A context and semantic enhanced UNet for semantic segmentation of high-resolution aerial imagery

A context and semantic enhanced UNet for semantic segmentation of high-resolution aerial imagery

Fang Wang[1,3] and Jindong Xie[1,2*]

[1] School of Electronic and Information Engineering, Beihang University, Beijing, 100191, China

[2] Hefei Innovation Research Institute, Beihang University, Hefei, Anhui, 230013, China

[3] AVIC Aerospace System Co., Ltd., Beijing, 100028, China

[*]Corresponding author's e-mail: jindong.xie@buaa.edu.cn

Abstract. Semantic segmentation of high-resolution aerial images is of paramount importance in a wide range of remote sensing applications. The ever-increasing spatial resolution of aerial imagery brings about two specific challenges that incur labelling ambiguities: intra-class heterogeneity and inter-class homogeneity. To address these two challenges, a novel end-to-end semantic segmentation network for high-resolution aerial imagery, namely Context and Semantic Enhanced UNet (CSE-UNet), is proposed in this paper. Specifically, we exploit multi-level Receptive Field Block (RFB) based skip pathways to enhance the representational power of multi-scale contextual information, and therefore tackle the issue of intra-class heterogeneity. To solve the inter-class homogeneity problem, we propose a dual-path encoder where an auxiliary multi-kernel based feature encoding path is embed to produce strong semantic features at all levels to enlarge the inter-class differences. Experimental results shows that our proposed CSE-UNet achieves competitive performance and outperforms UNet and several other deep networks on the ISPRS Potsdam and Vaihingen datasets.

1. Introduction

With the rapid technological advancement of the aerial platform and airborne sensor, the increased accessibility to high-resolution aerial images has opened up new horizons in various remote sensing applications [1], such as intelligent agriculture, urban planning and disaster management. Towards automatic interpretation of aerial imagery, semantic segmentation (i.e., semantic labelling) is a vital step to extract valuable information from the region of interest by inferring every pixel in the image with the category of the object instance.

High resolution aerial imagery provides detailed structural information and exhibits high diversity of objects on the land surface, which leads to the issues of intra-class heterogeneity and inter-class homogeneity. On the one hand, intra-class heterogeneity means that two objects belonging to the same category but with different visual appearances or characteristics should be assigned the same semantic label. For instance, figure 1(a) shows that the colors of cars vary widely. On the other hand, inter-class homogeneity is defined that two ground objects that have different semantic labels but with similar appearances ought to be categorized into two separate classes. In figure 1(b), buildings and impervious surfaces belonging to different categories are very similar in appearance.

The intra-class heterogeneity issue is mainly derived from the lack of contextual information [2-3]. Multi-level receptive fields and contextual information are useful to learn the discriminating features with the aim of correctly categorizing objects with large intra-class variances. Likewise, the problem of inter-class homogeneity results from poor semantic information at all levels [4-5]. To remove the semantic ambiguity, encoding feature maps with strong multi-level semantics provides an effective way to segment similar objects belonging to different categories.

(a) (b)

Figure 1. Examples of intra-class heterogeneity and inter-class homogeneity in aerial images.

Nowadays, numerous Deep Convolutional Neural Network (DCNN) based methods have been proposed in the field of semantic segmentation. Fully Convolutional Network (FCN) [6] becomes the first end-to-end, pixel-to-pixel DCNN-based segmentation network. Subsequently, the encoder-decoder structures, such as SegNet [7], UNet [8] and GCN [9], are further studied, and these methods have made their way into the aircraft-based remote sensing domain [10-12]. As the representative model, UNet utilizes skip connections to concatenate the upsampled features from the decoder with corresponding downsampled feature maps from the encoder at every level.

Regarding UNet, there exists two shortcomings that make the issues of intra-class heterogeneity and inter-class homogeneity remain challenging, which limits its application to semantic segmentation of high-resolution aerial images. On the one hand, UNet encodes insufficient multi-level contextual information, and therefore generates less discriminating features to decrease the intra-class variances. On the other hand, UNet fails to enlarge the inter-class differences due to its incapability of fully exploring semantic information of all levels.

In this paper, we present a novel deep convolutional model for semantic segmentation of high-resolution aerial images, namely Context and Semantic Enhanced UNet (CSE-UNet), to effectively resolve the two above-mentioned challenges. In order to overcome the intra-class heterogeneity problem, we develop multi-level Receptive Field Block (RFB) based skip pathways to strengthen the representational capacity of the network for multi-scale contextual features, by means of exploiting varying convolution kernels and dilated convolutions to adjust the sizes and eccentricities of different receptive fields respectively. Moreover, we equip the model with a dual-path encoder where an auxiliary multi-kernel based feature encoding path is embed to extract and fuse multi-level features with rich semantic information to tackle the issue of inter-class homogeneity.

To summarize, the contributions of this article are three-fold. (i) We propose CSE-UNet, an end-to-end semantic segmentation architecture for high-resolution aerial imagery. (ii) We propose multi-level RFB-based skip connections to encode abundant contexts to mitigate the problem of intra-class heterogeneity. (iii) We develop a multi-kernel dual-path encoder to extract abundant semantic information at all levels to tackle the inter-class homogeneity issue.

2. Method

2.1. Network architecture

The architecture of CSE-UNet is illustrated in figure 2. In high-resolution aerial images, multi-level contextual and semantic information is essential for resolving the issues of intra-class heterogeneity and inter-class homogeneity respectively [2-5]. Therefore, we equip UNet with the RFB-based skip

connections to encode sufficient multi-level contextual information. We also apply the dual-path encoder which includes the auxiliary multi-kernel based feature encoding path to extract and fuse features with rich semantics at both of high and low levels.

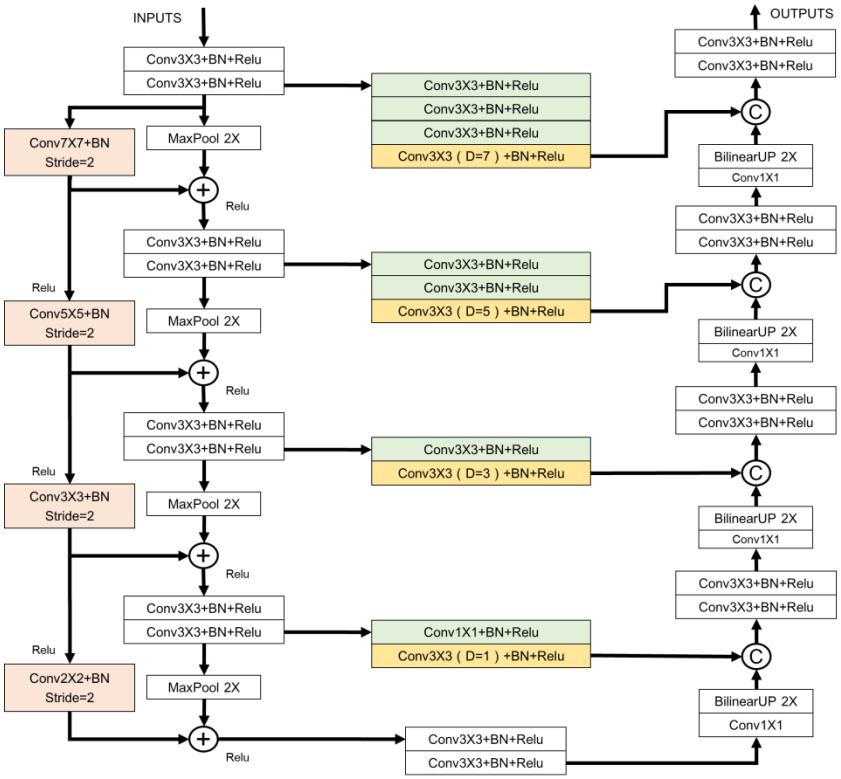

Figure 2. The architecture of CSE-UNet.

2.2. Multi-level RFB-based skip pathways

Motivated by RFBNet [13] that is based on the concept of receptive fields in human visual systems, we propose multi-level RFB-based skip pathways to produce strong multi-scale contextual features from the encoder. Then these feature maps are concatenated with the corresponding upsampled ones in the decoder. Thus, the representational power of the model for contextual information is reinforced.

It is stated that the size of the receptive field is a function of eccentricity in the retinotopic map of human, and the receptive field size combined with the appropriate eccentricity is effective in highlighting informative regions [13]. In addition, the study on RFBNet proposed that the convolutional kernel size and dilation rate have a similar positive functional relation as that of the size and eccentricity of the receptive field in the human visual cortex [13]. Therefore, we take advantage of deep convolutional layers with varying kernels corresponding different sizes of the receptive fields, and we apply dilated convolution layers with different dilation rates to control the eccentricities.

Specially, we introduce a stack of three and two convolution layers with 3 × 3 kernel in replace of 7 × 7 and 5 × 5 convolution kernels separately, and one 3 × 3 dilated convolution kernel with dilation rate set to 7 and 5 accordingly to match the receptive field sizes of 7 × 7 and 5 × 5. Similarly, one convolution layer with 3 × 3 kernel and one dilated convolution kernel with dilation rate of 3, together with one convolution layer with 1 × 1 kernel and one dilated convolution kernel with dilation rate of 1 are also applied to match the receptive field sizes of 3 × 3 and 1 × 1 respectively.

2.3. Multi-kernel dual-path encoder

To obtain multi-level features with strong semantic information, we employ a dual-path encoder that contains an auxiliary multi-kernel based feature encoding path to provide additional rich semantics during the downsampling process in the encoder part of the CSE-UNet.

To better encode multi-level semantic information, we employ a 4-level hierarchy in the auxiliary feature encoding path where strided convolutions are applied as the downsampling method. We use 7 × 7, 5 × 5, 3 × 3 and 2 × 2 kernels from low to high levels respectively. At each level, feature outputs from the original encoding path of UNet and the auxiliary multi-kernel based feature encoding path are fused via element-wise addition to generate feature representations with rich semantics at all level.

3. Results and discussion

3.1. Datasets

Our experiments are based on the ISPRS Potsdam and Vaihingen datasets. There are six object categories in both of the datasets, comprising impervious surfaces, buildings, low vegetation, trees, cars and clutter/background. The clutter/background class accounts for very little percentage of pixels, thus we only select the other five categories in our experiments. In addition, 24 true orthophoto (TOP) files from Potsdam that are provided with ground truth are used for training and validation. 16 TOP tiles from Vaihingen that involve labelled ground truth are used to train and validate the model.

3.2. Experiment details

We use several metrics including F1 score (*F1*), mean F1 (*mF1*), intersection over union (*IoU*), mean IoU (*mIoU*) and overall pixel accuracy (*OA*) to comprehensively assess the model performance.

To split the training and validation sets, we follow the protocol proposed in [11]. For Vaihingen, 11 out of the 16 annotated images are used as the training set. The remaining 5 images (with tile IDs 11, 15, 28, 30 and 34) are employed as the validation set. For Potsdam, 18 out of the 24 images are selected for model training. The tile 04-12 is discarded due to possible mislabelling. The remaining 5 images (with tile IDs 02-12, 03-12, 05-12, 06-12 and 07-12) are employed for the validation set.

All the experiments are based on the Pytorch framework, and performed on one GeForce GTX 1080Ti GPU (11GB RAM) with CUDA Toolkit 10.1. Ubuntu 18.04 is used as the operating system.

Images are cropped into 512 × 512 patches for Potsdam and into 256 × 256 for Vaihingen in a non-overlapping fashion for the model training and validation. Random horizontal flipping and random scaling (from 0.5 to 2.0) over the training images are adopted for data augmentation. The maximum number of training epochs is set to 200 for Potsdam, and set to 400 for Vaihingen.

Stochastic gradient descent (SGD) is used with batch size 4, momentum 0.99, weight decay 0.0005 and initial learning rate 0.001 for the training process.

3.3. Ablation study

We step-wise decompose our network and validate the effectiveness of each of the proposed modules. The results are shown in table 1. Experimental results indicate that CSE-UNet which integrates the dual-path encoder and multi-level RFB-based skip pathways collectively achieves the best accuracy in terms of *mIoU* and *mF1* on both datasets. The results also show that equipping UNet with each of the two modules alone achieves better accuracy performance than UNet respectively.

Table 1. Experimental results for ablation study.

Models	Potsdam		Vaihingen	
	mIoU	mF1	mIoU	mF1
UNet	75.88	86.09	63.60	76.96
UNet + Dual-path encoder	75.90	86.10	67.97	80.39
UNet + RFB-based skip pathways	76.16	86.27	68.38	80.76
CSE-UNet	**76.22**	**86.30**	**68.72**	**81.04**

3.4. Results and comparisons with other deep networks

We compare the performance of our proposed CSE-UNet with three other DCNN-based approaches, including UNet [8], FCN-8s [6] and CAN [10], based on the training from scratch.

3.4.1. Experimental results on Potsdam. The accuracy results are reported in table 2. It is observed that the proposed CSE-UNet achieves the highest overall accuracy with respect to *mIoU*, *mF1* and *OA*. In the meantime, the model outperforms the others in terms of per-class *IoU* and *F1* on the categories of impervious surfaces, buildings and trees. The competitiveness of our model is therefore proven.

Table 2. Experimental results for accuracy evaluation on Potsdam.

Model (Backbone)	Imp Suf		Building		Low Veg		Tree		Car		mIoU	mF1	OA
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1			
UNet	78.58	88.00	86.14	92.55	68.38	81.22	66.97	80.22	**79.33**	**88.47**	75.88	86.09	86.11
FCN-8s (VGG-16)	78.43	87.91	86.98	93.04	**68.76**	**81.49**	66.82	80.11	79.07	88.31	76.01	86.17	86.57
CAN (ResNet-50)	74.09	85.12	85.64	92.26	69.14	81.76	66.17	79.64	77.25	87.17	74.46	85.19	84.82
CSE-UNet	**79.40**	**88.52**	**87.34**	**93.24**	68.29	81.15	**67.60**	**80.67**	78.47	87.94	**76.22**	**86.30**	**86.86**

3.4.2. Experimental results on Vaihingen. The accuracy results are reported in table 3. The results in the table are organized in the same manner as in table 2. We observe that the proposed CSE-UNet also achieves the highest overall accuracy with respect to *mIoU*, *mF1* and *OA*. Meanwhile, the model outperforms the others in terms of per-class *IoU* and *F1* on all of the five categories by a considerable margin. Similarly, the numerical results show that CSE-UNet achieves the competitive accuracy performance compared with UNet and other deep networks.

Table 3. Experimental results for accuracy evaluation on Vaihingen.

Model (Backbone)	Imp Suf		Building		Low Veg		Tree		Car		mIoU	mF1	OA
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1			
UNet	71.77	83.57	78.65	88.05	54.26	70.35	69.97	82.33	43.36	60.50	63.60	76.96	81.08
FCN-8s (VGG-16)	71.81	83.59	79.86	88.80	55.49	71.37	67.37	80.50	47.99	64.86	64.50	77.82	81.06
CAN (ResNet-50)	72.36	83.97	81.21	89.63	53.54	69.74	69.50	82.00	55.51	71.39	66.42	79.35	81.56
CSE-UNet	**74.20**	**85.19**	**82.74**	**90.55**	**56.09**	**71.87**	**72.08**	**83.78**	**58.51**	**73.82**	**68.72**	**81.04**	**82.89**

3.5. Visualization Results.

We present the visualization results on both Potsdam and Vaihingen in figure 3.

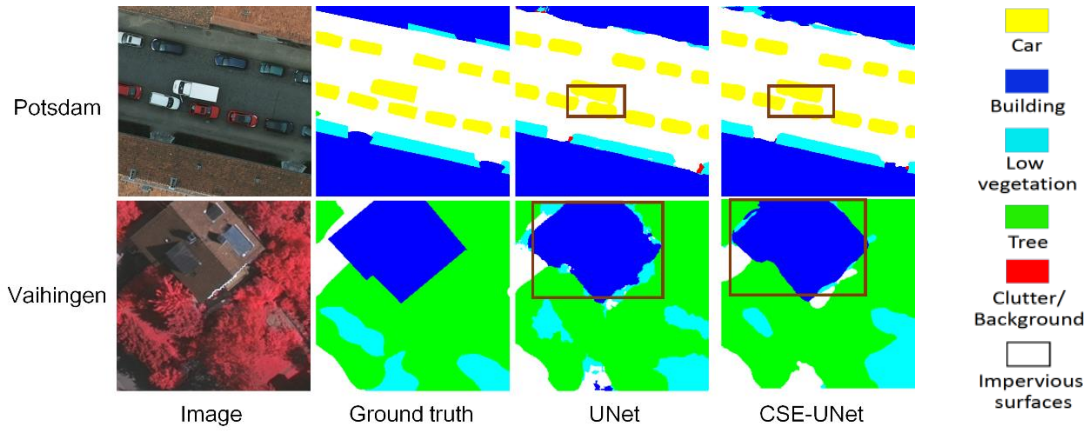

Figure 3. Examples of visual results for CSE-UNet.

It is observed that CSE-UNet obtains more accurate and coherent segmentation results compared with UNet. For Potsdam, UNet fails to fully distinguish the boundaries of cars that have different visual appearances, while CSE-UNet is competent to segment the cars with large intra-class variances. For Vaihingen, the segmentation result of the building outputted by UNet has a jagged edge, while CSE-UNet produces less blurring edge for the building. Meanwhile, areas of trees and low vegetation are better segmented by CSE-UNet. We conclude that owing to the proposed multi-level RFB-based skip pathways together with the multi-kernel dual-path encoder, the issues of intra-class heterogeneity and inter-class homogeneity in high-resolution aerial imagery are significantly alleviated.

4. Conclusion

We have proposed a notable CSE-UNet for semantic segmentation of high-resolution aerial images. The proposed architecture equips UNet with multi-level RFB-based skip pathways and a multi-kernel dual-path encoder to resolve the issues of intra-class heterogeneity and inter-class homogeneity separately. Our experiments demonstrate that CSE-UNet achieves superior performance over UNet and several other deep networks on the ISPRS Potsdam and Vaihingen benchmarks.

References

[1] Toth C and Jó´zków G 2016 *ISPRS J. Photogramm.* **115** 22–36
[2] Mboga N, Georganos S, Grippa T, Lennert M, Vanhuysse S and Wolff E 2019 *Remote Sens.* **11** 5 597
[3] Shang R, Zhang J, Jiao L, Li Y, Marturi N and Stolkin R 2020 *Remote Sens.* **12** 5 872
[4] Russell B, Freeman W, Efros A, Sivic J and Zisserman A 2006 *Proc. 2006 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* vol 2 pp 1605–14
[5] Borenstein E and Ullman S 2008 *IEEE T. Pattern Anal* **30** 12 2109–25
[6] Long J, Shelhamer E and Darrell T 2015 *Proc. 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* vol 1 pp 3431–40
[7] Badrinarayanan V, Kendall A and Cipolla R 2017 *IEEE T. Pattern Anal* **39** 12 2481–95
[8] Ronneberger O, Fischer P and Brox T 2015 *Proc. 2015 Medical Image Computing and Computer-Assisted Intervention (MICCAI)* vol 9351 pp 234–241
[9] Peng C, Zhang X, Yu G, Luo G and Sun J 2017 *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* vol 1 pp 4353–61
[10] Cheng W, Yang W, Wang M, Wang G and Chen J 2019 *Remote Sens.* **11** 10 1158
[11] Liu Y, Minh Nguyen D, Deligiannis N, Ding W and Munteanu A 2017 *Remote Sens.* **9** 6 522
[12] Wei X, Fu K, Gao X, Yan M, Sun X, Chen K and Sun H 2018 *Remote Sens. Lett.* **9** 3 199–208
[13] Liu S, Huang D and Wang Y 2018 *Proc. 2018 European Conf. on Computer Vision (ECCV)* vol 11215 pp 404–419