Pattern Recognition 126 (2022) 108549

Contents lists available at ScienceDirect

Pattern Recognition

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

SDUNet: Road extraction via spatial enhanced and densely connected UNet

Mengxing Yang, Yuan Yuan*, Ganchao Liu

School of Artificial Intelligence, OPtics and ElectroNics(iOPEN), Northwestern Polytechnical University, Xi'an, Shaanxi 710072, PR China

ARTICLE INFO

Article history:
Received 30 October 2020
Revised 8 November 2021
Accepted 22 January 2022
Available online 12 February 2022

Keywords:
Road extraction
Image segmentation
Remote sensing imagery
Spatial topology

ABSTRACT

Extracting road maps from high-resolution optical remote sensing images has received much attention recently, especially with the rapid development of deep learning methods. However, most of these CNN based approaches simply focused on multi-scale encoder architectures or multiple branches in neural networks, and ignored some inherent characteristics of the road surface. In this paper, we design a novel network for road extraction based on spatial enhanced and densely connected UNet, called SDUNet. SDUNet aggregates both the multi-level features and global prior information of road networks by combining the strengths of spatial CNN-based segmentation and densely connected blocks. To enhance the feature learning about prior information of road surface, a structure preserving model is designed to explore the continuous clues in the spatial level. Experimental results on two benchmark datasets show that the proposed method achieves the state-of-the-art performance, compared with previous approaches for road extraction. Code will be made available on https://github.com/MrStrangerYang/SDUNet.

1. Introduction

Automatically generating road maps from images is beneficial to a wide range of application domains [1,2], such as autonomous driving [3], land investigation [4], and urban planning [5]. One of the most appealing approaches for road map generation is by means of very high-resolution(VHR) satellite imaging technology, owing to its capability of large-area coverage. However, many factors, such as shadows induced by trees and complex road network topology, mitigate the accurate extraction of road maps, which makes this task challenging.

In early studies, researchers have designed various kinds of hand-crafted features for road extraction [6–8]. These conventional expert-knowledge based manners usually utilized texture structure [7,9], super-pixel partitioning [10,11], curve evolution techniques and region growing algorithms [12,13] to obtain integrated road areas. With local context information and prior neighborhood structures, these well-designed methods have achieved good segmentation results under certain conditions. In the past few years, rapid development in satellite imaging technology and the explosive growth of VHR image data bring about new challenges for road extraction due to complex structure of roads and diverse background distribution. In view of the above factors, conventional methods start to seriously underperform, since the models become poor generalization performance and feature expression limited.

Recently, convolutional neural networks (CNN) have made great success in tackling the semantic segmentation problem [14,15]. One of the most popular CNN-based methods is the fully convolutional network (FCN) [16], where an end-to-end semantic segmentation method was proposed based on the design of the fully convolutional layer. By combining low-level and high-level feature maps, Ronneberger et al. [17] developed the U-Net architecture, which achieved promising performance on semantic segmentation for biomedical images. Compared to hand-crafted features, properly trained CNNs is proved more effective in multi-level feature learning and generalization performance [18,19].

Although CNN based methods have significantly promoted research in the field of remote sensing semantic understanding, it is still not performing well in special features of remote sensing image analysis [2,20,21]. Unlike traditional visible images, road surface in remote sensing images usually have strong structure prior but less appearance clues. For instance, in Fig. 1, the roads and crossings have long continuous shape and may be occluded by street trees or cars. To address this issue, we design a framework based on spatial enhanced and densely connected UNet, named SDUNet, for multi-level context information learning in road map extraction task. The main contributions of our work are as follows:

1. A novel network called SDUNet is proposed for remote sensing road extraction. Considering the geometric topology of road

* Corresponding author.
E-mail address: y.yuan1.ieee@qq.com (Y. Yuan).

https://doi.org/10.1016/j.patcog.2022.108549
0031-3203/© 2022 Elsevier Ltd. All rights reserved.

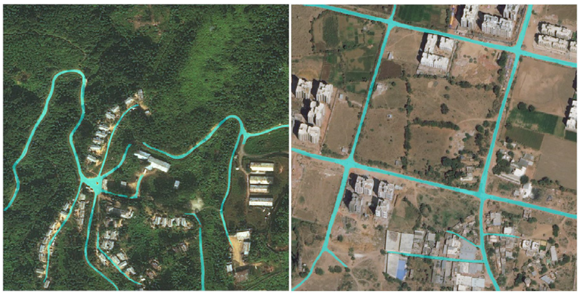


Fig. 1. Examples of road extraction of the DeepGlobe Road Extraction dataset. It can be seen that roads and crossings have long continuous shape and may be occluded by street trees or cars.

networks, densely connected encoder block and spatial intensifier (DULR Module) are designed to learn multi-level features and global prior information of road networks.

2. We introduced a structure preserving model, called DULR, to enhance the feature learning about the structure prior of road surface. Via slice-by-slice convolutions on encoding feature maps, it can guide the network to learn the continuous clues in four directions and mitigate the loss of spatial features during the encoding process.
3. To validate our approach, comprehensive experiments are carried out on Massachusetts dataset [18] and DeepGlobe dataset [22]. Experimental results show that our proposed approach achieves state-of-the-art performance in F1-score and IoU with a slight increase in the number of parameters, compared with previous approaches for road extraction.

This paper is organized as follows: Section 2 introduced previous representative works related to image segmentation and road extraction. In Section 3, the details of SDUNet are explained. Section 4 presents the experimental results and analyze the advance of SDUNet numerically and visually. Finally, our conclusions is given in Section 5.

2. Related work

Researches on remote sensing road extraction can be classified into two categories: conventional expert-knowledge based manners and CNN-based frameworks. In this section, we briefly review some representative approaches in these two categories.

With respect to conventional methods, researchers mainly model the dependencies of super-pixels around road areas by combining different types of prior texture information [7,10,12,23]. Then, discriminative or clustering methods are proposed to classify adjacent pixel blocks based on hand-crafted features. For instance, Miao et al. [24] absorbed the strengths of the geodesic method, KDE, and mean shift to trace the unbiased road centerlines between points of interest. In view of geometric and radiometric variability, SFS-SD texture analysis and beamlet transform based multi-scale reasoning is used in Sghaier and Lepage [25] to efficiently distinguish rectilinear structures of road surface. In [26], Shi et al. developed Zernike moments as discriminative characteristics to distinguish different objects in SAR images. More recently, Wegner et al. [27] modeled road networks via constructing higher-order conditional random fields by using super-pixel segments and the connected paths among them. Benediktsson et al. [28] formulated road extraction as a feature extraction and classification problem, then integrate HGAPSO and SVM to discriminate road and background pixels. These mentioned classical ML approaches take into account the context information and texture information of the image, and achieve better segmentation accuracy under certain conditions. However, in practical application, these algorithms rely on artificial feature construction, which not only takes time and energy, but are also difficult to adapt to the complex remote sensing environments.

Deep learning has attached great attention in a series of semantic understanding tasks [17,29,30], as its ability to cover multiple level features. LinkNet architecture [31] was proposed to reduce the number of network parameters to learn and improve the real-time processing ability of networks. Taking advantage of the powerful pretrained VGG encoders [32] and the U-Net architecture [17], Iglovikov et al. proposed TernausNet [29], which was part of the winning solution (first out of 735) in the Carvana Image Masking Challenge. In the field of road extraction, Zhou et al. [33] proposed D-LinkNet for VHR remote sensing road extraction. The D-linknet, which was delivered from Linknet [31], embed dilated convolution layers into the middle part to ensemble multi-scale features and reserve the detailed information simultaneously. Zhang et al. [34] proposed a road extraction framework based on FCN with an ensemble strategy, called spatial consistency (SC), to solve the imbalance of road and background areas in aerial images.

Compared with natural images, targets in remote sensing images usually have the characteristics of lack of appearance clues, variable scales, and complex backgrounds, which makes it difficult to obtain ideals by directly applying deep convolutional neural networks. Considering the above factors, other attempts were carried out to exploit spatial information via neural networks. Recurrent neural networks were used in Visin et al. [35], Bell et al. [36] to exploit information transfer mechanisms in spatial level. Specially, each pixel position in every RNN layer could only exchange information from the corresponding row or column. Wang et al. [37] introduced global attention modules and improved category classification by using pooling operation to maintain global context information and enhance high-level features to improve segmentation performance. Xie et al. proposed an improved Linknet model in Xie et al. [38], called HsgNet. Taking into account the long span, connectivity, and complexity of roads, HsgNet utilized bilinear pooling based attention modules to preserve global, second-order context cues. In [20], a two-arm convolutional network, named as Y-Net, was proposed to obtain rich road detailed features and eliminate complex background interference simultaneously for road extraction.

In summary, despite numerous attempts on CNN-based road extraction, most of them simply focused on multi-scale encoder architectures or multiple branches in neural networks, but ignored some inherent characteristics of road surface. Meanwhile, conventional expert-knowledge based manners and RCNN based lane detection in Visin et al. [35], Bell et al. [36] inspires us that specific contextual priors of roads and structural association in spatial level are of great significance during feature learning. Thus, we propose SDUNet, based on spatial CNN and densely connected blocks, to further improve completeness and smoothness of road extraction in remote sensing imagery.

3. Our method

3.1. Dense-UNet

One of the most prevalent architectures for binary semantic segmentation is UNet [17], which fuses multi-level feature maps to hierarchically increase the spatial resolution of the output probability map. The encoder exploited in the original UNet is built by plain neural units, while the feature representation capability of plain neural unit is overtaken by residual neural unit proposed in He et al. [39]. More recently, a multi-layer dense block was proposed in DenseNet architecture [40], where each layer is connected to every other layer. Taking advantage of this type of connection, the forward feature propagation ability is strengthened, and the number of parameters can be reduced as the network going deeper. Inspired by this, we seek to investigate whether the UNet encoders can be made of dense blocks and if an improvement of semantic segmentation can be achieved. Its encoder part is mainly composed of four dense blocks and three transition layers, and the decoder part is constructed based on five plain neural units. To be specific, the dense block and transition layer are explained as follows.

(1) Dense block. A dense block is mainly consisted of multiple layers, which are built through their connections between each layer and subsequent layers, as presented in Fig. 4(a). Specifically, the output of lth layer can be represented by:

$$X_l = H_l([x_0, x_1, \ldots, x_{l-1}]), \tag{1}$$

where $[x_0, x_1, \ldots, x_{l-1}]$ represents concatenation of the feature maps generated from all the preceding layers, and $H_l(\cdot)$ refers to the nonlinear mapping function of the layers. Different from the residual block utilized in Huang et al. [40], the dense block exploits the concatenation operator to combine the learned feature maps, which can increase the variation in the input of the following layers and improve efficiency.

(2) Transition layer. Due to the concatenation operator utilized inside dense blocks, spatial resolution of feature maps cannot be down-sampled through dense blocks. Therefore, as illustrated in Fig. 4(b), transition layers consisting of batch normalization, convolutional and pooling layers are introduced between separate dense blocks for the spatial down-sampling.

3.2. Spatial intensifier: DULR module

The above densely connected encoder-decoder framework (Dense-UNet) is able to extract more expressive local features than the original UNet. However, It is not effective enough to handle the strong structure priors, like the spatial relationship and continuous surface of roads. To address these issues, Spatial CNN [41] is introduced to enhance the feature prior of roads on four directions, which are downward, upward, leftward and rightward respectively.

The motivation of Spatial CNN, also called DULR module, is that the network can explicitly construct the spatial relationship between different location features. Unlike universal spatial context algorithms, such as SE attention, which establish the spatial relationship between all pixels and the current pixel, the DULR module takes rows and columns in the feature map as the layer of the network and performs slice-by-slice convolution.

As shown in Fig. 3, a feature map X with size of $C \times H \times W$ is generated via encoder distribution, where C, H and W denote the size of channel, height and width respectively. The spatial CNN module performs four similar and cascading stages, and individually enhances features in four directions(upward, downward, leftward, and rightward). Taking CNN_D as an example, the feature map X firstly is separated into H slices to further exploit local connection relationship on the spatial level. Then, define the element of these slices as $X\_D_H^i (i \in [1, H])$, and the downward iterative computation of CNN_ D is defined as follows:

$$(X\_D)_H^i = \begin{cases} X_H^i, & i = 1 \\ X_H^i + f\left(X_H^{i-1} \oplus k_{C\times\omega}\right), & i = 2, 3, \ldots, H \end{cases} \tag{2}$$

where X_H^i and $X\_D_H^i$ are respectively the feather slices before and after computation. Operator $\oplus$ represents the convolution with C kernels and f is a ReLU activation function. Note that each convolution kernel is of size $C \times \omega$, and parameters shared throughout the whole process.

As roads in VHR remote sensing images are narrow and continuous through large areas, we used the DULR Module to maintain topological structure of roads and ease loss of spatial features. Fig. 3 detailed our implementation of DULR Module. Usually, topper hidden layers contain richer and more representative semantics. Considering for the trade-off between effect and complexity, two DULR blocks are deployed to replace the directly concatenated operation, as shown in Fig. 2. During the encoder process, four dense blocks are performed to extract multiple features, with DULR blocks followed the second one and the third one. Then, the low-level features after corresponding transition layers are concatenated with features enhanced through the DULR blocks to form a new input during the decoder process.

3.3. Joint loss function

The binary cross-entropy loss function is applied in most pixel-level segmentation task, such as medical image segmentation based on UNet. However, it has the drawback of misleading the model seriously biased to the background, when the number of pixels on target is much smaller than the number of pixels in background.

To address this problem, joint Loss Function, which blends the Dice coefficient and the pixel-wise binary cross entropy is employed to guide the network for which class to focus on. Given the input images Y_i and the associated ground truth maps G_i, the joint loss function is exploited for learning networks, which combines pixel-wise binary cross entropy (BCE) loss and the Dice coefficient. In particular, the Dice coefficient is determined by the true positive (TP), false positive (FP), and false negative (FN) based on the prediction and ground truth, which can be written as:

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{3}$$

Correspondingly, the joint loss function is formulated as:

$$L = \frac{1}{N} \sum_{i=1}^{N} (BCE(F(Y_i), G_i) + 1 - Dice(F(Y_i), G_i)) \tag{4}$$

where N is the number of images in a batch, and $F(Y_i)$ represents the output probability map of the trained network, given the input image Y_i.

4. Experiments and results

To validate the effectiveness of our approach, comprehensive experiments are carried out on two benchmark datasets, which are Massachusetts dataset [18] and DeepGlobe dataset [22]. We perform Dense-UNet, SUNet and SDUNet for ablation study, and compare the proposed method with other road extraction techniques, including classical ML based model HGAPSO + SVM [28], the baseline model UNet[17], embed dilated convolution based model D-LinkNet [33]and attention based model HsgNet [38]. Hyper-parameters and protocols are exactly the same as the original article and available source code. In this section, the experimental setup and results are illustrated.

4.1. Dataset descriptions

In this experiment, we have used two datasets, namely, (i) Massachusetts road dataset [18] and (ii) DeepGlobe Road Extraction dataset [22].

(1) **Massachusetts road dataset**: The Massachusetts roads dataset was built by Mihn and Hinton [18]. The dataset provides 1171 images with size of 1500 × 1500, including 1108 for training, 14 for validation, and 49 for testing. The resolution of these images is 1m per pixel. Images in Massachusetts roads dataset totally cover 500 km^2 space crossing from cities, suburban to countryside, and diverse ground objects including multiple roads,

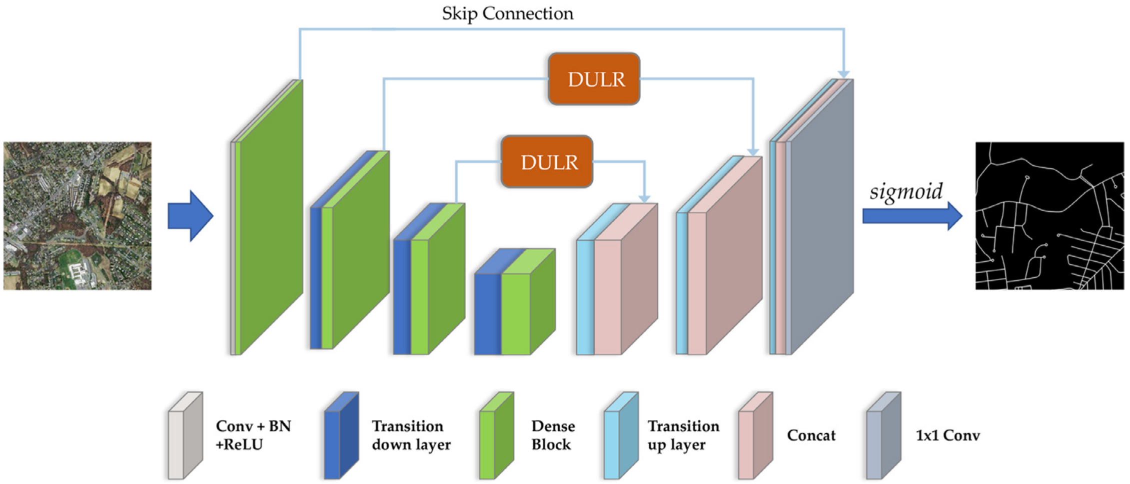


Fig. 2. The proposed architecture of SDUNet.

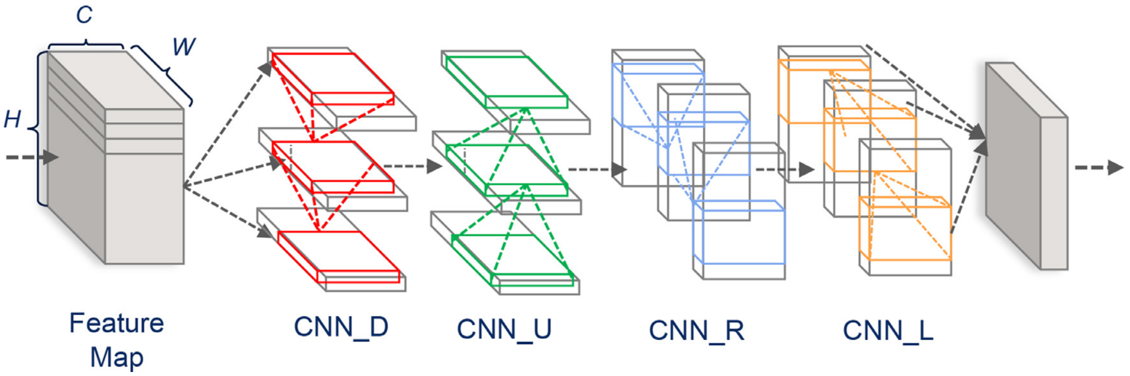
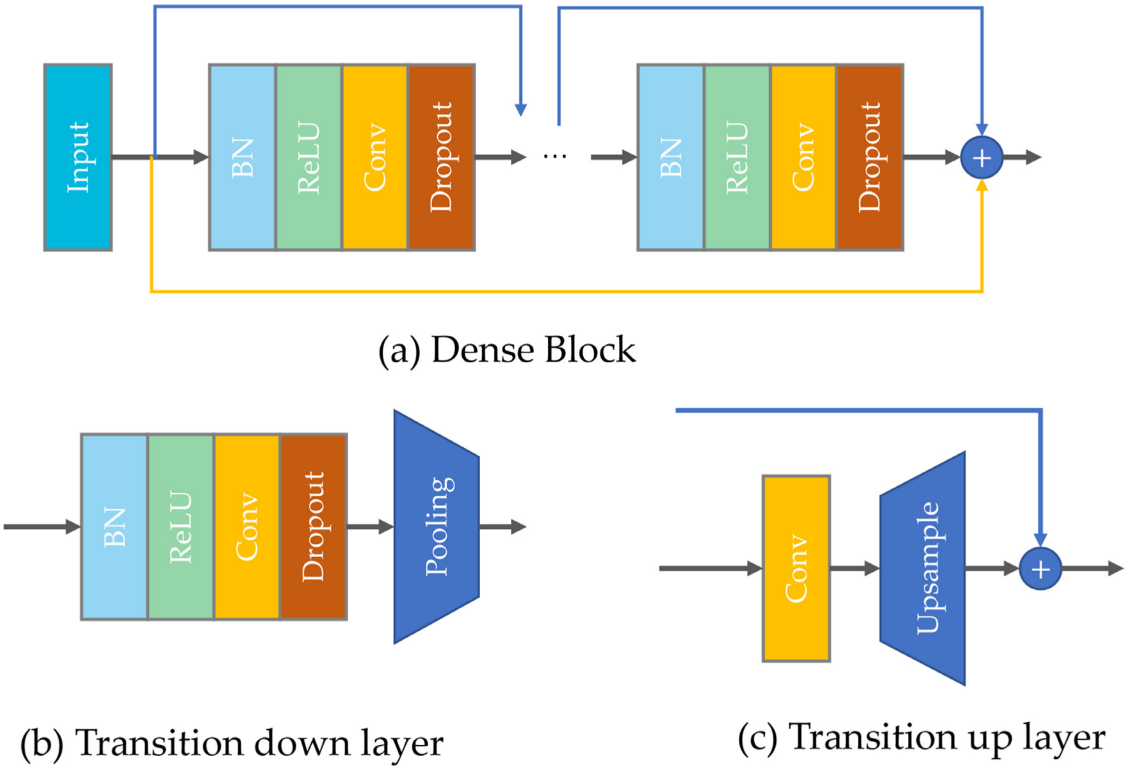


Fig. 3. Spatial Intensifier: DULR Module.



(a) Dense Block

(b) Transition down layer

(c) Transition up layer

Fig. 4. A dense block is mainly composed of multiple layers that are built through their connections between each layer and subsequent layers, and the learned feature maps inside the block are combined by the concat operator. The transition layer is exploited for down-sampling the spatial size of feature maps.

Table 1
Quantitative evaluation of seven methods conducted on the Massachusetts Dataset.

Network	Precision	Recall	F1	IoU
HGAPSO + SVM [28]	0.560	0.679	0.629	0.532
UNet [17]	0.747	0.721	0.722	0.682
D-LinkNet [33]	0.767	0.741	0.737	0.717
HsgNet [38]	0.769	0.752	0.749	0.720
Dense-UNet	0.780	0.731	0.739	0.714
SUNet	0.798	0.736	0.753	0.721
SDUNet	**0.812**	**0.757**	**0.784**	**0.741**

Table 2
Quantitative evaluation of seven methods conducted on the Deepglobe road extraction dataset.

Network	Precision	Recall	F1	IoU
HGAPSO + SVM [28]	0.531	0.653	0.622	0.493
UNet [17]	0.732	0.701	0.769	0.627
D-LinkNet [33]	0.741	0.719	0.780	0.643
HsgNet [38]	0.735	0.732	0.790	0.660
Dense-UNet	0.761	0.731	0.786	0.639
SUNet	0.772	0.727	0.788	0.657
SDUNet	**0.784**	**0.742**	**0.794**	**0.668**

buildings, trees, and crops, etc. The ground truth of the images are binary images containing two classes: roads and non-roads. In this paper, we train our network on training set of this data set and evaluate the models on its test set.

(2) **DeepGlobe Road Extraction dataset**: The DeepGlobe Road Extraction dataset [22] is also adopted. It covers images captured over Thailand, Indonesia, and India. The ground resolution of the image pixels is 50 cm/pixel. During the development phase, the dataset totally contains 1243 high-resolution remote sensing images with a spatial size of 1024 × 1024 pixels. For the following analysis, we randomly split the whole data into training, validation, and test samples with the percentages of 80%, 10%, and 10%, respectively.

4.2. Implementation details

We perform the proposed model and five other architectures in the Pytorch framework. In training phase, random crops, vertical/horizontal flips, and random rotations are adopted for data augmentation. The learning rate for all the methods was set at 5×10^{-4} and decayed by a factor of 0.1 for every 20 epochs. The total number of epochs was 200. The Adam optimizer [20] was exploited for minimizing the joint loss. Online hard negative mining procedure was also adopted for training the networks. In testing phase, we evaluate the inference performances using four metrics, including precision, recall, F1-score and IoU. Specifically, the metrics are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1{-}score = \frac{2 \times precision \times recall}{precision + recall} \tag{7}$$

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

where TP, FP and FN are separately the true positive, false positive, and false negative based on the prediction and ground truth.

4.2.1. Test on massachusetts road extraction dataset

Quantitative results on Massachusetts Road Extraction dataset are shown in Table 1. It proves that our proposed approach achieves better performance both in F1-score and IoU, compared with other road extraction methods. SDUNet are respectively 6.2% and 5.9% better than baseline model(UNet) in F1-score and IoU. Meanwhile the F1-score by SDUNet is 4.7% and better than D-Linknet and 3.5% better than HsgNet, and the IoU is 2.4% higher than D-Linknet and 2.1% higher than HsgNet.

As illustrated in Fig. 5, we demonstrate the produced road maps based on the comparison of networks. Most road areas can be classified correctly by all of these seven methods. In addition, compared with the classical ML based HGAPSO + SVM, deep learning methods can improve robustness and reduce the misdetection of pixels. It can be observed from the areas indicated by red rectangles in first and second rows that traditional U-shape networks (UNet, HsgNet and Dense-UNet) tend to break roads. Approaches based on the DULR module(SUNet and SDUNet) can profit from topological structure enhanced on the spatial level, and make a remarkable improvement in connectivity and smoothness of road extraction. The third and fourth rows are examples for evaluating the effectiveness of the methods under the complex background and occlusion of buildings or trees. All these seven methods may produce a certain extent of erroneous segmentation results, when disturbed by a series of similar semantic blocks. Embed dilated convolution in D-Linknet can slightly optimize the segmentation of local patches in complex background. Meanwhile, note that SDUNet can still perform a better result, compared to the rest. For one thing, networks may get more multiple features during the encoder process by densely connected blocks, as the effectiveness is shown in second and third row. On the other side, the plausible reasons may be that contextual information can be considered in the final decision of SDUNet.

4.2.2. Test on DeepGlobe road extraction dataset

Table 2 reports the quantitative comparison of seven methods, conducted on the Deepglobe road extraction dataset. As shown in Table 2, SDUNet achieves the best performance in four metrics, and gets 4.1% better than baseline model(UNet), 2.5% and 0.8% better than D-LinkNet, HsgNet in IoU. It proves the superiority of our method on DeepGlobe Road Extraction dataset.

Some examples in testing set from DeepGlobe dataset are presented in Fig. 6, which consists of original images, ground truth, HGAPSO + SVM, UNet, HsgNet, D-Linknet, Dense-UNet, and SUNet from left to right. With the enhancement of remote sensing image resolution, the interference factors of complex background are greatly increased, and it is difficult for classical ML approaches to adapt to these challenges. The first row in Fig. 6 presents an example of blur road, traditional CNN-based methods (UNet, HsgNet and D-Linknet) can not extract the single road indicated by red rectangles, as these methods are trained mainly depending on texture features. Attention module in HsgNet can help to find more global information and slightly alleviate the discontinuity of the extraction results. Dense-UNet can identify the road completely, but brings several erroneous extractions. SUNet and SDUNet are robust in topological structure and generate better segmentation results. Examples with slender roads and similar texture interference are reported in second and third row, the result of SDUNet still performs the most accurate results segmentation results, while other methods produce lack of the necessary connectivity. A kind of complex road network is shown in the fourth row in Fig. 6, known as overpass. It is observed that the baseline UNet generates more break points in predicted road map. HsgNet, D-Linknet and Dense-UNet tent to split the dual lanes in red rectangles, while the DULR based SUNet and SDUNet aggregate these lanes under the influence of structure preserving model.

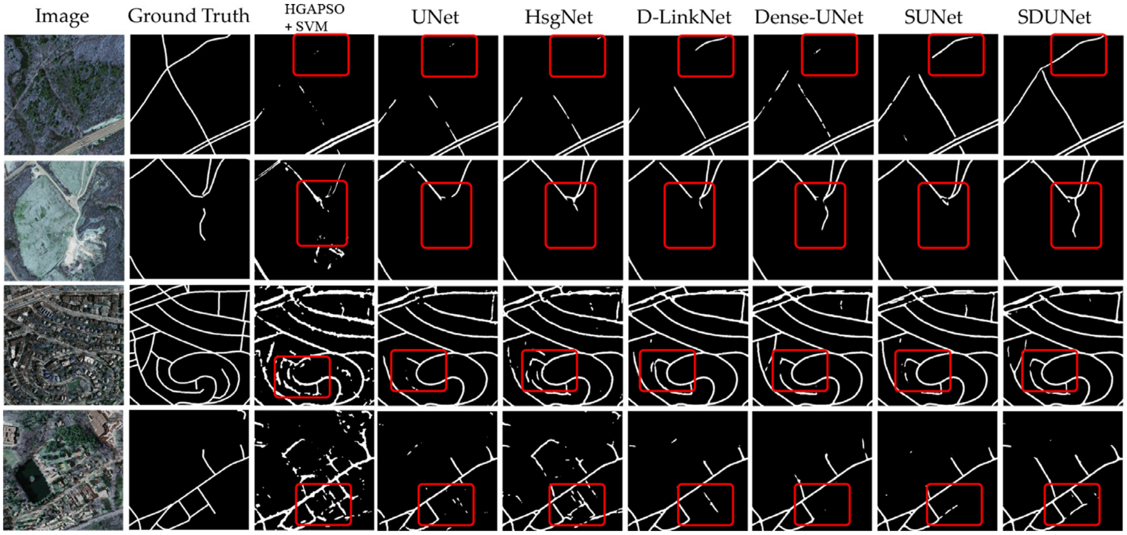


Fig. 5. Visual comparison of seven methods, conducted on the Massachusetts road dataset.

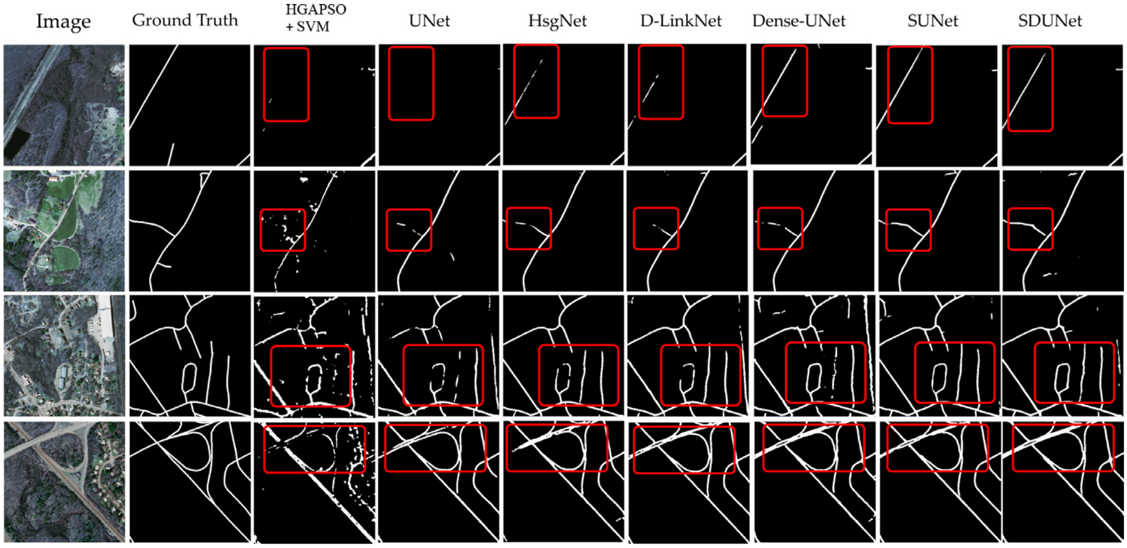


Fig. 6. Visual comparison of seven methods, conducted on the DeepGlobe road extraction dataset.

4.2.3. *Ablation study*

In this section, we evaluate the effects of different modules in the performance of the proposed SDUNet Framework. We use UNet as a backbone. Dense-UNet and SUNet are densely connected UNet and spatial enhanced UNet respectively.

As shown in Table 1, taking the results of UNet (First row) as a baseline, the IoU improved 3.2% and F1-score improved 1.7% when using our multi-scale dense block on the Massachusetts data set (Dense-UNet in fourth row). When our Spatial CNN model was used (SUNet in fifth row), IoU improved 3.9% and F1-score improved 3.1%. The IoU and F1 score of fusion method (SDUNet in sixth row) were 5.9% and 6.2% higher than the baseline, respectively.

In Table 2, on the DeepGlobe dataset, the results of Dense-UNet increased by 1.2% in IoU and 1.7% in F1-score compared to UNet. The Spatial CNN model can improve the performance over baseline method by 3.0% in IoU and 1.9% in F1-score. Overall, the fusion method (SDUNet) can improve the performance of UNet baseline significantly from 62.7% to 66.8% in IoU and 76.9% to 79.4% in F1-score.

Some examples for ablation study are presented in Fig. 7. Row(a) shows input image with narrow and long road. UNet can not predict the single road throughout the full image completely. With densely connected blocks, dense-UNet can extract the road more effectively, but there are still a few breaking points. Compared with UNet and Dense-UNet, the segmentation results of SUNet and SDUNet are complete and smooth, which proves that DULR model is effective for exploring the spatial relationship in road extraction. Row(b) in Fig. 7 presents road occluded by street trees and cars. The predicted road maps by UNet and Dense-Unet tend to be curved around these disturbances. The DULR can alleviate these issues by learning to maintain the topological structure of roads, hence the results of SUNet and SDUNet are more straight. In row(c), it can be observed that Dense-UNet and SDUNet are more discriminating for equivocal roads via densely connected blocks, correspondingly compared with UNet and SUNet. In row(d), all of these four methods are not desirable for complex transportation network with dense traffic roads. Compared to the baseline and two modified approaches, SDUNet significantly improves performance results by aggregate

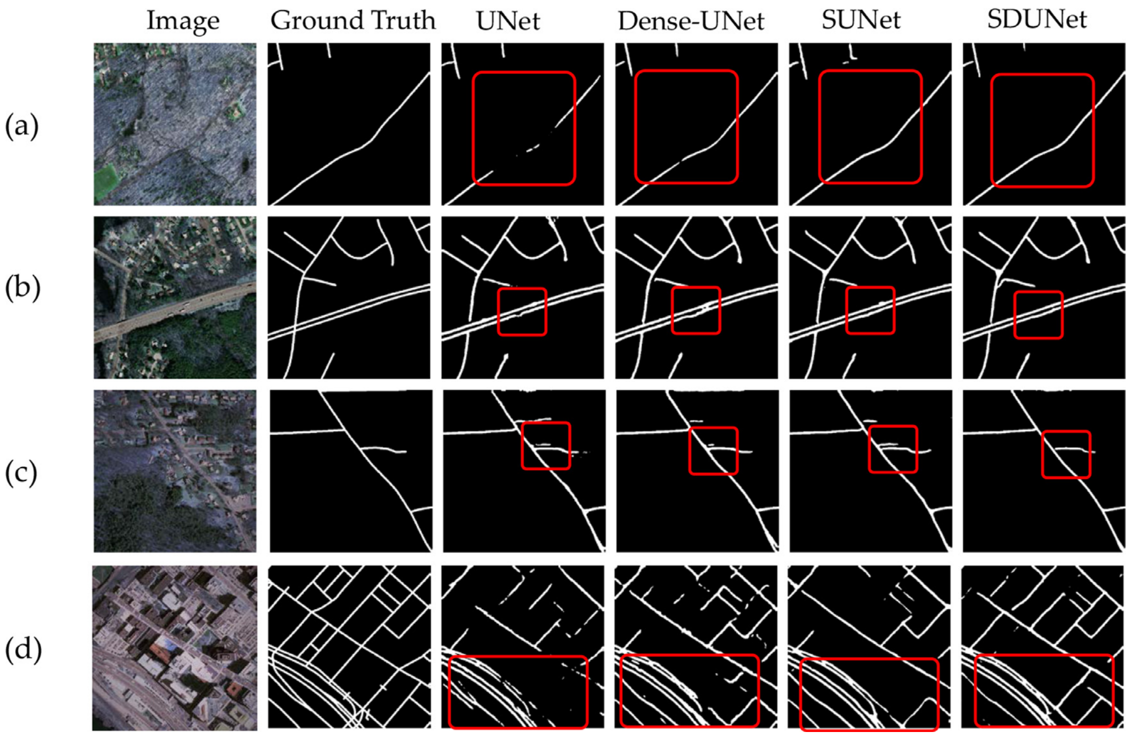


Fig. 7. Visual comparison for ablation study. Column 1: original images; Column 2: ground truth; Column 3: the backbone network (UNet); Column 4: densely connected based UNet (Dense-UNet); Column 5: spatial enhanced network (SUNet); Column 6: our proposed method (SDUNet). Row (a) and (b) are results Massachusetts Road Extraction dataset, meanwhile Row (c) and (d) are results Massachusetts Road Extraction dataset.

Table 3
Comparison of computational efficiency of different methods.

Network	Params (M)	MACs (GMac)	Time cost (s)
HGAPSO + SVM [28]	–	–	1.30
UNet [17]	28.95	193.18	1.25
D-LinkNet [33]	83.64	139.71	1.40
HsgNet [38]	66.24	290.75	2.05
Dense-UNet	55.73	255.62	1.35
SUNet	51.14	310.90	1.46
SDUNet	80.24	353.26	1.75

both the multi-level features and global prior information of road networks.

4.2.4. Computational efficiency

Table 3 shows comparison of computational efficiency of different methods. It consists of parameters of the models Multiply-accumulate operations (MACs) on the pytorch platform. Meanwhile, the time required for each sample in the test phase is also shown in the Table 3.

These experimental results demonstrate that the proposed framework achieves better performance in road extraction than other state-of-the-art methods, with an increase in a few numbers of parameters. In terms of running speed, our approach is at the same level compared with the previous algorithms.

5. Conclusions

This paper proposes a novel network based on spatial enhanced Dense-UNet, named as SDUNet, to perform the road extraction task in high-resolution remote sensing images. SDUNet takes advantage of the prominent ability of feature encoding based on dense blocks and spatial context information to predict the high-resolution road masks. In SDUNet, densely connected blocks are proposed to extract multi-level local features in the encoding stage, and a structure preserving model called DULR is constructed to explore the continuous clues in the spatial level and mitigate the loss of information during the encoding process. Based on the Massachusetts Road Extraction datas et and DeepGlobe Road Extraction dataset, SDUNet achieves better performance than other state-of-the-art networks according to the metrics of F1-score and IoU with a slight increase in the number of parameters. Especially, with a slight increase in the number of parameters, SDUNet significantly maintains the geometric structure of the road network and improves the segmentation performance for remote sensing road extraction.

For future work, we would like to investigate whether multi-task learning can be adopted for road extraction. Besides the consideration of road topology, we seek to combine more prior information into the training of networks.

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.

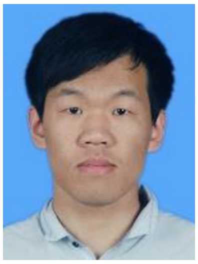
Acknowledgments

This work was supported in part by the National Key Research and Development Project under Grant 2020YFB2103902, in part by the National Science Fund for Distinguished Young Scholars under Grant 61825603, in part by the Key Program of National Natural Science Foundation of China under Grant 61632018 and in part by the National Natural Science Foundation of China under Grant 62001390.

References

[1] W. Wang, N. Yang, Y. Zhang, F. Wang, T. Cao, P. Eklund, A review of road extraction from remote sensing images, J. Traffic Transp. Eng. 3 (3) (2016) 271–282.

[2] Y. Long, G. Xia, S. Li, W. Yang, M. Yang, X. X. Zhu, L. Zhang, D. Li, DiRS: on creating cenchmark datasets for remote sensing image interpretation, arXiv: 2006.12485[cs] (2020).
[3] J. Senthilnath, N. Varia, A. Dokania, G. Anand, J. Benediktsson, Deep TEC: deep transfer learning with ensemble classifier for road extraction from UAV imagery, Remote Sens. 12 (2) (2020) 245.
[4] C. Tian, C. Li, J. Shi, Dense fusion classmate network for land cover classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 192–196.
[5] A. Van Etten, City-scale road extraction from satellite imagery v2: road speeds and travel times, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1775–1784.
[6] S. Hinz, A. Baumgartner, Automatic extraction of urban road networks from multi-view aerial imagery, ISPRS J. Photogramm. Remote Sens. 58 (1–2) (2003) 83–98.
[7] X. Huang, L. Zhang, Road centreline extraction from high-resolution imagery based on multiscale structural features and support vector machines, Int. J. Remote Sens. 30 (8) (2009) 1977–1987.
[8] J. Hu, A. Razdan, J. Femiani, M. Cui, P. Wonka, Road network extraction and intersection detection from aerial images by tracking road footprints, IEEE Trans. Geosci. Remote Sens. 45 (12) (2007) 4144–4157.
[9] B. Han, Y. Wu, A novel active contour model based on modified symmetric cross entropy for remote sensing river image segmentation, Pattern Recognit. 67 (2017) 396–409.
[10] Q. Wang, X. He, X. Li, Locality and structure regularized low rank representation for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 57 (2) (2018) 911–923.
[11] S. Choy, S. Lam, K. Yu, W. Lee, K. Leung, Fuzzy model-based clustering and its application in image segmentation, Pattern Recognit. 68 (2017) 141–157.
[12] C. Unsalan, B. Sirmacek, Road network detection using probabilistic and graph theoretical methods, IEEE Trans. Geosci. Remote Sens. 50 (11) (2012) 4441–4453.
[13] A. Saglam, A. Baykan, Sequential image segmentation based on minimum spanning tree representation, Pattern Recognit. Lett. 87 (2017) 155–162.
[14] C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.
[15] Y. Li, L. Song, Y. Chen, Z. Li, X. Zhang, X. Wang, J. Sun, Learning dynamic routing for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8553–8562.
[16] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[17] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention, 2015, pp. 234–241.
[18] V. Mnih, Machine Learning for Aerial Image Labeling, Citeseer, 2013.
[19] G. Liu, L. Li, L. Jiao, Y. Dong, X. Li, Stacked Fisher autoencoder for SAR change detection, Pattern Recognit. 96 (2019) 106971.
[20] Y. Li, L. Xu, J. Rao, L. Guo, Z. Yan, S. Jin, A Y-Net deep learning method for road segmentation using high-resolution visible remote sensing images, Remote Sens. Lett. 10 (4) (2019) 381–390.
[21] G. Liu, Y. Yuan, Y. Zhang, Y. Dong, X. Li, Style transformation-based spatial-spectral feature learning for unsupervised change detection, IEEE Trans. Geosci. Remote Sens. 60 (2022) 5401515, doi:10.1109/TGRS.2020.3026099.
[22] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, R. Raska, Deepglobe 2018: a challenge to parse the earth through satellite images, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 172–17209.
[23] X. Li, M. Chen, F. Nie, Q. Wang, A multiview-based parameter free framework for group detection, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017, pp. 4147–4153.
[24] Z. Miao, B. Wang, W. Shi, H. Zhang, A semi-automatic method for road centerline extraction from VHR images, IEEE Geosci. Remote Sens. Lett. 11 (11) (2014) 1856–1860.
[25] M. Sghaier, R. Lepage, 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) (2015) 1946–1958.
[26] W. Shi, Z. Miao, J. Debayle, An integrated method for urban main-road centerline extraction from optical remotely sensed imagery, IEEE Trans. Geosci. Remote Sens. 52 (6) (2013) 3359–3372.
[27] J. Wegner, J. Montoya, K. Schindler, Road networks as collections of minimum cost paths, ISPRS J. Photogramm. Remote Sens. 108 (2015) 128–137.
[28] P. Ghamisi, J. Benediktsson, Feature selection based on hybridization of genetic algorithm and particle swarm optimization, IEEE Geosci. Remote Sens. Lett. 12 (2) (2015) 309–313.
[29] V. Iglovikov, A. Shvets, Ternausnet: U-Net with VGG11 encoder pre-trained on imagenet for image segmentation, arXiv preprint arXiv:1801.05746(2018).
[30] Y. Yuan, J. Fang, X. Lu, Y. Feng, Spatial structure preserving feature pyramid network for semantic image segmentation, ACM Trans. Multimed. Comput., Commun., Appl. (TOMM) 15 (3) (2019) 1–19.
[31] A. Chaurasia, E. Culurciello, Linknet: exploiting encoder representations for efficient semantic segmentation, in: Proceedings of the IEEE Visual Communications and Image Processing (VCIP), 2017, pp. 1–4.
[32] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556(2014).
[33] L. Zhou, C. Zhang, M. Wu, D-LinkNet: linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 192–1924.
[34] X. Zhang, W. Ma, C. Li, J. Wu, X. Tang, L. Jiao, Fully convolutional network-based ensemble method for road extraction from aerial images, IEEE Geosci. Remote Sens. Lett. 17 (10) (2019) 1777–1781.
[35] F. Visin, K. Kastner, K. Cho, M. Matteucci, A. Courville, Y. Bengio, Renet: a recurrent neural network base alternative to convolutional networks, arXiv preprint arXiv:1505.00393(2015).
[36] S. Bell, C. Zitnick, K. Bala, R. Girshick, Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2874–2883.
[37] S. Wang, H. Yang, Q. Wu, Z. Zheng, Y. Wu, J. Li, An improved method for road extraction from high-resolution remote-sensing images that enhances boundary information, Sensors 20 (7) (2020) 2064.
[38] Y. Xie, F. Miao, K. Zhou, J. Peng, HsgNet: a road extraction network based on global perception of high-order spatial information, ISPRS Int. J. Geoinf. 8 (12) (2019) 571.
[39] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[40] G. Huang, Z. Liu, L. Maaten, K. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708.
[41] X. Pan, J. Shi, P. Luo, X. Wang, X. Tang, Spatial as deep: spatial CNN for traffic scene understanding, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI), 2018, pp. 7276–7283.



Mengxing Yang is currently working toward the Ph.D. degree with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and machine learning.

Yuan Yuan is currently a Full Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. She has authored or co-authored over 150 papers, including about 100 in reputable journals, such as the IEEE Transactions and Pattern Recognition, and the conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image video content analysis.

Ganchao Liu received the Ph.D. degree from Xidian University, Xi'an, China, in 2016. He is currently an associate professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His current interests include image processing and pattern recognition.