IDANet: Iterative D-LinkNets with Attention for Road Extraction from High-Resolution Satellite Imagery

Benzhu Xu, Shengshuai Bao, Liping Zheng, Gaofeng Zhang, and Wenming Wu(✉)

School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China
wwming@hfut.edu.cn
http://ci.hfut.edu.cn/

Abstract. Road information plays a fundamental role in many application fields, while satellite images are able to capture a large area of the ground with high resolution. Therefore, extracting roads has become a hot research topic in the field of remote sensing. In this paper, we propose a novel semantic segmentation model, named IDANet, which adopts iterative D-LinkNets with attention modules for road extraction from high-resolution satellite images. Our road extraction model is built on D-LinkNet, an effective network which adopts encoder-decoder structure, dilated convolution, and pretrained encoder for road extraction task. The attention mechanism can be used to achieve a better fusion of features from different levels. To this end, a modified D-LinkNet with attention is proposed for more effective feature extraction. With this network as the basic refinement module, we further adopt an iterative architecture to maximize the network performance, where the output of the previous network serves as the input of the next network to refine the road segmentation and obtain enhanced results. The evaluation demonstrates the superior performance of our proposed model. Specifically, the performance of our model exceeds the original D-LinkNet by 2.2% of the IoU on the testing dataset of DeepGlobe for road extraction.

Keywords: Road extraction · Semantic segmentation · Convolutional neural network · Attention mechanism · Iterative architecture

1 Introduction

Accurate and real-time road information update is of great significance in many applications, such as urban planning and construction, vehicle navigation, natural disaster analysis, etc. Benefiting from the development of remote sensing

This work was supported in part by the National Natural Science Foundation of China (No. 61972128, 61906058) and the Natural Science Foundation of Anhui Province, China (No. 1808085MF176, 1908085MF210) and the Fundamental Research Funds for the Central Universities, China (No. PA2021KCPY0050).

H. Ma et al. (Eds.): PRCV 2021, LNCS 13020, pp. 140–152, 2021.
https://doi.org/10.1007/978-3-030-88007-1_12

technology, continuous ground observations via remote sensing satellites have been achieved, and a large number of high-resolution satellite images can be easily obtained, providing a reliable and abundant data source with rich spatial structure and geometric texture for extracting various ground targets.

Traditional methods aim to detect roads by carefully designing multi-level and multi-scale features to distinguish between the road and non-road. However, it is challenging to choose effective features, especially in complex heterogeneous regions where surroundings such as buildings, vegetation are all contextual features that also affect the original features of the road. Recently, with the rise of deep learning, convolutional neural networks (CNN) have shown considerable development in feature extraction and have been applied in many computer vision tasks, including image recognition [11] and semantic segmentation [14]. The "era of deep learning" also popularizes the use of deep neural networks in the field of remote sensing, which has enabled effective automatic road extraction.

CNN-based approaches have been proposed for extracting roads, and remarkable improvements have been made. Most of them consider road extraction as semantic segmentation. However, this problem is far from solved. On the one hand, ground information provided by satellite images has increased dramatically with altitude, which reduces the differences of ground targets. On the other hand, roads often have differences in width, color, and structure. Surroundings usually cause shadow and occlusion issues. The above-mentioned has undoubtedly increased the difficulty of road extraction with inherent complexity and variability. CNN-based methods extract road parts with some important parts missed and poor connectivity. The road is sparser compared to other ground targets, and the imbalance of categories increases the difficulty of semantic segmentation. Key features can not be well extracted from various feature information.

To this end, we propose a novel semantic segmentation model named IDANet, which adopts iterative D-LinkNets [20] with attention modules [19] for road extraction with much better accuracy, connectivity, and completeness than previous methods. We equip the original D-LinkNet with more advanced dilated convolution modules, which can further enhance the ability of feature learning. To fix the issue of category imbalance, a powerful loss function with focal loss [13] is adopted. We also introduce recent popular attention modules into our model for more effective feature extraction. Furthermore, we use an iterative architecture to obtain enhanced results. The output of the previous network serves as the input of the next network to refine the results of semantic segmentation. The main contributions of our work are summarized as follows:

- A improved encoder-decoder network with attention is introduced as the basic module of our model, which can significantly improve the results of semantic segmentation.
- A novel dilated convolution module which has fewer parameters but better performance than that in original D-LinkNet.
- A novel iterative framework is proposed for road extraction from high-resolution satellite images, which can further refine and enhance the results of road extraction.

2 Related Work

Our work relates to a line of research on semantic segmentation in the field of deep learning. We refer the reader to the recent survey [8] on semantic segmentation in deep learning for discussions on a variety of methods. CNN has achieved unprecedented success in many computer vision tasks, which also provides various powerful tools for semantic segmentation. Fully convolutional networks (FCN) [14] has made milestone progress in image segmentation, which extends the classification at the image level into the pixel level with small storage and high segmentation efficiency. Since then, various FCN-based methods [2,21] continuously refresh the record of semantic segmentation. U-Net [16] adopts a symmetric U-shaped structure, which lays the foundation for the following design of the segmentation network. To increase the receptive field without information loss due to decreased resolution, DeepLab [5,6] is based on dilated convolution and has shown strong abilities to increase the segmentation accuracy.

Inspired by the research of image segmentation, CNN-based methods provide new chances for road extraction. Excellent works [1] have been proposed for road extraction. Mattyus et al. [15] develop a variant of FCN using ResNet as an encoder with a fully deconvolutional decoder to estimate road topology. Zhang et al. [20] introduce a deep residual U-Net for road extraction. Skip connections are used to obtain improved performance by information propagation. Bastani et al. [3] propose RoadTracer to extract road network using an iterative search process guided by a patch-based CNN decision function. Based on this, Lian and Huang [12] develop a road network tracking algorithm for road extraction. Zhou et al. [22] propose D-LinkNet for road semantic segmentation, which contains an encoder–decoder structure with a dilated convolution part in the center. D-LinkNet achieves obvious improvement in road extraction but retains several issues concerning road connectivity and recognition. Based on this, Huang et al. [10] propose D-CrossLinkNet by adding cross-resolution connections in D-LinkNet. Our proposed model also makes full use of the outstanding extraction capability of the D-LinkNet architecture. Long-distance spatial information learning is very important in road extraction. Except for dilated convolution, attention mechanism [9,18,19] can also achieve good results in global information learning. Therefore, we introduce recent popular attention modules [19] into our model for more effective feature extraction, which could also reduce the loss of short-distance spatial features caused by dilated convolution.

3 Methodology

3.1 Overview

We propose a novel semantic segmentation model IDANet, which adopts iterative D-LinkNets with attention modules for road extraction from high-resolution satellite images. The whole network is designed in an iterative architecture, as shown in Fig. 1, which can strengthen the learning of semantic segmentation by

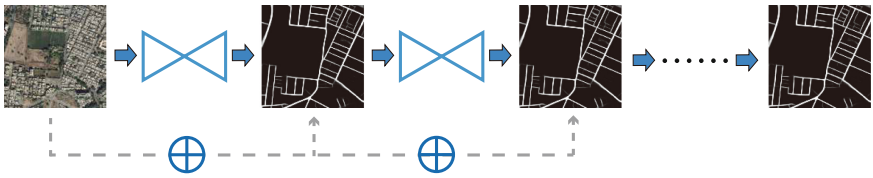


Fig. 1. Iterative architecture. IDANet adopts iterative D-LinkNets with attention modules for road extraction.

fusing the original input and the intermediate result generated in each iteration. IDANet uses D-LinkNet as the basic iteration module, but some effective modifications have been made. First of all, to enhance the effect of the dilated convolution, we modify the dilation module in the original D-LinkNet to make better use of feature information at different levels. We also introduce attention modules into our model for more effective feature extraction. Finally, a powerful loss function with focal loss is adopted to fix the issue of category imbalance. The iteration module is repeated to achieve self-correction and further improve the segmentation output. Therefore, IDANet can extract road information with much better accuracy, connectivity, and completeness than previous methods.

3.2 Basic Iteration Module

D-LinkNet is a classical encoder-decoder network that receives high-resolution images as input. The encoder part and decoder part of our model remain the same as the original D-LinkNet, so our model can also work with high-resolution images. The encoder part reduces the resolution of the feature map through the pooling layers, if an image of size 1024×1024 goes through the encoder part, the output feature map will be of size 32×32. The decoder part uses several transposed convolution layers to do upsampling, restoring the resolution of the feature map from 32×32 to 1024×1024, as shown in Fig. 2.

Dilation Module. Having a large receptive field is important for road extraction, as roads in most satellite images span the whole image with some natural properties such as narrowness, connectivity, complexity. The highlight of D-LinkNet is its focus on increasing the receptive field of feature points by embedding dilation convolutions in the network without decreasing the resolution of feature maps. In the original D-LinkNet, different dilated convolutions are stacked both in cascade mode and parallel mode. This will inevitably lead to the loss of short-distance spatial information due to the "holes" introduced in the computation of dilated convolution. A better approach is to use the original input to further strengthen the output of the dilated convolution to achieve the supplement of information features and decrease the information loss in the computation of dilated convolution. Therefore, we only keep the backbone dilation network where the dilation rates of the stacked dilated convolution layers are 1, 2, 4, 8, respectively, and discard other parallel branch networks. Inspired

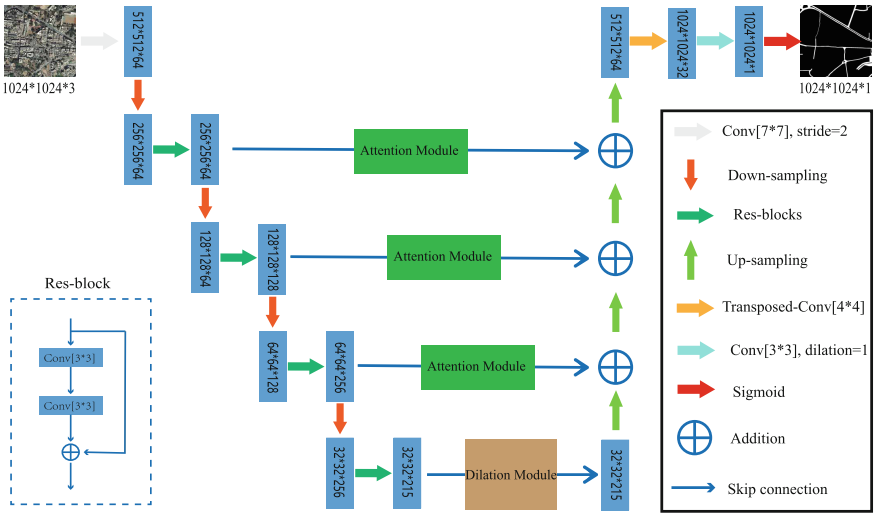


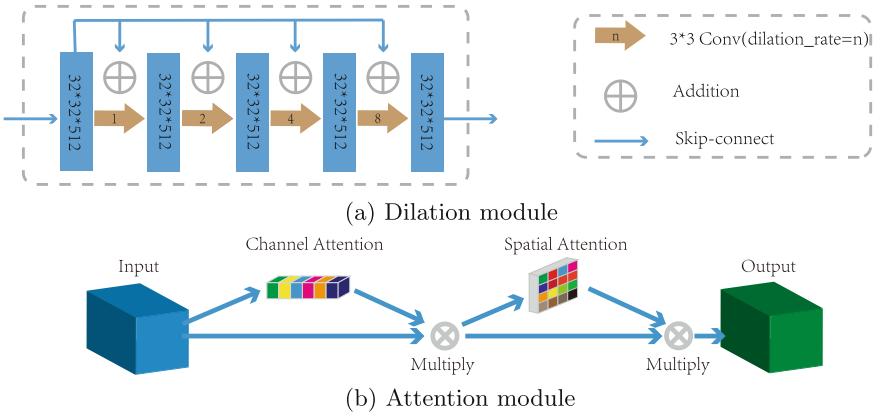
Fig. 2. The basic iteration module including the dilation module and attention module.

by the identity mapping, we also add identity mapping between different dilated convolution layers, as shown in Fig. 3(a).

Attention Module. Upsampling can supplement some lost image information, but it is certainly incomplete. Therefore, the feature map from upsampling of the decoder part and which from the corresponding layer of the encoder part are concatenated together by skip connection which bypasses the input of each encoder layer to the corresponding decoder. This operation can bring in the features of lower convolution layers which contain rich low-level spatial information. To achieve better fusion between these two feature maps, we introduce attention modules into our model. We choose recent popular attention architecture CBAM (Convolutional Block Attention Module) [19] as our attention module, as shown in Fig. 3(b). CBAM can achieve good results in global and long-distance spatial information learning by combining the channel attention module and spatial attention module. Specifically, the channel attention module squeezes the spatial dimension of the input feature map and focuses on which channels are meaningful. The spatial attention module generates a spatial attention map by utilizing the inter-spatial relationship of features and focuses on where is an informative part given an input image. Then these two attention maps are multiplied to the input feature map for adaptive feature refinement.

Loss Function. D-LinkNet uses BCE (Binary Cross Entropy) and dice coefficient loss as loss function. Binary Cross-Entropy performs pixel-level classification and defined as:

$$L_{BCE}(y, \hat{y}) = -(ylog(\hat{y}) + (1 - y)log(1 - \hat{y})) \tag{1}$$



(a) Dilation module

(b) Attention module

Fig. 3. Dilation module(a) and attention module(b) used in IDANet.

Here, $\hat{y}$ is the predicted value. Dice coefficient loss is widely used to calculate the similarity between two images and defined as:

$$DL_{(}y, \hat{p}) = 1 - \frac{2y\hat{p} + 1}{y + \hat{p} + 1} \tag{2}$$

Here, 1 is added to ensure that the function is well defined when $y = \hat{p} = 0$.

However, we have a category imbalance problem, since most of the areas in satellite images are non-road pixels (more than 90% of the area pixels are non-road pixels), which also causes that it is difficult to train sparse samples and seriously affect the training effect. To this end, the focal loss [13] is introduced. It can not only alleviate the imbalance of categories but also down-weight the contribution of easy-training examples and enable the model to focus more on learning hard-training samples. Focal Loss is defined as:

$$FL_{(}p_t) = -\alpha_t(1 - p_t)^{\gamma} log(p_t) \tag{3}$$

$$p_t = \begin{cases} p, & \text{if y} = 1 \\ 1 - p, & otherwise \end{cases} \tag{4}$$

Here, $\gamma > 0$ and $0 \leq \alpha_t \leq 1$. Since the real goal of image segmentation is to maximize IoU metrics, and dice coefficient is calculated based on IoU, dice coefficient loss is especially suitable for the optimization of IoU. Therefore, we adopt focal loss and dice coefficient loss as our loss function:

$$L = \alpha_f FL + \alpha_d DL \tag{5}$$

Here, α_f and α_d are weights for focal loss and dice coefficient loss, respectively.

3.3 Iterative Architecture

Using the above modified D-LinkNet as the basic module, we adopt an iterative architecture to refine the road segmentation and obtain enhanced results, as

shown in Fig. 1. From the perspective of data processing, our iterative algorithm can be regarded to enhance the post-processing of output results. In our iterative architecture, the output of the previous network serves as the input of the next network to refine the road segmentation and obtain enhanced results. The input image of our model is expressed as I, the input of the i^{th} iteration is expressed as I_i, and the output of the i^{th} iteration is expressed as O_i. Each iteration step can be defined as follows:

$$D\text{-}LinkNet(I_{i+1}) \longrightarrow O_{i+1}, i = 1, ..., n \tag{6}$$

I_{i+1} is the splicing result of I and O_i along the channel in the t^{th} basic iterative module, which is defined as follows:

$$I_{i+1} = concat(I, O_i), i = 1, ..., n \tag{7}$$

where n is the number of basic iteration modules. The iterative architecture integrates information through repeated iterative enhancement learning of the splicing results of D-LinkNet output results and original images, which can further refine and enhance the results of road extraction. The proposed iterative architecture is very useful and efficient. The training time for a single iteration module is decreased with the iteration increasing, while impressive performance can be obtained. It contributes about 0.5% at IoU in our experiment.

4 Experiment

4.1 Datasets

We have performed our experiments on two diverse datasets: 1) the DeepGlobe dataset [7] and 2) the Beijing-Shanghai dataset [17], as shown in Fig. 4. The DeepGlobe dataset consists of 6226 annotated satellite images with an image resolution of 1024×1024. Among them, 5226 images are used for training, and the left is used as the testing set. There are 348 satellite images (298 Beijing maps and 50 Shanghai maps) in the Beijing-Shanghai dataset. Each image has a size of 1024×1024. During the experiments, 278 images are used for training, and the left is used as the testing set. For these benchmark datasets, the road labels on the image are manually marked. Specifically, roads are labeled as foreground, and other objects are labeled as background.

4.2 Implementation Details

We use PyTorch to implement and train our networks. All models are trained and tested on an NVIDIA RTX 2080TI with 11 GB memory. Limited by the hardware, we can only train IDANet step by step, which means that the results of the previous network serve as the training data for the next network. During the training process, we find that only the first iteration network takes about 200 epochs to converge, and the following iteration modules have faster convergence

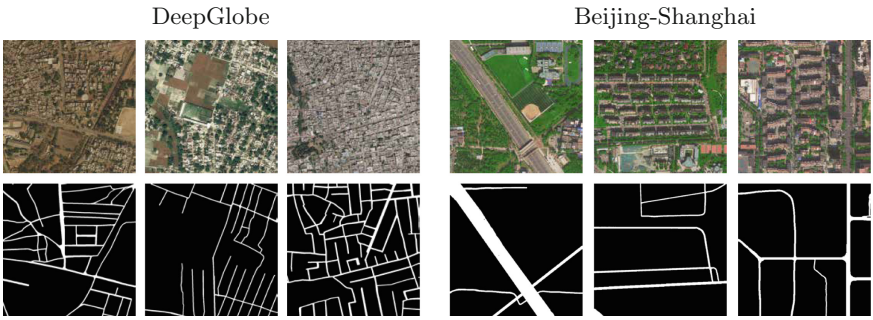


Fig. 4. Visualization of two benchmark datasets.

with less than 100 epochs. The number of the iteration network is an important architecture parameter that can affect the final segmentation results. The number of basic iteration modules is set to $n = 3$, which is an experimental value and enables to obtain the final stable segmentation results. For more discussion and analysis of the iteration number, please refer to Sect. 5.3. In our experiment, the loss weights for focal loss and dice coefficient loss are empirically set for different datasets to achieve the best results. Specifically, $\alpha_f = 30$ & $\alpha_d = 1$ for the DeepGlobe dataset and $\alpha_f = 20$ & $\alpha_d = 1$ for the Beijing-Shanghai dataset in our experiments. We also implement data augmentation similar to D-LinkNet, which can make full use of the limited amount of training data.

5 Results

Accuracy, *Recall*, and *IoU* (Intersection over Union) are commonly used as the evaluation indicators for semantic segmentation. Specifically, *Accuracy* is the ratio of the number of correctly predicted samples to the total number of predicted samples. *Recall* is the ratio of the number of correctly predicted positive samples to the total number of positive samples. *IoU* refers to the ratio between the intersection of the road pixels predicted and the true road pixels and the result of their union. In our experiments, road pixels are labeled as foreground, and other pixels are labeled as background, so we also adopt *Accuracy*, *Recall*, and *IoU* to evaluate the segmentation results at the pixel level.

5.1 Comparison of Road Segmentation Methods

To evaluate our model, we select U-Net [16], LinkNet [4], D-LinkNet [22], and D-CrossLinkNet [10] as competitors. Specifically, U-Net is trained with 7 pooling layers, and LinkNet is trained with pretrained encoder but without dilated convolution in the center part. We also compare our proposed model with the non-iterative version of IDANet, namely BaseNet. We have trained these models on two benchmark datasets. *Accuracy*, *Recall*, and *IoU* of each method on

Table 1. Comparison results on the testing dataset of different models

Dataset	Method	*Accuracy*	*Recall*	*IoU*
DeepGlobe	U-Net	0.9652	0.5637	0.5202
	LinkNet	0.9779	0.7948	0.6681
	D-LinkNet	0.9752	0.8005	0.6700
	D-CrossLinkNet	0.9749	0.8011	0.6710
	BaseNet	0.9802	0.8066	0.6876
	IDANet	**0.9813**	**0.8201**	**0.6921**
Beijing-Shanghai	Unet	0.9359	0.7131	0.5478
	LinkNet	0.9404	0.7458	0.5800
	D-LinkNet	0.9429	0.7282	0.5835
	D-CrossLinkNet	**0.9431**	0.7294	0.5841
	BaseNet	0.9419	0.7502	0.5891
	IDANet	0.9416	**0.7710**	**0.5948**

testing datasets are calculated. The results of these models on different datasets are shown in Table 1.

For the DeepGlobe dataset, it can be observed that, compared with U-Net, LinkNet, D-LinkNet, and D-CrossLinkNet, our model (both BaseNet and IDANet) achieves the best performance in all of evaluation metrics. Both our BaseNet and IDANet exceed D-LinknNet and D-CrossLinkNet considerably. Taking the results of D-LinkNet as a baseline, the *Accuracy* improves 0.61%, the *Recall* improves 1.96%, and the *IoU* improves 2.21%. The experimental results on the DeepGlobe dataset are shown Fig. 5.

In terms of the Beijing-Shanghai dataset where the amount of data is much smaller than the DeepGlobe dataset, the results are much poor for all indicators. Even so, IDANet achieves the best performance in both *Recall* and *IoU*. The *Recall* and *IoU* of IDANet are 4.28% and 1.13% higher than the results of D-LinkNet, respectively. For the results of *Accuracy*, IDANet is slightly lower than D-LinkNet's 0.9429 and D-CrossLinkNet's 0.9416. This small difference in *Accuracy* can be attributed to insufficient training data. Furthermore, *Recall* and *Accuracy* are mutually restricted in general. In the case of small data sets, the pursuit of high *Recall* will lead to lower *Accuracy*, which is a normal phenomenon. This result can be understood that our method can further improve the *Recall* and *IoU* while maintaining the *Accuracy* compared with other models.

5.2 Ablation Experiment

This section aims to further certify the effectiveness and universality of the modules introduced in IDANet, including the dilation module, attention module,

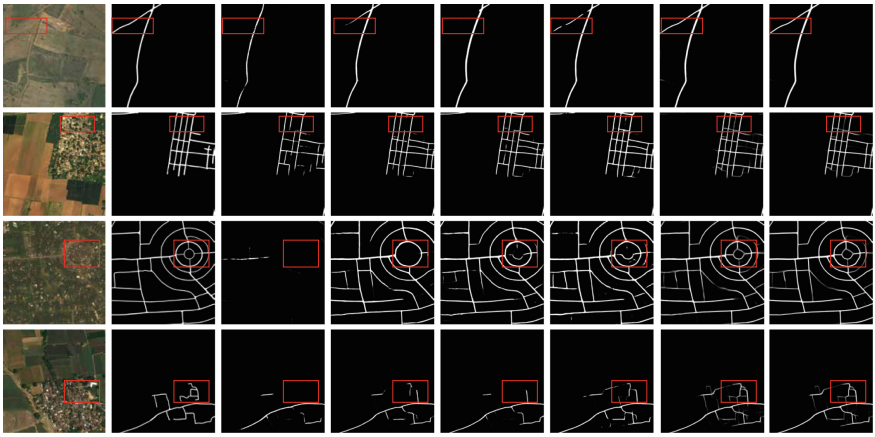


Fig. 5. Visualization results on the DeepGlobe dataset. From left to right: (a) Satellite image, (b) Ground truth, (c) U-Net, (d) LinkNet, (e) D-LinkNet, (f) D-CrossLinkNet, (g) BaseNet, (h) IDANet.

and loss function. Therefore, we have done an ablation experiment on the DeepGlobe dataset. We only perform one iteration for IDANet as well as other models derived from IDANet. We first remove the dilation and attention module from IDANet, denoted as IDANet-D and IDANet-A, respectively. Then we remove both the dilation and attention module from IDANet, denoted as IDANet-DA. We denote the network where BCE loss and dice coefficient loss are used to train our IDANet as IDANet+diceBCE. The results of the ablation experiment are shown in Table 2. For the DeepGlobe dataset, IDANet outperforms other derived versions in terms of the *Accuracy* and *IoU* in the ablation experiment. For the results of *Recall*, the performance of IDANet is slightly lower than IDANet-Dilation. As discussed in Sect. 5.1, *Recall* and *Accuracy* are mutually restricted, and high *Accuracy* results in lower *Recall*, which means that the network is more capable of correcting errors. Compared with IDANet+diceBCE, our loss function can alleviate the imbalance of training samples and focus on learning hard-training samples, which can effectively improve the segmentation results. Experiments show that the dilation module, attention module, and loss function are effective for the semantic segmentation of road extraction.

5.3 The Influence of Network Iteration

In this section, we evaluate the effects of different iterations of IDANet on the performance of road extraction for the DeepGlobe dataset. The results are described in Fig. 6. When the iteration number n is increased, the *IoU* of IDANet is also increasing until the iteration number reaches $n = 3$. After that, the performance of IDANet changes slightly to converge. The results prove that the performance of the model is robust to parameter $n = 3$ for the DeepGlobe dataset and achieves a tradeoff between accuracy and efficiency.

Table 2. Ablation experiment

Method	Accuracy	Recall	IoU
IDANet	**0.9802**	0.8066	**0.6876**
IDANet-D	0.9791	**0.8086**	0.6845
IDANet-A	0.9788	0.8053	0.6842
IDANet-DA	0.9764	0.7999	0.6792
IDANet+diceBCE	0.9758	0.8007	0.6769

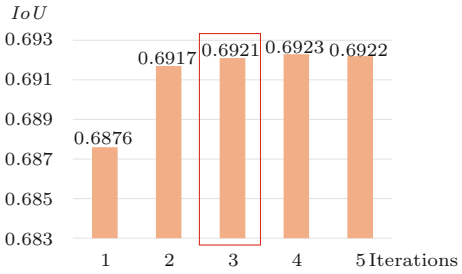


Fig. 6. Different iterations of IDANet

6 Conclusion

This paper aims to improve the accuracy of road extraction from high-resolution satellite images by a novel encoder-decoder network, IDANet, which adopts an iterative architecture to refine the road segmentation and obtain enhanced results. For more effective feature learning, a modified D-LinkNet with attention is proposed as the basic iteration module. We evaluate IDANet on two benchmark datasets. Experimental results show that our model can extract road information from high-resolution satellite images with much better accuracy, connectivity, and completeness than previous methods.

Nevertheless, there is still a big gap between the results we have obtained and the expected. The main issues of current results are the missed identification and wrong recognition. The overall segmentation accuracy is still unsatisfactory. Therefore, our future work will address two aspects. On the one hand, data is always the foundation of deep learning. We plan to adopt more advanced data augmentation techniques and more effective data post-processing methods to enable supervised learning. On the other hand, network is always the key to deep learning. We plan to do more research on the design of refinement networks to realize global and local improvement of semantic segmentation.

References

1. Abdollahi, A., Pradhan, B., Shukla, N., Chakraborty, S., Alamri, A.: Deep learning approaches applied to remote sensing datasets for road extraction: a state-of-the-art review. Remote Sens. **12**(9), 1444 (2020)
2. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
3. Bastani, F., et al.: Roadtracer: automatic extraction of road networks from aerial images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4720–4728 (2018)

4. Chaurasia, A., Culurciello, E.: Linknet: exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE Visual Communications and Image Processing (VCIP), pp. 1–4. IEEE (2017)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2017)
6. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)
7. Demir, I., et al.: Deepglobe 2018: a challenge to parse the earth through satellite images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 172–181 (2018)
8. Hao, S., Zhou, Y., Guo, Y.: A brief survey on semantic segmentation with deep learning. Neurocomputing **406**, 302–321 (2020)
9. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
10. Huang, K., Shi, J., Zhang, G., Xu, B., Zheng, L.: D-CrossLinkNet for automatic road extraction from aerial imagery. In: Peng, Y., et al. (eds.) PRCV 2020. LNCS, vol. 12305, pp. 315–327. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60633-6_26
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural. Inf. Process. Syst. **25**, 1097–1105 (2012)
12. Lian, R., Huang, L.: Deepwindow: sliding window based on deep learning for road extraction from remote sensing images. IEEE J. Sel. Top Appl. Earth Obs. Remote Sens. **13**, 1905–1916 (2020)
13. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
14. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
15. Máttyus, G., Luo, W., Urtasun, R.: Deeproadmapper: extracting road topology from aerial images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3438–3446 (2017)
16. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
17. Sun, T., Di, Z., Che, P., Liu, C., Wang, Y.: Leveraging crowdsourced GPS data for road extraction from aerial imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7509–7518 (2019)
18. Wang, F., et al.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2017)
19. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
20. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. IEEE Geosci. Remote Sens. Lett. **15**(5), 749–753 (2018)

21. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
22. Zhou, L., Zhang, C., Wu, M.: D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 182–186 (2018)