LOG-CAN: LOCAL-GLOBAL CLASS-AWARE NETWORK FOR SEMANTIC SEGMENTATION OF REMOTE SENSING IMAGES

Xiaowen Ma, Mengting Ma, Chenlu Hu, Zhiyuan Song, Ziyan Zhao, Tian Feng, Wei Zhang

Zhejiang University

ABSTRACT

Remote sensing images are known of having complex backgrounds, high intra-class variance and large variation of scales, which bring challenge to semantic segmentation. We present LoG-CAN, a multi-scale semantic segmentation network with a global class-aware (GCA) module and local class-aware (LCA) modules to remote sensing images. Specifically, the GCA module captures the global representations of class-wise context modeling to circumvent background interference; the LCA modules generate local class representations as intermediate aware elements, indirectly associating pixels with global class representations to reduce variance within a class; and a multi-scale architecture with GCA and LCA modules yields effective segmentation of objects at different scales via cascaded refinement and fusion of features. Through the evaluation on the ISPRS Vaihingen dataset and the ISPRS Potsdam dataset, experimental results indicate that LoG-CAN outperforms the state-of-the-art methods for general semantic segmentation, while significantly reducing network parameters and computation. Code is available at https://github.com/xwmaxwma/rssegmentation.

Index Terms— Semantic segmentation, remote sensing, class representations

1. INTRODUCTION

Semantic segmentation of remote sensing images aims to assign definite classes to each image pixel, which makes important contributions to land use, yield estimation, and resource management [1–3]. Compared to natural images, remote sensing images are coupled with sophisticated characteristics (e.g., complex background, high intra-class variance, and large variation of scales) that potentially challenge the semantic segmentation.

Existing methods of semantic segmentation based on convolutional neural networks (CNN) focus on context modeling [4–7], which can be categorized into spatial context modeling and relational context modeling. Spatial context modeling methods, such as PSPNet [4] and DeepLabv3+ [8], use spatial pyramid pooling (SPP) or atrous spatial pyramid pooling (ASPP) to integrate spatial contextual information. Although these methods can capture the context dependencies with homogeneity, they disregard the differences of classes. Therefore, unreliable contexts may occur when a general semantic segmentation method processes remote sensing images with complex objects and large spectral differences.

Regarding the relational context modeling, non-local neural networks [5] compute the pairwise pixel similarities in the image using non-local blocks for weighted aggregation, and DANet [6] adopts spatial attention and channel attention for selective aggregation. However, the dense attention operations used by these methods enable a large amount of background noise given the complex background of remote sensing images, leading to the performance degradation in semantic segmentation. Recent class-wise context modeling methods, such as ACFNet [9] and OCRNet [10], integrate class-wise contexts by capturing the global class representations to partially prevent the background inference caused by dense attentions. Despite the fact that these methods have achieved ideal performance in semantic segmentation on natural images, the performance on remote sensing images remains problematic, specifically for high intra-class variance that leads to the large gap between pixels and the global class representations. Therefore, introducing local class representations may address this issue.

Given the above observations, we design a global classaware (GCA) module to capture the global class representations, and local class-aware (LCA) modules to generate the local class representations. In particular, local class representations are used as intermediate aware elements to indirectly associate pixels with global class representations, which alleviates the complex background and the high intra-class variance of remote sensing images. Both modules are integrated into LoG-CAN, a semantic segmentation network with a multi-scale design that improves the large variation of scales issue of remote sensing images.

The primary contributions of this paper are summarized as follows:

This work was supported in part by the National Natural Science Foundation of China under Grant 62202421; in part by Zhejiang Provincial Key Research and Development Program under Grant 2021C01031; in part by Ningbo Yongjiang Talent Introduction Programme under Grant 2021A-157-G; and in part by the Public Welfare Science and Technology Plan of Ningbo City under Grant 2022S125.



--> Feature mapping --> Class mapping 🕀 Concatenation 🕀 Element-wise Sum 🛞 Matrix multiplication

Fig. 1. Architecture of LoG-CAN with GCA and LCA modules.

- a novel local class-aware module using the local class representations for class-wise context modeling;
- a multi-scale semantic segmentation network integrating both local and global class-aware modules;
- the state-of-the-art performance on two benchmark datasets for aerial images and a significant reduction of the number of parameters and computational efforts.

2. METHOD

2.1. Overall Architecture

The proposed LoG-CAN has an encoder-decoder architecture (as shown in Fig. 1). The encoder uses ResNet50 [11] as the backbone for multi-scale feature extraction, and the decoder consists of a global class-aware (GCA) module and local class-aware (LCA) modules to refine multi-scale feature representations from the backbone via class-wise context modeling. Specifically, each residual block i of the four extracts multi-scale feature representations \mathcal{R}_i from the input image; the feature representations \mathcal{R}_q from the last residual block are processed by the GCA module to obtain the intermediate global class representations \mathcal{C}'_{q} . Then, each \mathcal{R}_{i} and the i+1-th LCA module's output are processed with feature mapping and concatenation to reach intermediate feature representations \mathcal{R}' . In addition, the feature representations \mathcal{R} and the class representations C_q input to the LCA module are obtained via feature mapping and class mapping from \mathcal{R}' and \mathcal{C}'_{q} . Being refined by the cascaded LCA modules, the feature representations at different spatial scales are element-wisely

summed and quadruply upsampled for the semantic segmentation output.

Note that our design of feature mapping and class mapping, which are implemented respectively by a 3×3 convolution layer and a 1×1 convolution layer, enables the following two effects: (1) the multi-scale feature representations and class representations further interact with each other in a specific feature space after mapping; (2) mapping reduces the feature channels of both representations, creating a lighter structure that contains fewer model parameters and computation without degrading the model performance.

2.2. Global Class-Aware Module

Motivated by [10], we design a GCA module to capture the global class representations. With feature representations $\mathcal{R}_g \in \mathbb{R}^{C' \times H' \times W'}$ that contain rich semantic information, the distribution of class probability \mathcal{D}_g is obtained as follows,

$$\mathcal{D}_q = \mathcal{H}(\mathcal{R}_q),\tag{1}$$

where \mathcal{D}_g is a matrix of size $K \times H' \times W'$ and K is the number of classes. \mathcal{H} is implemented by two consecutive 1×1 convolution layers. Then, the global class representations \mathcal{C}'_g is defined as follows,

$$\mathcal{C}'_g = \mathcal{D}_g^{K \times (H' \times W')} \otimes \mathcal{R}_g^{(H' \times W') \times C'}, \qquad (2)$$

where \mathcal{C}'_{q} is a matrix of size $K \times C'$.

2.3. Local Class-Aware Module

For remote sensing images, class-wise context modeling that only uses the global class representations circumvents the in-

Method	Imp. Sur.	Building	Low Veg.	Tree	Car	AF	mIoU	OA
PSPNet [4]	91.38	94.20	83.05	88.71	75.02	86.47	76.78	89.36
DeepLabv3+ [8]	91.63	94.09	82.51	88.00	77.66	86.77	77.13	89.12
DANet [6]	91.38	94.10	83.09	89.02	76.80	86.88	77.32	89.47
Semantic FPN [12]	91.78	94.37	82.87	89.44	79.45	87.58	77.94	89.86
FarSeg [13]	92.13	94.57	82.87	88.74	81.11	87.88	79.14	89.57
OCRNet [10]	92.87	95.14	84.32	89.23	84.52	89.22	81.71	90.47
LANet [14]	92.41	94.90	82.89	88.92	81.31	88.09	79.28	89.83
BoTNet [15]	92.22	94.48	83.97	89.57	82.93	88.63	79.89	90.16
MANet [16]	93.02	95.47	84.64	89.98	88.95	90.41	82.71	90.96
UNetFormer [17]	92.70	95.30	84.90	90.60	88.50	90.40	82.70	91.00
LoG-CAN (Ours)	93.71	96.64	85.89	90.93	90.16	91.46	84.13	91.97

 Table 1. Effectiveness comparison with the state-of-the-art methods on the test set from the ISPRS Vaihingen dataset. Per-class best performance is marked in bold.

terference of noise caused by intensive attention operations. However, it can potentially lead to considerable semantic differences between pixels and the global class representations due to the insufficient consideration of high intra-class variance, which degenerates the semantic segmentation performance. In this regard, we exploit the local class representations as an intermediate awareness element to capture the relationship between pixels and the local class representations and aggregate this relationship with the global class representations for class-wise context modeling.

For the feature representations $\mathcal{R} \in \mathbb{R}^{C \times H \times W}$, we deploy a pre-classification operation for the corresponding distribution $\mathcal{D} \in \mathbb{R}^{K \times H \times W}$. In particular, we split \mathcal{R} and \mathcal{D} along the spatial dimension to get \mathcal{R}_l and \mathcal{D}_l , followed by calculating the local class representations \mathcal{C}_l as follows,

$$C_{l} = \mathcal{D}_{l}^{(N_{h} \times N_{w}) \times K \times (h \times w)} \otimes \mathcal{R}_{l}^{(N_{h} \times N_{w}) \times (h \times w) \times C}, \quad (3)$$

where h and w represent the height and width of the selected local patch, $N_h = \frac{H}{h}$, and $N_w = \frac{W}{w}$. The corresponding affinity matrix \mathcal{R}_r , which represents the similarity between the pixel and the local class representations, is obtained as follows,

$$\mathcal{R}_r = \mathcal{R}_l^{(N_h \times N_w) \times (h \times w) \times C} \otimes \mathcal{C}_l^{(N_h \times N_w) \times C \times K}.$$
 (4)

Finally, we utilize \mathcal{R}_r to associate the global class representations \mathcal{C}_g and acquire the augmented representations \mathcal{R}_o

$$\mathcal{R}_o = \psi(\mathcal{R}_r^{(N_h \times N_w) \times (h \times w) \times K} \otimes \mathcal{C}_g^{K \times C}), \qquad (5)$$

where ψ is a function that puts the per-local enhanced representations back in place in \mathcal{R} .

3. EXPERIMENTS

We implemented the proposed method and evaluated LoG-CAN on the ISPRS Vaihingen dataset and the ISPRS Potsdam dataset using three common metrics: average F1-score (AF), mean Intersection-over-Union (mIoU), and overall accuracy (OA). ISPRS Vaihingen dataset [18] includes 33 true orthophoto (TOP) tiles and the corresponding digital surface model (DSMs) collected from a small village, where the image size varies from 1996×1995 to 3816×2550 pixels and the ground truth labels comprise six land-cover classes (i.e., impervious surfaces, building, low vegetation, tree, car, and clutter/background). We used 16 images for training and the remaining 17 for testing. ISPRS Potsdam dataset [18] includes 38 TOP tiles and the corresponding DSMs collected from a historic city with large building blocks. All images have the same size of 6000×6000 pixels and the ground truth labels comprise the same six land-cover classes as the ISPRS Vaihingen dataset. We used 24 images for training and the remaining 14 for testing.

3.1. Implementation Details

We selected ResNet-50 [11] pretrained on ImageNet as the backbone for all experiments. The optimizer was SGD with batch size of 8, and the initial learning rate was set to 0.01 with a poly decay strategy and a weight decay of 0.0001. Following previous work [14, 16], we randomly cropped the images from both datasets to produce 512×512 patches, and the augmentation methods, such as random scale ([0.5, 0.75, 1.0, 1.25, 1.5]), random vertical flip, random horizontal flip and random rotate, were adopted in the training process. The number of epochs was set to 150 with the ISPRS Vaihingen dataset.

3.2. Evaluation and Analysis

As shown in Table 1, the proposed method outperformed other state-of-the-art methods on the ISPRS Vaihingen dataset in AF, mIoU, and OA. In particular, our LoG-CAN achieved the AF of 91.46% and the mIoU of 84.13%, even higher than MANet [16] and UNetFormer [17], showing that our design

as:-	s best performance is marked in bold.					
	Method	AF	mIoU	OA		
-	PSPNet [4]	89.98	81.99	90.14		
	DeepLabv3+ [8]	90.86	84.24	89.18		
	DANet [6]	89.60	81.40	89.73		
	Semantic FPN [12]	91.53	84.57	90.16		
	FarSeg [13]	91.21	84.36	89.87		
	OCRNet [10]	92.25	86.14	90.03		
	LANet [14]	91.95	85.15	90.84		
	BoTNet [15]	91.77	84.97	90.42		
	MANet [16]	92.90	86.95	91.32		
	UNetFormer [17]	92.80	86.80	91.30		
-	LoG-CAN (Ours)	93.53	87.69	92.09		

Table 2. Effectiveness comparison with the state-of-the-artmethods on the test set from the ISPRS Potsdam dataset. Per-class best performance is marked in bold.



Fig. 2. Example outputs from the LoG-CAN and other methods on the ISPRS Vaihingen dataset. Best viewed in color and zoom in.

on class-wise context modeling has greater effectiveness. As shown in Table 2, our LoG-CAN also reached outstanding performances in all metrics on the ISPRS Potsdam dataset. Fig. 2 shows example result outputs from our LoG-CAN, PSPNet, and MANet. In particular, the proposed method not only better preserves the integrity and regularity of semantic objects, but also improves the segmentation performance of small objects.

To validate the lightness of our method, we compare our LCA module with several classical context aggregation modules, including the number of parameters measured in million (M), the floating-point operations per second (FLOPs) measured in giga (G), and the memory consumption measured in megabytes (MB). All inputs were set to the size of $2048 \times 128 \times 128$ to ensure the comparison's fairness. As shown in Table 3, the LCA module enables significantly less number of parameters and lower computation compared to PPM [4]. From the perspective of the entire network's structure, our LoG-CAN only needs 60% of the parameters and 25% of the GFLOPs compared to PSPNet [4], which suggests its design as a lightweight method.

We investigated if the number of patches in the LCA mod-

 Table 3. Computational complexity comparison with other popular context aggregation modules. Per-class best performance is marked in bold.

Method	Params (M)	FLOPs (G)	Memory (MB)	
PPM [4]	23.1	309.5	257	
ASPP [8]	15.1	503.0	284	
DAB [6]	23.9	392.2	1546	
OCR [10]	10.5	354.0	202	
PAM+AEM [14]	10.4	157.6	489	
ILCM+SLCM [19]	11.0	180.6	638	
KAM [16]	5.3	85.9	160	
LCA (Ours)	0.8	11.9	53	



Fig. 3. Plot of AF against the number of patches on the ISPRS Vaihingen dataset (yellow) and the ISPRS Potsdam dataset (blue)

ule has any impact on the results. As shown in Figure 3, the best result was obtained on each dataset with the number of patches being set to 16. Besides, when the number of patches was set to 1, the local class representations degenerated to the global class representations, resulting into relatively unsatisfactory performances. These findings indicate that local class awareness can effectively improve class-wise context modeling.

4. CONCLUSION

In this paper, we introduce LoG-CAN for semantic segmentation of remote sensing images. Our method effectively resolves the problems due to complex background, high intraclass variance, and large variation of scales in remote sensing images by combining the global and local class representations for class-wise context modeling with a multi-scale design. According to the experimental results, LoG-CAN has greater effectiveness than the state-of-the-art general methods for semantic segmentation, while requiring less network parameters and computation. The proposed method provides a better trade-off between efficiency and accuracy.

5. REFERENCES

- Georgios N Kouziokas and Konstantinos Perakis, "Decision support system based on artificial intelligence, gis and remote sensing for sustainable public and judicial management," *European Journal of Sustainable Devel*opment, vol. 6, no. 3, pp. 397–397, 2017.
- [2] Xin Huang, Dawei Wen, Jiayi Li, and Rongjun Qin, "Multi-level monitoring of subtle urban changes for the megacities of china using high-resolution multi-view satellite imagery," *Remote sensing of environment*, vol. 196, pp. 56–75, 2017.
- [3] Michele Volpi and Vittorio Ferrari, "Semantic segmentation of urban scenes by learning local class interactions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 1–9.
- [4] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vi*sion and pattern recognition, 2017, pp. 2881–2890.
- [5] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *Proceed*ings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.
- [6] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.
- [7] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu, "Ccnet: Crisscross attention for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 603–612.
- [8] 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.
- [9] Fan Zhang, Yanqin Chen, Zhihang Li, Zhibin Hong, and Errui Ding, "Acfnet: Attentional class feature network for semantic segmentation," *IEEE*, 2019.
- [10] Yuhui Yuan, Xilin Chen, and Jingdong Wang, "Objectcontextual representations for semantic segmentation," in *European conference on computer vision*. Springer, 2020, pp. 173–190.

- [11] 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.
- [12] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár, "Panoptic feature pyramid networks," in *Proceedings of the IEEE/CVF conference on computer* vision and pattern recognition, 2019, pp. 6399–6408.
- [13] Zhuo Zheng, Yanfei Zhong, Junjue Wang, and Ailong Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4096–4105.
- [14] Lei Ding, Hao Tang, and Lorenzo Bruzzone, "Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 426–435, 2021.
- [15] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani, "Bottleneck transformers for visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16519–16529.
- [16] Rui Li, Shunyi Zheng, Ce Zhang, Chenxi Duan, Jianlin Su, Libo Wang, and Peter M. Atkinson, "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.
- [17] Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M Atkinson, "Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 196–214, 2022.
- [18] Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sbastien Bnitez, and U Breitkopf, "International society for photogrammetry and remote sensing, 2d semantic labeling contest," Accessed: Oct. 29, 2020., Available: https://www.isprs.org/education/ benchmarks/UrbanSemLab.
- [19] Zhenchao Jin, Bin Liu, Qi Chu, and Nenghai Yu, "Isnet: Integrate image-level and semantic-level context for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7189–7198.