Class Attention Network for Semantic Segmentation of Remote Sensing Images

Zhibo Rao[*], Mingyi He[*†], Yuchao Dai[*],
[*] Northwestern Polytechnical University, Xian 710129, China
[‡] Email address: myhe@nwpu.edu.cn (Mingyi He)

***Abstract*—Semantic segmentation in remote sensing images is beneficial to detect objects and understand the scene in earth observation. However, classical networks always failed to obtain an accuracy segmentation map in remote sensing images due to the imbalanced labels. In this paper, we proposed a novel class attention module and decomposition-fusion strategy to cope with imbalanced labels. Based on this motivation, we investigate related architecture and strategy by follows. (1) we build a class attention module to generate multi-class attention maps, which forces the network to keep attention to small sample categories instead of being flooded by large sample data. (2) we introduce salient detection, which decomposes semantic segmentation into multi-class salient detection and then fuses them to produce a segmentation map. Extensive experiments on popular benchmarks (e.g., US3D dataset) show that our approach can serve as an efficient plug-and-play module or strategy in the previous scene parsing networks to help them cope with the problem of imbalance labels in remote sensing images.**

I. Introduction

Semantic segmentation is a research spot of computer vision tasks aiming to estimate pixel-wise classification on the images [1]. Due to the booming of deep learning in recent years [2]–[6], the performance of semantic segmentation has made significant achievements [7], [8], promoting various applications, such as object detection [9], [10], autonomous driving [11], [12], and disease diagnosis [13]. It is notable that many applications in remote sensing image segmentation tasks, helping us complete large-scale 3D scene reconstruction [14]–[16].

Many new networks have been improved the accuracy of semantic segmentation [1], [11], [17], which can be divided into the methods based on the receptive field or the attention mechanism.

One way is to enhance the receptive field for modeling the long-range dependencies in convolutional neural networks (CNNs). Zhao *et al.* applied spatial pyramid pooling (SPP) module [18] to exploit the capability of global context information [19]. Kirillov *et al.* used feature pyramid networks (FPN) [7] as a shared backbone and endowed Mask R-NN [1] to yield a lightweight top-performing method for panoptic segmentation [6]. Chen *et al.* proposed an atrous convolution to enlarge the receptive field and put this structure into an encoder-decoder model [11]. Hou *et al.* employed a novel pooling structure called strip pooling to capture both global and local contexts, which is beneficial for models to collect long-range dependencies [20]. These approaches are based on improving the receptive field by using the multi-scale feature or deform convolution.

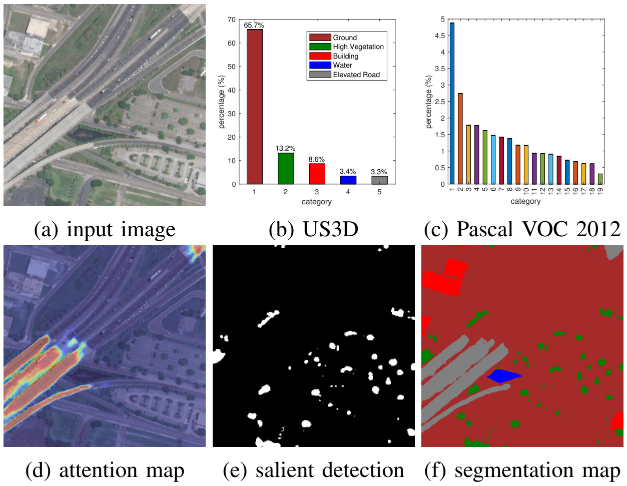
(a) input image (b) US3D (c) Pascal VOC 2012
(d) attention map (e) salient detection (f) segmentation map

Fig. 1: **Illustration of imbalance data and our solution in remote sensing images.** Compared with the traditional semantic datasets (e.g., Pascal VOC 2012 [22]), the remote sensing datasets (e.g., US3D [23]) face a serious problem of imbalance labels, which means data flooding by large sample labels. In this paper, we propose class attention to generate attention maps of each category, as shown in (d). Moreover, we introduce the concept of salient detection that also face imbalance labels. We decompose semantic segmentation task into multi-class salient detection and fuse them to produce a segmentation map, as shown in (e) and (f).

Another way is to collect the region of interest (RoI) by the attention module or multi-task learning. He *et al.* presented a flexible network that can detect objects in an image and generate a high-quality segmentation mask simultaneously [1]. Li *et al.* employed a dual attention structure to distinguish the foreground at the instance level and background at the semantic level. They used this structure to guide the network by object-level and pixel-level attention mechanism [17]. Huang *et al.* proposed a novel criss-cross attention module to exploit the contextual information of all pixels on its criss-cross path [21]. These methods gather vital information to improve accuracy by attention operation.

Although the above methods have succeeded in the traditional semantic segmentation, these methods still exist some limitations in remote sensing images, as shown in Fig. 1. (1) Extreme imbalanced data distribution; Because remote sensing images are obtained at a high altitude by the satellite, most

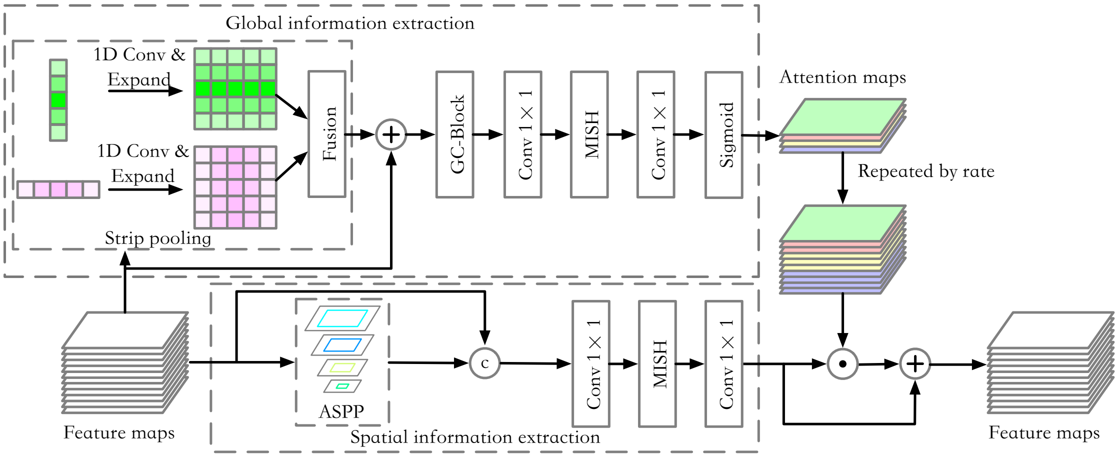

Fig. 2: **Our class attention module, CA module.** The CA module is a parallel structure: the top part extracts global information to generate attention maps, and the bottom part extracts spatial information.

regions are labeled as grounds, and few areas are marked as others (e.g., road, water, building, etc.). It leads that network accessible to ignores the small sample areas. (2) Small targets in the remote sensing images; It is also caused by the high altitude, leading the objects are very small (e.g., building). These factors cause that traditional segmentation methods can't perform well with remote sensing images.

In this paper, we exploit multi-class attention to collect informative contexts for efficiently capturing the small object in the remote sensing images. Then, semantic segmentation is regarded as a multi-object saliency detection to alleviate the problem of data imbalance. In summary, our main contributions are two-fold:

- We design a novel class attention module in this work, which can leverage to capture contextual information and generate multi-class attention maps.
- We introduce salient detection to the semantic segmentation of remote sensing images, which helps network handle imbalanced labels.

II. Our Method

Our approach, class attention network, is a simple network whose goal is to achieve better performance in remote sensing images. We will give a detailed design and motivation, including the entire architecture, loss functions, etc. Moreover, we will discuss why the class attention mechanism can handle the imbalanced data and improve the scene parsing. Note that, in this paper, we replace batch normalization (BN) [24] and rectified linear unit (ReLU) [25] by group normalization (GN) [26] and Mish activate function [27].

A. Backbone Selection

In the recent years, many researchers had done much research about the semantic segmentation of the daily images on the many datasets (e.g., Cityscapes [28], COCO [29], PASCAL VOC [22], etc.). We first briefly review these papers [2], [11], [13], and find that many researchers use feature pyramid networks (FPN) [7] as a standard network. These network architectures consist of the feature with multiple spatial resolutions (e.g., ResNet [30]), and then add a light top-down pathway with lateral connections, as shown in Fig. 3. The top-down starts the deepest layer and uses the de-convolutions to progressively up-sample the features maps while adding the same resolution features from the bottom-top pathway. Thus, FPN generates pyramid features or multi-scale features to promote the performance of semantic segmentation. Through disciplined study, we decide to follow these works and select the FPN as the backbone.

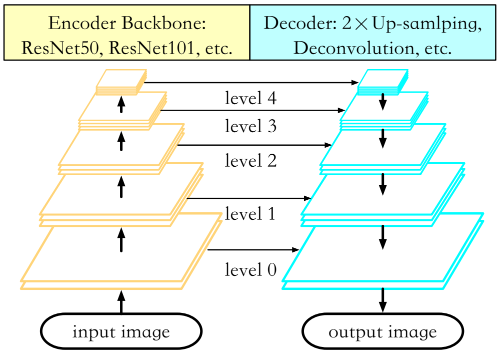

Fig. 3: **Classical feature pyramid network.** We add our class attention module to level 1-4 for enhancing perception.

B. Class Attention Module

In the past deep learning methods, many novel attention mechanisms were introduced to the semantic segmentation. In general, their approaches only generate an attention map, then multiply it by the feature maps, as shown in Fig. 4. This operation can help extract the object from the background. However, when used in multi-objects, this operation tends to distinguish foreground or background instead of providing instance-level attention. Thus, we employ a class attention strategy to observe each object separately in each pyramid structure level.

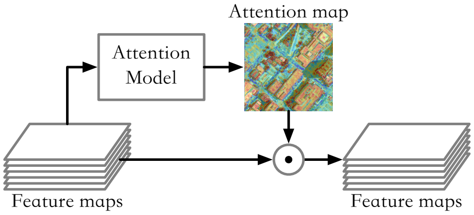

Fig. 4: **An Example of Classical Attention model.** It provides semantic-level attention to images.

In our class attention module, we introduce an effective way to help network capture long-range context and assign different meaningful weights to feature maps, as shown in Fig. 2. The primary motivation is to make the networks pay more attention to the minority class (e.g., road, water, tree, etc.). Our design consists of the parallel structure: the global and spatial information extractions. In this module, we first employ global information to capture long-range context for generating multi-class attention maps. Then, we repeatedly copy and reconstruct attention maps by the ratio. Finally, we apply the reconstructed attention maps to feature maps with spatial information by the attention operations. For the convenience of description, we assume that H_i, W_i, and F_i mean height, width, and channel number of the i-th level feature map.

(1) Global information extraction

To help the network pay attention to small samples' objects, we extract global information and convert them to attention maps. Meanwhile, it has been demonstrated in previous work [20], [31] that collecting long-range dependencies is beneficial to capture global information. Therefore, we first use the strip pooling [20] to build long-range dependencies. Then, we adopt a global context block [31] (GC-block) to capture global information. Finally, we apply two 1D convolutions and a sigmoid function to encode them. After this part, we get each level's multi-class attention maps $G_i \in \mathbb{R}^{H_i \times W_i \times \{C-1\}}$, where C represents the total number of classes.

(2) Attention map reconstruction

In this step, we repeatedly copy multi-class attention maps G_i from $C-1$ to F_i by ratio. To distribute the weight reasonably, we assign a ratio according to the size of the dataset distribution. e.g., in the US3D dataset [23], the different categories account for 13.2%, 8.6%, 3.4%, and 3.3% respectively. Thus, the ratio is close to $4:2:1:1$. Then, we construct the attention maps by this ratio. For example, the channel number of the 5-th level feature maps is 2048, and we can copy each attention map by the ratio of 1024, 512, 256, and 256.

(3) Spatial information extraction

In the bottom branch, we want to build a relationship with various positions in the input feature maps. As we know, enlarging the receptive fields is an excellent way to capture spatial information. Thus, we first use atrous spatial pyramid pooling (ASPP) [11] to enlarge the receptive fields. Then, we concatenate them and the input feature maps, and we adopt 1D convolution to reduce dimension. After this process, we can get the spatial information $S_i \in \mathbb{R}^{H_i \times W_i \times F_i}$.

(4) Global and Spatial information fusion

In this part, we link the attention maps and spatial information by attention operation. The process can be described as:

$$S_i' = S_i \odot G_i \oplus S_i \tag{1}$$

where $\odot$ and $\oplus$ denotes element-wise multiplication and sum respectively, and S_i' indicates the output feature maps.

C. Classification and Fusion

To cope with the challenge of imbalanced data, we covert the multi-classification problem to $\{C-1\}$ binary classification, and then we fuse the result of each class to get the final result. Thus, we have two-steps in this part: $\{C-1\}$ binary classification and fusion.

(1) $\{C-1\}$ binary classification

We can observe that the ground's label number is far more than others from the statistical results. It means the classification of the ground is a simple problem, and it drowns other data. In this case, we choose to ignore these data of ground and only classify different categories (if one point is not in other categories, it must be ground). Therefore, we convert C classification to $\{C-1\}$ binary classification like salient object detection of each class, as shown in Fig. 5. First, we obtain the probability from the feature map by using the sigmoid function. Then, we do a salient object detection (binary classification) for each category, except for the ground.

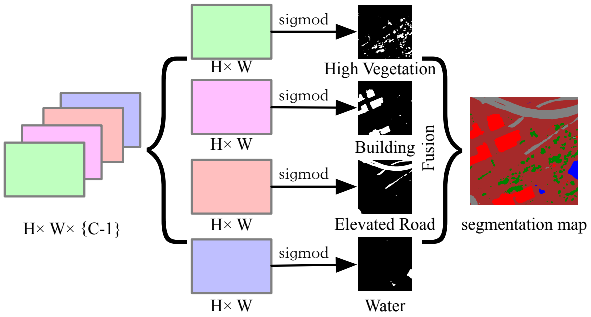

Fig. 5: **The decomposition-fusion strategy in remote sensing images.** We first decompose semantic segmentation into multi-class salient detection and then fuse each category's results to produce a segmentation map.

There are several advantages to encourage us to do this way. First, we detect only four classes, which have a similar number of labels. It can reduce the impact of imbalanced data. Second, we view each small class as a problem of salient object detection, making us use the detection field's loss functions to detect it and improve performance.

(2) Fusion

To obtain the final results, we need to fuse all categories' outputs to get the segmentation map. We follow the fusion guidelines as follows: First, if one point only in one class, we set this point as this class. Second, if one point in two or more categories, we choose the class of the max probability in these classes (this process like ArgMax operation). Third, if one point not in any classes, we set it as ground. After the fusion, we can get a segmentation map.

D. Loss Function

To handle the imbalanced remote sensing data, we introduce the focal loss function [32] to balance the importance of positive/negative examples. First, we view our output as $\{C-1\}$ binary classification instead of traditional methods. Next, we use the focal loss function to calculate the loss of each category and attention maps. Finally, we fuse the different losses of each category and attention maps to get the total loss.

In this case, we first introduce the focal loss function as follows:

$$\mathrm{Focal}(p_i) = \sum_{i=1}^{C-1} -\alpha(1-p_i)^{\gamma}\log(p_i), \quad (2)$$

where α, γ mean the hyper-parameter of the focal loss function, and p_i denotes the distance of i-th class between the estimated probability and label. It can be obtained as follows:

$$p_i = \begin{cases} c_i & \text{if } \mathrm{y_i} = 1 \\ 1-c_i & \text{otherwise}, \end{cases} \quad (3)$$

where c_i means the estimated probability for i-th class and $c_i \in [0,1]$, and y_i denotes the label for i-th class and $y_i \in \{0,1\}$. In this paper, we use the defaults hyper-parameter of focal loss function ($\alpha = 0.25, \gamma = 2$).

Then, we use the focal loss function to calculate the total loss. Therefore, we define our total loss $\mathcal{L}$ as follows:

$$\mathcal{L} = \mathrm{Focal}(B) + \sum_{i=1}^{L=4} \mathrm{Focal}(G_i), \quad (4)$$

where B represents the output of the network, and G_i denotes multi-class attention maps of i-th level in FPN. Here, we resize G_i to the original size to calculate the focal loss.

III. Experiments

To explore the class attention network's performance, we test our method on the large dataset of remote sensing: US3D dataset [23], [33]. First, we introduce our structure's implementation details, as shown in Sec. III-A. Then, we conduct a comprehensive ablation analysis on our approach's effect on the US3D dataset, as presented in Sec. III-B. Moreover, we test our model on the evaluation website to prove our method and reveal our attention maps to demonstrate our approach's effectiveness, as shown in Sec. III-C.

A. Implementation Details

In this paper, we implement our network by Tensorflow [34] with 88.4 M trainable parameters. For achieving the high performance, we introduce the dataset, hyper-parameter choice, learning strategy, and evaluation criteria, as presented in follows:

(1) Dataset

The urban semantic 3D (US3D) dataset [23], [33] is a large-scale public dataset including approximately 100 square kilometer coverage for two large cities of the United States (Jacksonville, Florida and Omaha, Nebraska). The US3D dataset contains four tasks: single-view height estimation, pairwise semantic stereo, multi-view semantic 3D reconstruction, and point cloud semantic segmentation. Because we only do a single task about semantic segmentation. We use the semantic data of single-view height estimation and pairwise semantic stereo, which contains $6,977$ images and the corresponding labels with the size 1024×1024. The classification labels are acquired by LSA specification, and all pixels are classified into the following six categories: ground, high vegetation/trees, building roof, elevated road/bridge, water, and unlabeled.

(2) Hyper-parameter choice

We apply the ADAM optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$) to train the model. For hyper-parameters, we set a mini-batch size of 1 image per GPU (so 2 on 2 GPUs) and the number of category $C = 5$. Moreover, we standardize the input image before feeding into the model, and we apply $\alpha = 0.25$ and $\gamma = 2$ in the focal loss functions. For data augmentation, we first randomly rescale (from 0.5 to 2.0) the input images and use random left-right or transpose flip to process them, then we randomly crop images into 512×512.

(4) Learning strategy

To evaluate our model offline, we divide the data into training data and validation data. We randomly choose $6,477$ images as the training data and 500 images as validation data. Thus, we can use these validation data to choose the best hyper-parameters of the focal loss function. After this process, we employ all images as training data to get the final model and results, and we upload our result to the Codalab website to evaluate our model. For all models in the paper, we use the same learning strategy to train them, which has two stages:

- In the first learning stage, the learning rate is initially set to 1×10^{-3} for 200 epochs on the US3D dataset;
- In the second learning stage, the learning rate is set to 1×10^{-4} for 100 epochs on the US3D dataset;

(4) Evaluation criteria

The evaluation metrics are following previous works. The mean intersection over union (mIou) and pixel accuracy (PixAcc) are employed to assess the performance of the proposed model and other baseline models. The formulas are expressed as follows:

$$\mathrm{mIoU} = \frac{1}{C}\sum_i \frac{tp_i}{tp_i + fp_i + fn_i}, \mathrm{PixAcc} = \frac{n_{correct}}{N_{all}}, \quad (5)$$

where C denotes the number of categories, i means i-th class, tp_i represents the true positive of i-th class, fp_i represents the false positive of i-th class, fn_i represents the true negative of i-th class, $n_{correct}$ expresses the number of correct pixels, and N_{all} is the number of all pixels.

B. Ablations

To verify our design, we run several experiments with different settings to obverse and analyze our networks, including class attention structure, loss functions, etc. Results are shown in Tab. I and discussed in detail next.

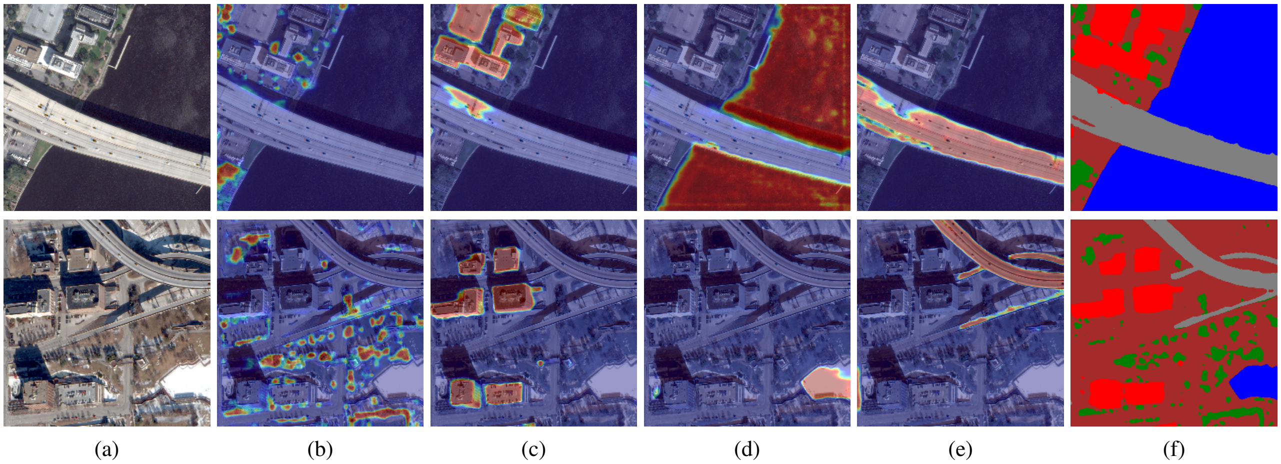

Fig. 6: **The results of our approach on the US3D dataset.** Here, (a) means the input images, (b) - (e) represent the attention maps of level 3, and (f) denotes the segmentation map.

TABLE I: **Evaluation of the CA-Net with different settings.** We compare the mIoU and PixAcc on the US3D validation set. Here, CA means the class attention, and we apply cross entropy loss function (CE) as a contrast. Here, C4 means water, and C5 denotes road.

Setting			**Category (IoU)**		**mIoU**	**PixAcc**
backbone	CA	focal loss	C4	C5		
ResNet-101	-	-	0.849	0.652	0.742	0.923
ResNet-101	√	-	0.934	0.783	0.779	0.941
ResNet-101	-	√	0.927	0.779	0.772	0.938
ResNet-101	√	√	0.956	0.803	0.787	0.953

As presented in Sec. II, our class attention module is based on the FPN, and we choose the classical Res-Net 101 as the backbone. As shown in Tab. I, we list the results of all settings. 1) When no CA module and focal loss function, we achieve a result of 0.742 in terms of mIoU and 92.3% in terms of PixAcc. 2) When we add the CA module, we have an effect of 0.779 and 94.1%, i.e., around 3.7% and 1.8% improvement. 3) When we use the focal loss function to replace the CE loss function, a performance gain of 3.0% and 1.5% can be obtained. 4) Furthermore, if we employ focal loss and CA module together, the model can get the best performance (mIoU: 0.787 and PixAcc: 0.953). It is worth noting that the small samples (e.g., C4: water and C5: elevated road) can get a significant improvement.

C. US3D

Here, we compare the proposed approach with the previous state-of-the-art methods. The results can be found in Tab. II and Fig. 6. As shown in Tab. II, our approach can reach a mIoU score of 0.7801, which is already better than most previous methods, even some fusion methods (e.g., Pop-Net and SDBF-Net). As shown in Fig. 6, we list the heat-maps of our attention map from the CA module in level 3, and it shows that our attention can help the network to understand the scene at a semantic level.

TABLE II: **Performance comparison with other state-of-the-art methods on the US3D.** Here, we get the results of ICNet and DeepLab from the US3D [23], † represents the results we get by running their source code, and $*$ means the methods fuse the additional information (e.g., height, disparity, and etc.)

mIoU	**Category (IoU)**					**mIoU**
	C1	C2	C3	C4	C5	
ICNet	-	-	-	-	-	0.7000
ICNet†	0.7884	0.5357	0.7681	0.9281	0.6605	0.7362
DeepLab v3	-	-	-	-	-	0.7500
DeepLab v3†	0.7992	0.5494	0.7655	0.9131	0.7304	0.7515
Pop-Net	0.8278	0.5851	0.7863	0.8045	0.6520	0.7301
Pop-Net* [16]	0.8053	0.5416	0.7926	0.9438	0.8057	0.7778
SDBF-Net* [14]	0.8104	0.5618	0.7861	0.9175	0.7655	0.7682
CA-Net (Our)	0.8117	0.5605	0.7824	0.9432	0.8026	**0.7801**

IV. Conclusions

In this paper, we present a new module and decomposition-fusion strategy to cope with imbalanced labels of remote sensing images for enhancing performance. First, it allows the model to collect global contextual information by class attention module and promote the model to keep attention to small sample data. Then, we design a decomposition-fusion strategy to cope with imbalanced data as salient detection. Experiments on the US3D dataset demonstrate that the proposed approach can be natural to add the previous methods to improve performance in remote sensing images.

V. Acknowledgment

This work was supported by Natural Science Foundation of China (61671387, 61420106007, and 61871325). The authors would like to thank the Johns Hopkins University Applied Physics Laboratory and IARPAT for providing the data used in this study, and the IEEE GRSS Image Analysis and Data

Fusion Technical Committee for organizing the Data Fusion Contest.

References

[1] K. He, G. Gkioxari, P. Dollr, and R. Girshick, "Mask r-cnn," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018. 1

[2] Y. Li, X. Chen, Z. Zhu, L. Xie, G. Huang, D. Du, and X. Wang, "Attention-guided unified network for panoptic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7026–7035. 1, 2

[3] Z. Rao, M. He, Y. Dai, Z. Zhu, B. Li, and R. He, "Msdc-net: Multi-scale dense and contextual networks for stereo matching," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 578–583. 1

[4] Z. Rao, M. He, Y. Dai, Z. Zhu, B. Li, and R. He, "Nlca-net: a non-local context attention network for stereo matching," *APSIPA Transactions on Signal and Information Processing*, vol. 9, no. e18, pp. 1–13, 2020. 1

[5] Z. Rao, M. He, and Z. Zhu, "Input-perturbation-sensitivity for performance analysis of cnns on image recognition," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 2496–2500. 1

[6] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic Feature Pyramid Networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6399–6408. 1

[7] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944. 1, 2

[8] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," in *The European Conference on Computer Vision (ECCV)*, 2018. 1

[9] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *IEEE international conference on computer vision (ICCV)*, 2017, pp. 5209–5217. 1

[10] J. Zhang, Y. Dai, F. Porikli, and M. He, "Multi-scale salient object detection with pyramid spatial pooling," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018. 1

[11] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *The European Conference on Computer Vision (ECCV)*, 2018, pp. 833–851. 1, 2, 3

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241. 1, 2

[14] Z. Rao, M. He, Z. Zhu, Y. Dai, and R. He, "Sdbf-net: Semantic and disparity bidirectional fusion network for 3d semantic detection on incidental satellite images," in *1*. IEEE, 2019, pp. 438–444. 1, 5

[15] Z. Rao, M. He, Z. Zhu, Y. Dai, and R. He, "Bidirectional guided attention network for 3-d semantic detection of remote sensing images," in *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 2020, pp. 1–16. 1

[16] Z. Zheng, Y. Zhong, and J. Wang, "Pop-net: Encoder-dual decoder for semantic segmentation and single-view height estimation," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2019, pp. 4963–4966. 1, 5

[17] Y. Li, X. Chen, Z. Zhu, L. Xie, G. Huang, D. Du, and X. Wang, "Attention-guided Unified Network for Panoptic Segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–16, 2014. 1

[19] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6230–6239. 1

[20] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng, "Strip pooling: Rethinking spatial pooling for scene parsing," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4003–4012. 1, 3

[21] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 603–612. 1

[22] M. Everingham, S. A. Eslami, L. Van G., C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015. 1, 2

[23] M. Bosch, K. Foster, G. Christie, S. Wang, G. D. Hager, and M. Brown, "Semantic stereo for incidental satellite images," in *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019, pp. 1524–1532. 1, 3, 4, 5

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *arXiv preprint*, 2015. 2

[25] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *International conference on artificial intelligence and statistics*, 2011, pp. 315–323. 2

[26] Y. Wu and K. He, "Group normalization," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19. 2

[27] D. Misra, "Mish: A self regularized non-monotonic neural activation function," in *arXiv preprint*, 2019. 2

[28] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223. 2

[29] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision (ECCV)*, 2014, pp. 740–755. 2

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. 2

[31] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *IEEE International Conference on Computer Vision Workshops*, 2019. 3

[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988. 4

[33] B. L. Saux, N. Yokoya, R. Hnsch, M. Brown, and G. Hager, "2019 ieee grss data fusion contest: Large-scale semantic 3d reconstruction [technical committees]," *IEEE Geoscience and Remote Sensing Magazine*, 2019. 4

[34] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," in *arXiv preprint*, 2016. 4