is fed to a softmax for producing probability of pixel to be road class.

\[
p(y = k) = \frac{\exp(b(x))}{\sum_{k=1}^{K} \exp(b_k(x))}, \tag{2}
\]

where the \( b(x) \) is output of last convolutional layer at pixel \( x \). \( y \) is label of \( x \). \( p(y = k) \) is probability of pixel \( x \) being \( k \)th class. There are two channels in output images. The first channel includes probability that the road class is assigned to pixel at same location. And the second channel contains probability that the pixels are predicted to be background class.

### 3.2 Training

Given a set of training images and corresponding ground truth segmentations \( \{S_i, m_i\} \), our goal is to estimate parameters of DRPPNet, such that it can produce accurate and robust road areas in testing dataset. This is achieved through minimizing loss between segmentations generated by DRPPNet and ground truth \( m_i \). In this section, the authors use cross-entropy as loss function to train DRPPNet.

\[
L_{ce} = -\frac{1}{N} \sum_{j=1}^{N} (m_j \log p(y_j = 1|x_j) + (1 - m_j) \log p(y_j = 0|x_j)), \tag{3}
\]

where \( m_j = 1 \) denotes that the \( j \)th pixel is road class. Otherwise it is background class. The \( p(y_j = 1|x_j) \) is probability that the pixel \( x_j \) is classified to road class and it belongs to 0 to 1.

To train DRPPNet with end-to-end manners, \( L_{ce} \) need to be minimized with respect to DRPPNet parameter \( \theta \). The authors can obtain all derivative of loss in Equation (3) to different layers using chain rule, and then update \( \theta \) layer-by-layer with backpropagation strategies. For clarity, the authors only show derivative of loss to output of last convolutional layer in DD. The derivative of \( L_{ce} \) to \( b(x) \) of last convolutional layer in DD is defined as

\[
\frac{\partial L_{ce}}{\partial b(x)} = -\frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial b(x)} b(x) \log p(y_i = 1|x_i) + (1 - m_i) \log p(y_i = 0|x_i)), \tag{4}
\]

and

\[
\log p(y_i = 1|x_i) = \log \frac{1}{1 + e^{-b(x)}} = -\log(1 + e^{-b(x)}), \tag{5}
\]

where

\[
\log p(y_i = 0|x_i) = \log(1 - p(y_i = 1|x_i)),
\]

\[
= \log \left(1 - \frac{1}{1 + e^{-b(x)}}\right), \tag{6}
\]

\[
= \log(e^{-b(x)}/(1 + e^{-b(x)})),
\]

\[
= -b(x) - \log(1 + e^{-b(x)}).
\]

## 4 EXPERIMENT

To assess effectiveness of proposed method, the authors conduct experiments on two remote sensing image road datasets, that is, Cheng-Roads dataset and Mnh-Roads dataset. First, the authors provide a short description of both datasets and detail of experiment setup. Then, the authors compare proposed network with other state-of-the-art methods both visually and quantitatively.

### 4.1 Datasets and experiment setup

#### 4.1.1 Datasets

The authors employ two public datasets in this paper to evaluate proposed method. The information of these datasets is presented in the following.

The Cheng-Roads consists of 224 images and corresponding road maps is labeled by hand, which was published by Cheng et al.\cite{46}. One of 180 images are used for training and 14 images are used for validation and remaining 30 for testing. There are three spectral bands in each image (red, green and blue). The image has different spatial size range from 600 × 600 to 1200 × 1500 pixels. The spatial resolution of all images is
1.2 m². Figure 3 shows some examples in Cheng-Roads dataset, which includes street roads, rural roads in complex residential areas and highways in city areas.

The Mnih-Roads was published by Mnih [19], which consists of 1157 high resolution aerial images and covers an area of 2603 km². Each image is downloaded from google maps and the size is 1500 × 1500 pixels, covering an area of 2.25 km². The corresponding maps are collected from OpenStreetMap. The spatial resolution is 1.0 m². To train and evaluate network, the authors utilize 1108 images for training and build testing dataset with remaining 49 images. Figure 4 shows some examples in Mnih-Roads dataset. The dataset covers a wide variety of urban, suburban, and rural regions, which is a challenging road extraction dataset.

4.1.2 Experimental setup

The authors implement proposed method using tensorflow 1.13 and Python 3.6. Theoretically, the proposed DRPPNet can take arbitrary size image as input. However, the large size image needs to cost amount of GPU memory to store the feature maps. In this paper, the authors fix image size to 128 × 128 for training model. The mini-batch size is set to 8 and the authors train model on an GeForce GTX 980 Ti GPUs with 6GB memory. The SGD with initially learning rate 0.00001 is used to optimize the model. In addition, the learning rate drops by a factor of 0.1 every 100000 iterations. The network is trained until no further performance increase is observed.

4.2 Evaluation metrics

The road extraction task can be considered as binary classification, where road pixels are positive and the remaining non-road pixels are negatives. Let $TP$, $FP$, and $FN$ represent the true positive, false positive, and false negative, respectively. The authors present four benchmark metrics, that is, Precision (Pre), Recall (Rec) F1 and IoU to evaluate quantitative performance about road extraction.

The Pre measures fraction of matched areas in reference map. Rec represents percentage of matched road areas in segmentation map. F1 is an overall metric, which is computed using Pre and Rec. The IoU is percentage of overlapping area to union area between label map and predicted map. The authors calculate all metrics with following formula:

$$
Pre = \frac{TP}{TP + FP},
$$

$$
Rec = \frac{TP}{TP + FN},
$$

$$
F1 = \frac{2 \times Pre \times Rec}{Pre + Rec},
$$

$$
IoU = \frac{TP}{TP + FP + FN}.
$$
The examples from Minh-Roads dataset, where a column refers to each image. The first row is original images. The second row is label maps.

**Figure 4**

| Table 3 | The analysis of kernels size of PPM on Cheng-Roads and Mnih-Roads |
|---------|---------------------------------------------------------------|
|         | Cheng-Roads          | Mnih-Roads          |
|         | F1  | IoU  | F1  | IoU  |
| 1 × 1, 2 × 2, 3 × 3, 6 × 6 | 0.8772 | 0.7775 | 0.7432 | 0.5852 |
| 3 × 3, 4 × 4, 5 × 6, 6 × 6 | 0.8503 | 0.7453 | 0.7030 | 0.5456 |
| 1 × 1, 3 × 3, 4 × 4, 7 × 7 | 0.8810 | 0.7798 | 0.7160 | 0.5586 |

**4.3 Parameters analysis**

In this section, the authors analyze effects of different scales of PPM. The performance of PPM with different maxpooling size on Cheng-Roads and Mini-Roads is shown in Table 3. The best value is showed in bold in each metric term. In Table 3, the first column represents setup of kernel size of PPM. For example, taking first row parameters for example, the authors set four maxpooling kernels in PPM to 1 × 1, 2 × 2, 3 × 3 and 6 × 6, respectively. It can be found that if we set maxpooling kernels to small size, the DRPPNet shows high accuracy on Mnih-Roads. Specially, when 1 × 1 is used in PPM, the DRPPNet can obtain 0.7432 and 0.7160 in F1 on Mnih-Roads, which is higher 1% at least than without 1 × 1 maxpooling kernel. But when the maxpooling kernel is set to slight large size, such as third row, the authors find that DRPPNet can obtain 0.8810 in F1 on Cheng-Road, which is higher about 0.04% ~3% than other rows. The main reason for these results is that average road width in Cheng-Roads is larger than one in Mnih-Roads. In this case, the kernel with slight larger size is more fit to extract road in Cheng-Roads, but the small is more prefect for Minh-Roads. Considering the overall accuracy for road extraction, the authors set 1 × 1, 2 × 2, 3 × 3 and 6 × 6 to default parameter setting of PPM. The following experimental results show that satisfactory performance of DRPPNet can be obtained on Cheng-Roads and Mnih-Roads.

**4.4 Experimental results on Cheng-Roads**

This section illustrates details of our experiments on Cheng-Roads Dataset. To evaluate effectiveness of proposed DRPPNet, the authors compare road extraction methods with other state-of-the-art methods, including SegNet[44], Vgg19[25] and Unet[31]. All experiments adopt same postprocessing to obtain final road maps.

Table 4 lists quantitative performance of road extraction results with different state-of-the-art methods. The F1 and IoU is an overall metric than Pre and Rec. In Table 4, the first three rows are performance of three sampled images, and the last two rows are average performance of all methods in testing dataset. The best value is showed with bold in each metric term. It can be seen from Table 4 that our proposed DRPPNet achieves the best performance compared with other methods. Specially, the DRPPNet significantly increases Rec, with at least 4.5% improvement. Meanwhile, the average F1 of DRPPNet is 0.8772, which is far better 4.6% than of SegNet. And
TABLE 4  Performance evaluation with different compared methods on Cheng-Roads dataset. The best values are showed in bold and last two rows are average performance of all images in test set

|       | SegNet       | Vgg19         | Unet         | DRPPNet        |
|-------|--------------|---------------|--------------|----------------|
|       | Pre  | Rec  | F1  | Pre  | Rec  | F1  | Pre  | Rec  | F1  | Pre  | Rec  | F1  |
| Image 1 | 0.9327 | 0.8767 | 0.8271 | 0.8889 | 0.5071 | 0.6458 | 0.8440 | 0.6996 | 0.7650 | 0.8772 | 0.9087 | 0.8927 |
| Image 2 | 0.9194 | 0.8644 | 0.8156 | 0.8459 | 0.8070 | 0.8260 | 0.9527 | 0.5160 | 0.6626 | 0.9037 | 0.9068 | 0.9052 |
| Image 3 | 0.8578 | 0.8192 | 0.7840 | 0.8113 | 0.5706 | 0.6700 | 0.8405 | 0.7282 | 0.7803 | 0.8905 | 0.9107 | 0.9005 |
| Average | 0.8122 | 0.8314 | 0.8515 | 0.7762 | 0.8239 | 0.7993 | 0.8306 | 0.8334 | 0.8320 | 0.8579 | 0.8973 | 0.8772 |
| mean-IoU | 0.7262 | 0.6801 | 0.7289 | 0.7262 | 0.6801 | 0.7289 | 0.8579 | 0.8973 | 0.8772 |               |           |     |

FIGURE 5  Visual comparisons of extracted road region with different comparison algorithms on Cheng-Roads. White color represents true positive (TP), yellow color represents false positive (FP), and blue color represents false negative (FN).

the IoU obtained by DRPPNet is 5.1% higher than SegNet. Besides SegNet, the average F1 of DRPPNet is nearly 8% higher than Vgg19. These experimental results demonstrate validity of DRPPNet on road extraction task.

Figure 5 shows qualitative and quantitative results with state-of-the-art methods. The first column shows original images. The remaining column shows road extraction results of different methods. The white color denotes region that is true road region in label map. The yellow color represents false positive and the blue color represents false negative. It should be noted that all compared methods are DCNN-based methods, which has great performance in road extraction. From Figure 5, it can be seen that Unet is sensitive to occlusions of trees since its predicted maps have too much region with blue color. The maps of Vgg19 and SegNet show more coherent results than Unet, which effectively alleviates effectiveness of occlusions to some extent. However, the Vgg19 and SegNet tend to bring in too much FPs. To alleviate this shortcoming, the authors propose DRPPNet and introduce pyramid pooling module to extract road feature from multiple views. Thus, the DRPPNet


shows more satisfactory and coherent road detection results and brings in less FPs than other compared methods. In addition, our method is more robust against occlusions.

Finally, the authors list training and inference time for SegNet, Vgg19, Unet and DRPPNet on Cheng-Roads in Table 5. The inference time is time that network extract all road region in testing dataset. And the training time is time that network needs to cost for achieving convergence. From Table 5, it can be seen that DRPPNet only cost 0.8 h before convergence, which reduces 0.2 h at least than other road extraction networks. Although the PPM is added in DRPPNet, which may bring a little complexity into network, the authors find DRPPNet can have an appropriate inference time on testing dataset, about 0.9 min.

4.5 Experimental results on Mnih-Roads

In this section, the authors conduct experiments on Mnih-Roads to further verify effectiveness of DRPPNet. The Mnih-Roads is a more complex road dataset than Cheng-Roads since this dataset covers a wide variety regions, such as urban, suburban, and rural regions. The compared methods are FCN[47], CIS-CNN[21] and Vgg19. The road extraction results are presented in Table 6 and Figure 6.

The authors list quantitative results of all compared methods in Table 6. The best results are marked in bold. Although metric value on Mnih-Roads is not comparable with Cheng-Roads, DRPPNet achieves all of the best performances on Mnih-Roads. Specially, the average score of DRPPNet achieves 0.7446, 0.7419, 0.7432 and 0.5852 in terms of Pre, Rec, F1 and IoU, respectively, which is 0.9%, 1.2%, 0.6% and 1.5% higher than Vgg19. Compared with CIS-CNN that the performance only achieves 0.7109, 0.5157, 0.5978 and 0.405 in Pre, Rec, F1 and IoU; the performance of DRPPNet is improved to 3.4%, 22.6%, 14.5% and 15.5%, respectively. Moreover, the road extraction results of FCN on Mnih-Roads are not compared with DRPPNet and the average F1 score of FCN only is 0.5215. These experimental results confirm that DRPPNet indeed can improve road extraction from high resolution remote sensing image. The reason is that DRPPNet has the ability to integrate multiple views road features through PPM.

Furthermore, in order to directly compare road extraction performance of different methods, Figure 6 present a visual comparison of different methods about road extraction results on Mnih-Roads dataset. The method with visual results from first column to fourth column are FCN, CIS-CNN, Vgg19, respectively and the last column is proposed DRPPNet. It is clear that proposed DRPPNet retains most details and achieves the best road extraction result. The road extraction results of FCN exhibit most disconnection, compared to other methods.

4.6 Effectiveness of PPM

To better explore effectiveness of PPM among different road extraction methods based on same feature extraction network,
the authors provide visual and quantitative performance of different road extraction methods with PPM in Tables 8 and 9 and Figures 7 and 8.

It is noted that SegNet and Unet are encoder-decoder network. Therefore, the authors directly put PPM on last of encoder but before of decoder, which aims to capture multiple views road features. For FCN and Vgg19 that are not encoder-decoder structure network, the authors view them as feature extractor and put PPM and DD on the last. By comparing average performances in Tables 8 and 9, it can be seen that all modified methods (i.e. SegNet+PPM, Unet+PPM, FCN+PPM and Vgg19+PPM) achieve better quantitative performance than their original methods (i.e. SegNet, Unet, FCN and Vgg19) to some degree. Specially, the SegNe+PPM achieves about 1.95%
TABLE 9  Performance Evaluation of PPM on Mnih-Roads dataset. The left part shows road extraction accuracies of original networks and the right part shows accuracies with PPM.

|       | FCN   | FCN+PPM | Vgg19 | Vgg19+PPM |
|-------|-------|---------|-------|-----------|
|       | Pre   | Rec     | F1    | Pre       | Rec     | F1    | Pre   | Rec     | F1    |
| Image 4 | 0.5971 | 0.7411  | 0.6614 | 0.7699    | 0.5762  | 0.6591 | 0.8192 | 0.7171  | 0.7648 |
| Image 5 | 0.4379 | 0.7048  | 0.5402 | 0.6292    | 0.5632  | 0.5944 | 0.6781 | 0.7900  | 0.7298 |
| Image 6 | 0.5815 | 0.7435  | 0.6526 | 0.7504    | 0.6061  | 0.6706 | 0.7849 | 0.7794  | 0.7821 |
| Average | 0.4343 | 0.6526  | 0.5215 | 0.5051    | 0.6638  | 0.5737 | 0.7455 | 0.7298  | 0.7376 |
| mean-IoU | 0.3706 |         |       | 0.4129    |         |       | 0.5700 |         |       |

**FIGURE 7**  Visual comparisons of extracted road region with PPM on Cheng-Roads. The white color represents true positive (TP), yellow color represents false positive (FP), and blue color represents false negative (FN).

and 1.94% improvement in F1 and IoU than SegNet, and other three gain about 1.2% improvement in F1 than original methods. These experimental results demonstrate that road features in different views obtained by PPM can help network achieve great performance.

To better illustrate this improvement in visual effects, Figures 7 and 8 show visual road extraction results. It is obvious that the maps obtained by modified methods bring in less FPs and FNs. These visual results show that the modified methods can achieve more satisfactory and coherent road extraction results than original network. Besides, the modified methods are more robust against occlusions of trees.

Finally, the authors make an experiment to compare DRPPNet with a state-of-the-art road extraction method, ResUnet[32]. The ResUnet is built with residual units and has similar architecture to Unet, which achieves great performance in road extraction task. From Table 10, it can be seen that the average F1 of ResUnet achieves 0.8592 in F1 on Cheng-Roads. This demonstrates that ResUnet can obtain relatively coherent road areas. This is because residual unit of ResUnet utilizes...
identity add to fuse deep features and shallow features. Thus, the ResUnet has ability to extract road network more effectively. Our proposed DRPPNet is also based on residual units, but we add PPM to further obtain road features from multiple views. Therefore, the DRPPNet can obtain higher road extraction performance than ResUnet. For example, the average F1 and IoU of DRPPNet are about 4.6% and 4.2% higher than ResUnet on Minh-Roads.

5 | CONCLUSION

In this paper, a novel end-to-end deep residual and pyramid network is proposed to improve performance of road extraction from high resolution remote sensing image. The authors first utilize deep residual network (DResNet) to extract deep road feature from input image, which can effectively avoid gradient vanish and enhance learning ability of DRPPNet. Furthermore, a pyramid pooling module (PPM) is proposed to fuse road features from multiple views using four maxpooling operations and it can help DRPPNet obtain smooth and coherent road extraction results. Finally, the deep decoder (DD) is used to recover feature map to same size as input image. The experimental results on two challenging road dataset, Cheng-Roads and Minh-Roads, verify the superiority of proposed method. In addition, the proposed method can be widely applied in server computing and edge computing because of its generalization ability and efficiency.

ACKNOWLEDGEMENTS

The authors would like to thank all colleagues in the lab, who generously provided their image dataset with the ground truth. The authors would also like to thank anonymous reviewers for their very competent comments and helpful suggestions.

ORCID

Pu Han https://orcid.org/0000-0001-9739-1084
REFERENCES
1. Sun, W., et al.: Pure endmember extraction using robust kernel archetypoid analysis for hyperspectral imagery. ISPRS J. Photogramm. Remote Sens. 131, 147–159 (2017)
2. Du, B., et al.: PLTD: Patch-based low-rank tensor decomposition for hyperspectral images. IEEE Trans. Multimedia 19, 67–79 (2016)
3. Li, X., et al.: A new global land-use and land-cover change product at a 1-km resolution for 2010 to 2100 based on human–environment interactions. Ann. Assoc. Geogr. 107, 1040–1059 (2017)
4. Li, D., et al.: Intelligent transportation system in Macao based on deep self-coding learning. IEEE Trans. Ind. Inf. 14, 3253–3260 (2018)
5. Abdollahi, A., Pradhan, B.: Integrated technique of segmentation and classification methods with connected components analysis for road extraction from orthophoto images. Expert Syst. Appl. 176, 114908 (2021)
6. Xu, Z., et al.: Road extraction in mountainous regions from high-resolution images based on DSDNet and terrain optimization. Remote Sens. 13, 90 (2021)
7. Shao, Z., et al.: MRENNet: Simultaneous extraction of road surface and road centerline in complex urban scenes from very-high-resolution images. Remote Sens. 13, 239 (2021)
8. Li, Y., et al.: Fusing taxi trajectories and RS images to build road map via DCNN. IEEE Access 7, 161487–161498 (2019)
9. Unsalan, C., Sirmacek, B.: Road network detection using probabilistic and graph theoretical methods. IEEE Trans. Geosci. Remote Sens. 50, 4441–4453 (2012)
10. Soni, P.K., et al.: Semiautomated Road Extraction Framework Based on Shape Features and LS-SVM from High-Resolution Images. J. Indian Soc. Remote Sens. 1–12 (2020)
11. Cheng, J., et al.: Road junction identification in high resolution Urban SAR images based on SVM. In: International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, pp. 597–606. Springer, (2019)
12. Wegner, J.D., et al.: A Higher-order CRF model for road network extraction. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1696–1705. IEEE, Piscataway (2013)
13. Krizhevsky, A., et al.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105. MIT Press, Cambridge (2012)
14. Girshick, R., et al.: Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 38, 142–158 (2015)
15. Redmon, J., et al.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788. IEEE, Piscataway (2016)
16. Shi, Q., et al.: Road detection from remote sensing images by generative adversarial networks. IEEE Access 6, 25486–25494 (2017)
17. Gao, X., et al.: An end-to-end neural network for road extraction from remote sensing images by multiple feature pyramid network. IEEE Access 6, 39401–39414 (2018)
18. Mnih, V., Hinton, G.E.: Learning to detect roads in high-resolution aerial images. In: Proceedings, Part VI, of the 11th European Conference on Computer Vision. Springer, Berlin (2010)
19. Volodymyr Mnih, G.H.: Learning to Label Aerial Images from Noisy Data. In: International Conference on Machine Learning. IEEE, Piscataway (2013)
20. Mnih, V.: Machine learning for aerial image labeling. University of Toronto (2013)
21. Saito, S., et al.: Multiple object extraction from aerial imagery with convolutional neural networks. J. Imaging Sci. Technol. 60, 010402 (2016)
22. Saito, S., Aoki, Y.: Building and road detection from large aerial imagery. Image Processing: Machine Vision Applications VIII. Int. Soc. Opt. Photonics 9405, 94050K (2015)
23. Shannahgam, L., Kaliperumal, V.: Junction-aware water flow approach for urban road network extraction. IET Image Proc. 10, 227–234 (2016)
24. Gaetano, R., et al.: Morphological road segmentation in urban areas from high-resolution satellite images. In: 2011 17th International Conference on Digital Signal Processing (DSP), pp. 1–8. IEEE, Piscataway (2011)
25. Simonovsky, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Computer Science (2014)
26. Minae, S., et al.: Image segmentation using deep learning: A survey. IEEE Trans. Pattern Anal. Mach. Intell. (2021)
27. Miao, Z., et al.: An object-based method for road network extraction in VHR satellite images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8, 4853–4862 (2015)
28. Huang, Z., et al.: Object-based conditional random fields for road extraction from remote sensing image. IOP Conf. Ser.: Earth Environ. Sci 17, 012276 (2014)
29. Cheng, G., et al.: Road extraction via adaptive graph cuts with multiple features. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 3962–3966. IEEE, Piscataway (2015)
30. He, K., et al.: Deep residual learning for image recognition [C]. Proc. IEEE Conf. Comput. Vision Pattern Recognit. 2016, 770–778 (2016)
31. Ronneberger, O., et al.: U-Net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer, Berlin (2015)
32. Zhang, Z., et al.: Road Extraction by Deep Residual U-Net. IEEE Geosci. Remote Sens. Lett. 15, 749–753 (2018)
33. Song, M., Greco, D.: Road extraction using SVM and image segmentation. Photogramm. Eng. Remote Sens. 70, 1365–1371 (2004)
34. Zhang, Q., Coudnier, I.: Benefit of the angular texture signature for the separation of parking lots and roads on high resolution multi-spectral imagery. Pattern Recognit. Lett. 27, 937–946 (2006)
35. Yuan, J., et al.: LEGON-based automatic road extraction from satellite imagery. IEEE Trans. Geosci. Remote Sens. 49, 4528–4538 (2011)
36. Du, S., et al.: 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 (2011)
37. Alvarez, J.M., et al.: Road scene segmentation from a single image. In: European Conference on Computer Vision, pp. 376–389. Springer, Berlin (2012)
38. Zhao, J., et al.: High-resolution image classification integrating spectral-spatial-location cues by conditional random fields. IEEE Trans. Image Process. 25, 4033–4045 (2016)
39. Liao, J., Wang, L.: Multiple Spatial Features Extraction and Fusion for Hyperspectral Images Classification. Can. J. Remote Sens. 46, 193–213 (2020)
40. Lu, Z., et al.: 3-D Channel and spatial attention based multiscale spatial–spectral residual network for hyperspectral image classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 4311–4324 (2020)
41. Pu, C., et al.: Spatial-Spectral Combination Convolutional Neural Network for Hyperspectral Image Classification. In: IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium, pp. 2037–2040. IEEE, Piscataway (2020)
42. Chen, L.C., et al.: DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40, 834–848 (2018)
43. Noh, H., et al.: Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision, pp. 1520–1528. IEEE, Piscataway (2015)
44. Badrinarayanan, V., et al.: SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39, 2481–2495 (2017)
45. Wegner, J.D., et al.: A higher-order CRF model for road network extraction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1698–1705. IEEE, Piscataway (2013)
46. Cheng, G., et al.: Automatic Road Detection and Centerline Extraction via Cascaded End-to-End Convolutional Neural Network. IEEE Trans. Geosci. Remote Sens. 55, 3322–3337 (2017) https://doi.org/10.1109/TGRS.2017.2669341.

47. Maggiori, E., et al.: Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification. IEEE Trans. Geosci. Remote Sens. 55, 645–657 (2017)

How to cite this article: Han, Y., Han, P., Jia, M. Road extraction from high resolution remote sensing image via a deep residual and pyramid pooling network. IET Image Process. 1–14 (2021).
https://doi.org/10.1049/ipr2.12296