GLOBAL EVOLUTION NEURAL NETWORK FOR SEGMENTATION OF REMOTE SENSING IMAGES

Xinzhe Geng[1,2], *Tao Lei*[1,2*], *Qi Chen*[1,2], *Jian Su*[1,2], *Xi He*[1,2], *Qi Wang*[3], *Asoke K. Nandi*[4]

[1]School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, P. R. China
[2]Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, P. R. China
[3]School of Computer Science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China
[4]Department of Electronic and Computer Engineering, Brunel University London, Uxbridge, Middlesex, UB8 3PH, United Kingdom

ABSTRACT

The popular convolutional neural networks (CNNs) have been successfully used in very high-resolution remote sensing image semantic segmentation. However, these networks often suffer from performance limitations. First, although deeper networks usually provide better feature representation, they may cause parameter redundancy and the inefficient use of prior knowledge. Secondly, attention-based networks often only focus on weighting different features of a single sample but ignore the correlation of all samples in training set, thus leading to the loss of global information. To address above issues, we propose two simple yet effective global evolution strategies. The first is knowledge enhancement. This strategy can reactivate invalid convolutional kernels through convergence of different models and make full use of prior knowledge from the network to improve its feature representation. The second is a dict-attention module that greatly enhances the generalization of networks by learning and inferring the global relationship among different samples through the dictionary unit. As a result, a novel global evolution network (GENet) is designed based on knowledge enhancement and dict-attention for remote sensing image semantic segmentation. Experiments demonstrate that the proposed GENet is not only superior to popular networks in segmentation accuracy.

***Index Terms*—** Deep learning, image segmentation, knowledge enhancement, attention mechanism

1. INTRODUCTION

The development of aviation facilitated the availability of very high-resolution (VHR) remote sensing images, and image semantic segmentation technology is an indispensable and important link. Semantic segmentation of remote sensing images aims to complete the pixel-level classification and segment different objects in remote sensing images according to their semantic interpretability [1]. With the rapid development of deep learning [2], relevant technology has been widely used in remote sensing tasks, such as end-to-end fully convolutional neural networks [3] and U shape networks [4]. But the limitations of these popular networks still exist. First, VHR remote sensing images contain richer useful information, which usually require a more complex network structure. As the layers of a network continues going deeper, the parameter redundancy of the network will be more serious. At present, a lot of studies pay more attention to simplifying the network structure by reducing the network connection [5] or designing compact structure [6]. However, currently there is no research shows that simply abandoning invalid convolutional kernels or connections is the best choice, which leaves the researchers in suspense as to whether these useless parts of the network are effective in a certain extent. Inspired by mutual learning [7], we speculate that it may be feasible to introduce valid information from other external networks to reactivate the dead kernels. Secondly, although attention-based neural networks, such as SENet [8] and Non-local [9], can better model the probability distribution of important features and play a key role in feature selection, they only model some features of a single sample's spatial-wise or channel-wise, which do not fully consider the global correlation among other samples, so leads to loss of global information. To address these issues mentioned above, we mainly made the following three contributions:

(1) We propose a new network learning strategy called knowledge enhancement. This strategy can reactivate the dead convolutional kernels using mutual learning between different networks, and thus improve the feature representation of the proposed network.

*Corresponding author: Tao Lei.

(2) We propose a new attention mechanism called dict-attention, which can effectively obtain the global correlation of samples and thus improve the generalization ability of the proposed network.

(3) Based on knowledge enhancement and dict-attention, we propose a global evolution network (GENet) for remote sensing image semantic segmentation. The experimental results show that the proposed network is superior to the currently popular models, and requires less memory and computational cost.

2. THE PROPOSED NETWORK

Since different semantic scenes in VHR remote sensing images are difficult to distinguish in the same band, the currently networks [10] [11] usually fuse different modals for feature encoding. It is obvious that these networks usually ignore the redundancy of the convolution kernels caused by too complex networks, and do not consider the global correlation among samples. Therefore, making full use of the prior knowledge of the network and the hidden information of the training samples is the key to the evolution of the network.

2.1. Knowledge enhancement

Instead of pruning dead kernels, we tend to reactivate them in a cost-efficient way called knowledge enhancement. As shown in Fig. 1, N_1 and N_2 are two networks from different initial training. In N_1, the parameters of the valid kernel in the i-th convolutional layer L_i are heuristically transfered to the dead kernels at the corresponding position of N_2 using the proposed knowledge enhancement strategy. Then we do the same operation on N_2, and finally get two networks with stronger feature representation ability without any cost of inferencing. We can define the knowledge enhancement as

$$W_i^{N_1'} = \lambda W_i^{N_1} + (1-\lambda)\, W_i^{N_2}, \tag{1}$$

where ${W_i^{N_1}}'$ denotes the parameter of the i-th layer in N_1 after knowledge enhancement, $W_i^{N_1}$ denotes the parameter of the i-th layer in the original N_1, and $W_i^{N_2}$ denotes the parameter of the i-th layer in the original N_2. λ denotes the intensity factor, which represents the strength of knowledge enhancement, and its value is dynamically adjusted in the range of $(0, 1)$ according to the amount of knowledge of the layer in the network. After the above process, $N_1{}'$ and $N_2{}'$ are obtained as the initial network of the next iteration, and $N_1{}''$ and $N_2{}''$ are obtained after subsequent training and knowledge enhancement, and so on.

It is worth noting that how to determine the amount of knowledge of the layer and how to adjust λ are two key points in knowledge enhancement. Since we pay more attention to the variation of the weight, we believe that the size of the information entropy determines the amount of knowledge of that layer. In order to convert the continuous distribution in the each layer into a discrete distribution, we divide the variation range of the parameter value into N regions with the same scale, and then calculate the probability of each region. Since the consistency of the layer, we use the information entropy in a convolutional layer to describe the amount of knowledge of this layer as

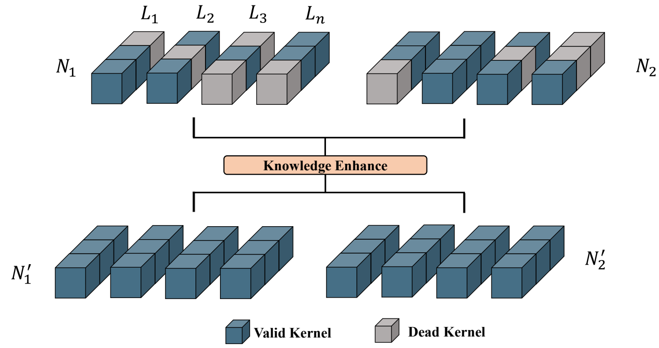

Fig. 1. Knowledge enhancement strategy.

$$K(W_i) = -\sum_{1}^{N} p_j \log p_j, \tag{2}$$

where $K(W_i)$ denotes the knowledge of i-th layer, N denotes the number of regions into which the layer is divided, and p_j denotes the probability of the j-th region. After defining the knowledge of one layer, we propose the method for adapting the weighting factor λ according to the difference of the amount of knowledge,

$$\lambda = 0.5 + A \times \left(\arctan\left(n \times \left(K\left(W_i^{N_1}\right) - K\left(W_i^{N_2}\right)\right)\right)\right), \tag{3}$$

$$n = \frac{100}{1 + e^{-10 \times \left|K\left(W_i^{N_1}\right) - K\left(W_i^{N_2}\right)\right|}}, \tag{4}$$

where A is a hyperparameter, and we set it to $0.4/\pi$ by experiments. n is used to scale the size of the amount of knowledge in the two layers to a reasonable interval. Suppose $K(W_i^{N_1})$ contains more knowledge than $K(W_i^{N_2})$, we believe that there are more valid kernels in the i-th layer in N_1, and λ should be larger than 0.5.

For the case of multiple networks $(N_1, N_2, \ldots, N_k)$, $(N_1', N_2', \ldots, N_{k-1}')$ is obtained through knowledge enhancement from $(N_1, N_2, \ldots, N_{k-1})$, then N_k' is obtained through knowledge enhancement from N_{k-1}' and N_k.

2.2. Dict-attention

In addition to knowledge enhancement, we believe that the effective extraction of hidden information from training samples can make a second order evolution. The current attention mechanism based on self-attention structure only focuses on the long-range relationships of one sample itself. To address this problem, we propose the dict-attention to describe the approximate representation of the same object in different samples by fully obtaining the global correlation of all samples as shown in Fig. 2, with less memory and computational cost. Specifically, let $F_{in} \in R^{N \times c}$ be an input feature map

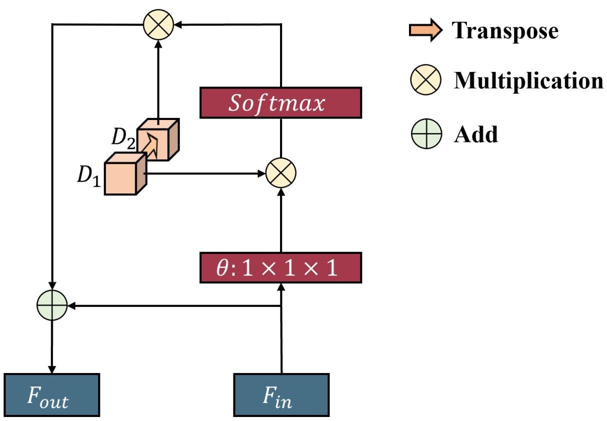

Fig. 2. The structure of dict-attention.

and $F_{out} \in R^{N \times c}$ be an output feature map, where N is the number of feature pixels and c is the number of feature dimensions. We used two dictionary units D_1 and $D_2 \in R^{D \times c}$ to obtain the essential features of samples, they can be analogous to the key and value in self-attention. The input feature map are computed to obtain the attention map inferred from previous samples. Then the feature map is normalized and similarity is computed with D_2 to update the input features. It is worth noting that attention map is often affected by the scale of the input feature, so we added $L1 - norm$ after softmax to normalize its different directions to avoid this problem. The operation can be represented as

$$F_{out} = Softmax\left(F_{in} \otimes {D_1}^T\right) \otimes D_2, \tag{5}$$

where D_1 and D_2 denote to the two dictionary units.

The essence of the dictionary unit is a base space that can represent the essential features of the same class in different samples. By the essential features, we can get any feature representation in the feature space, which is like a dictionary that stores the basic composition of all sentences. Since this base space considers the features of all samples in the dataset, it can evolve the generalization of the network representation.

In addition, compared with the computational complexity of $O(cN^2)$ in self-attention, the computational complexity of dict-attention is only $O(cDN)$ because it only needs two dictionary units to capture the global essential features of the samples, and D is a small hyperparameter. Therefore, dict-attention not only can capture the global features that represent the essence of samples, but also only needs $O(N)$ computational complexity.

2.3. Build global evolution network

As shown in Figure 3, based on the above knowledge enhancement and dict-attention, we build an efficient global evolution neural network (GENet) for VHR remote sensing image segmentation. The network contains encoder and decoder, where the encoder consists of double ResNet40 for feature extraction of DOM and DSM, while another ResNet40 is introduced for feature encoding after the fusion of the two modalities with the channel attention. The decoder part consists of linear interpolation upsampling and deconvolution. We introduce a multi-scale information fusion module Scale-Aware Pyramid Fusion (SAPF) [12], which can help the network to capture multi-scale objects in remote sensing images by adaptively fusing multi-scale features with spatial attention, while using the ghost operation [6] to reduce the overall module's memory and computational cost. Based

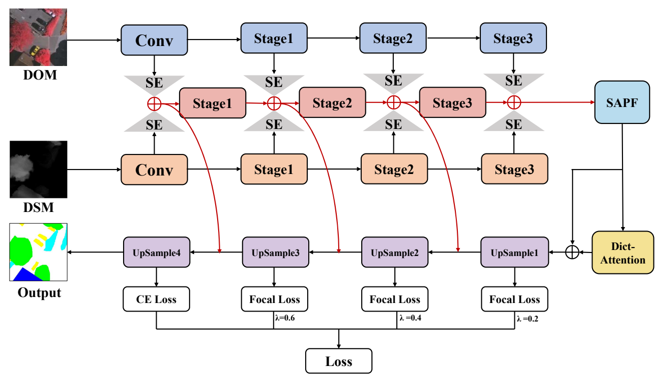

Fig. 3. The detailed structure of GENet.

on this structure, the proposed dict-attention is added to the deepest semantic feature extraction stage of the network as shown in Figure 3. It aims to better describe the approximate representation of the same object in different samples by fully capturing the global relevance of all samples. At the same time, we implement proposed knowledge enhancement strategy throughout the network training phase to reactivate dead kernels and fully exploit the potential expressive ability of the network.

3. EXPERIMENTS

To evaluate the performance of the proposed GENet in the task of semantic segmentation of VHR remote sensing images, we consider the ISPRS Vaihingen dataset [18] as experimental data. This dataset is provided by Committee III of ISPRS, consists of VHR true orthophoto (TOP) slices, DSM and a total of 33 images with an average size of 2494 × 2064 pixels are given. In 16 of these images have ground truth for 6 categories (impervious surfaces, buildings, low vegetation, trees, cars, and others). We scientifically selected 5 images numbered 11, 15, 28, 30, and 34 as the testing set and the

Table 1. Quantitative comparisons with different networks

Method	IoU(%)					MIoU(%)	OA(%)
	Car	*Building*	*Low veg.*	*Tree*	*Imp.surf.*		
DeepLabV3+ [13]	64.95	81.90	50.05	65.88	77.19	72.50	85.11
SCAttNet [14]	54.44	82.32	**66.73**	67.09	**80.40**	74.64	85.47
ACNet [15]	65.97	87.60	50.54	66.47	72.58	73.48	88.20
S-RA-FCN [16]	-	-	-	-	-	-	89.23
DSMFNet [17]	65.88	86.32	53.78	69.43	73.06	74.56	89.80
GENet(Ours)	**66.51**	**88.93**	58.52	**70.26**	78.35	**76.34**	**90.72**

remaining images for training and validating the model. We used data augmentation to enhance the model robustness. Experiments were performed on a server with NVIDIA GeForce RTX 3090 24GB and PyTorch 1.7.

3.1. Training details

Training hyper-parameters are set as follows. the batch size is set to 32. As for dict-attention, we set D to 64. As for knowledge enhancement method, we set the initial learning rate to 0.001 and 0.0001 for the two networks N_1 and N_2 parallelly, and the optimizers are SGD and ADAM, where weight decay of N_1 is set to 5e-4 while N_2 has no weight decay. We used CE loss as the main loss function and focal loss as the auxiliary loss function in order to enhance the network's ability to classify low-probability categories and to avoid network gradient disappearance, the final loss function is defined as

$$L_{total} = L_{ce} + \sum_{i=1}^{3} \eta_i L_i, \tag{6}$$

where L_{total} is the final loss function, L_{ce} is the main loss function, L_i is the auxiliary loss functions, η_i denotes to weight coefficients, i denotes the level of layers, $\eta_1 = 0.6$, $\eta_2 = 0.4$, and $\eta_3 = 0.2$. Training will be stopped after 200 epochs and the model parameters will be saved for further evaluation.

3.2. Evaluetion and results

We use two metrics IoU and OA to evaluate the segmentation quality of each network comprehensively in this experiment. Table 1 shows the performance on the test set using methods [13] [14] [15] [16] [17] and the proposed GENet. Since IoUs are not showed in [16], we only compare with OA provided for fairness. Obviously, our proposed GENet can achieve MIoU of 76.92% and OA of 90.55%, which are higher than other popular networks except for low vegetation and impervious ground , in which the IoU is slightly lower than the score given in paper [14]. Moreover, as shown in Fig. 4, the segmentation results from the proposed GENet show more complete edges and better segmentation results.

To verify the effectiveness of different modules in our proposed network, we conducted a series of ablation experiments

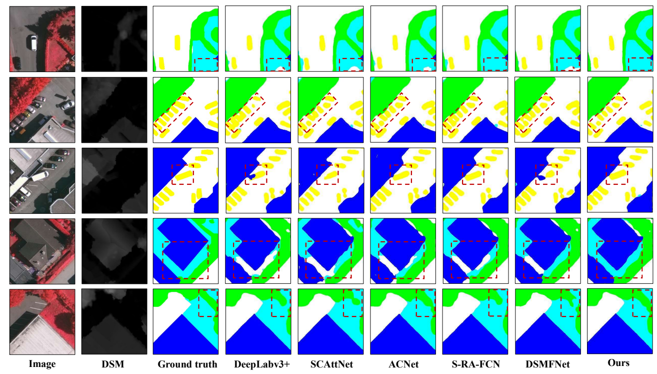

Fig. 4. Visual comparison of different methods.

on ISPRS Vaihingen dataset. As shown in Table 2, both the knowledge enhancement and dict-attention can improve the segmentation accuracy. The combination of the two evolutionary strategy for networks can achieve the highest segmentation accuracy.

Table 2. Ablation experiments

Backbone	Knowledge enhancement	Dict-attention	MIoU(%)	OA(%)
ResNet40+SAPF			74.21	88.99
ResNet40+SAPF	✓		75.56	90.52
ResNet40+SAPF		✓	75.25	90.06
ResNet40+SAPF	✓	✓	**76.34**	**90.72**

4. CONCLUTION

In this paper, we propose GENet for the classification task of VHR remote sensing images. The proposed GENet effectively solves two common problems existing in popular networks used for remote sensing image segmentation: excessive redundancy and the global information neglect of the samples. Specifically, the proposed knowledge enhancement significantly enhances feature representation of networks by reactivating the dead kernel. The proposed dict-attention can effectively extract the global features of the samples with low memory and computational cost. The experimental results show that our proposed GENet is superior to other currently popular networks for remote sensing image segmentation.

5. REFERENCES

[1] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–1934.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[3] Tao Lei, Yuxiao Zhang, Zhiyong Lv, Shuying Li, Shigang Liu, and Asoke K Nandi, "Landslide inventory mapping from bitemporal images using deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 6, pp. 982–986, 2019.

[4] Rongsheng Dong, Xiaoquan Pan, and Fengying Li, "Denseu-net-based semantic segmentation of small objects in urban remote sensing images," *IEEE Access*, vol. 7, pp. 65347–65356, 2019.

[5] Yihui He, Xiangyu Zhang, and Jian Sun, "Channel pruning for accelerating very deep neural networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1389–1397.

[6] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu, "Ghostnet: More features from cheap operations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1580–1589.

[7] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu, "Deep mutual learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4320–4328.

[8] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[9] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[10] Tao Lei, Linze Li, Zhiyong Lv, Mingzhe Zhu, Xiaogang Du, and Asoke K Nandi, "Multi-modality and multi-scale attention fusion network for land cover classification from vhr remote sensing images," *Remote Sensing*, vol. 13, no. 18, pp. 3771, 2021.

[11] Tao Lei, Jie Wang, Hailong Ning, Xingwu Wang, Dinghua Xue, Qi Wang, and Asoke K Nandi, "Difference enhancement and spatial-spectral nonlocal network for change detection in vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2021, doi:10.1109/TGRS.2021.3134691.

[12] Shuanglang Feng, Heming Zhao, Fei Shi, Xuena Cheng, Meng Wang, Yuhui Ma, Dehui Xiang, Weifang Zhu, and Xinjian Chen, "Cpfnet: Context pyramid fusion network for medical image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 10, pp. 3008–3018, 2020.

[13] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[14] Haifeng Li, Kaijian Qiu, Li Chen, Xiaoming Mei, Liang Hong, and Chao Tao, "Scattnet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 5, pp. 905–909, 2020.

[15] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang, "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1440–1444.

[16] Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 7557–7569, 2020.

[17] Zhiying Cao, Kun Fu, Xiaode Lu, Wenhui Diao, Hao Sun, Menglong Yan, Hongfeng Yu, and Xian Sun, "End-to-end dsm fusion networks for semantic segmentation in high-resolution aerial images," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 11, pp. 1766–1770, 2019.

[18] Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez, and Uwe Breitkopf, "The isprs benchmark on urban object classification and 3d building reconstruction," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences I-3 (2012), Nr. 1*, vol. 1, no. 1, pp. 293–298, 2012.