SEMANTIC SEGMENTATION FOR MULTI-SCENE REMOTE SENSING IMAGES WITH NOISY LABELS BASED ON UNCERTAINTY PERCEPTION

*Xinran Lyu and Libao Zhang**

School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China

ABSTRACT

As the annotation of remote sensing images requires domain expertise, it is difficult to construct a large-scale and accurate annotated dataset. Image-level annotation data learning has become a research hotspot. In addition, due to the difficulty in avoiding mislabeling, label noise cleaning is also a concern. In this paper, a semantic segmentation method for remote sensing images based on uncertainty perception with noisy labels is proposed. The main contributions are three-fold. First, a label cleaning method based on iterative learning is presented to handle noise labels such as missing or incorrect annotations. Second, a two-stage semantic segmentation model is proposed for image-level annotation, which eliminates the need for post-processing steps during testing. Lastly, a complementary uncertainty perception function is introduced to improve the utilization of dataset features and enhance the accuracy of segmentation. The effectiveness of this method was verified through comprehensive evaluation with 7 models on four datasets.

Index Terms— Remote sensing, weak annotation, label noise cleaning, iterative learning, uncertainty perception

1. INTRODUCTION

With the continuous development of remote sensing technology, the application of remote sensing images is becoming increasingly widespread. In recent years, deep learning technology has made significant progress in the field of computer vision, especially in image classification [1], target detection [2], and semantic segmentation [3-4]. This type of method is mainly based on convolutional neural networks (CNNs), which have advantages such as strong generalization ability and accurate segmentation results [5]. Chen et al. [6] proposed an end-to-end integrated full convolutional network to learn features at different scales.

However, the successful application of these technologies often requires a large amount of labeled data [7-8], which undoubtedly increases the cost and difficulty of data processing and system development. In addition, remote sensing images are often affected by various noises, such as sensor noise and atmospheric interference, which further increases the challenges of processing these data.

To address this issue, the weakly supervised semantic segmentation method has emerged as the times require. These methods enhance the value of remote sensing images by utilizing inaccurately annotated or incompletely annotated data for semantic segmentation. Therefore, how to use deep learning techniques to solve the problems of semantic segmentation has become a hot research topic.

Incomplete annotation mainly refers to the use of a small amount of accurately annotated data and a large amount of unlabeled data for training. Zhang et al. [9] proposed a semi-supervised deep semantic segmentation framework. Inaccurate annotation mainly refers to using annotation data with lower accuracy than the target result to train, such as scribble annotation [10-11], image-level annotation [12], etc.

A key step in semantic segmentation methods based on image-level annotation is to achieve spanning from image-level labels to pixel-level labels. Currently, the most widely used method is to calculate the class activation map (CAM) and its related improvement methods [13-15]. Li et al. [16] proposed a building segmentation algorithm, which introduces CAM to produce a pseudo mask. Liu et al. [17] proposed an uncertainty-aware self-attention method to extract regions of interest. Zhang et al. [18] proposed a hierarchical method to achieve salient object detection.

In summary, although some progress has been made in weakly supervised remote sensing image semantic segmentation, there are still some issues that need to be addressed, as follows: 1) Compared with fully supervised methods, there is still a gap in accuracy; 2) Rich feature information contained in remote sensing images has yet to be fully mined; 3) Little consideration is given to label noise.

In this paper, we propose a semantic segmentation method based on uncertainty perception with noisy labels. The main contributions are as follows:

(1) We propose a label cleaning method based on iterative learning for noise labels such as missing or incorrect annotations;

(2) For image-level annotation, we propose a two-stage semantic segmentation model that does not require post-processing steps during testing;

(3) For remote sensing image features, we propose a complementary uncertainty perception function to improve the utilization of features and increase the results.

* Corresponding author: libaozhang@163.com. This work was supported in part by the Beijing Natural Science Foundation under Grant 4222046, in part by the National Natural Science Foundation of China under Grant 62271060, Grant 61571050 and Grant 41771407.

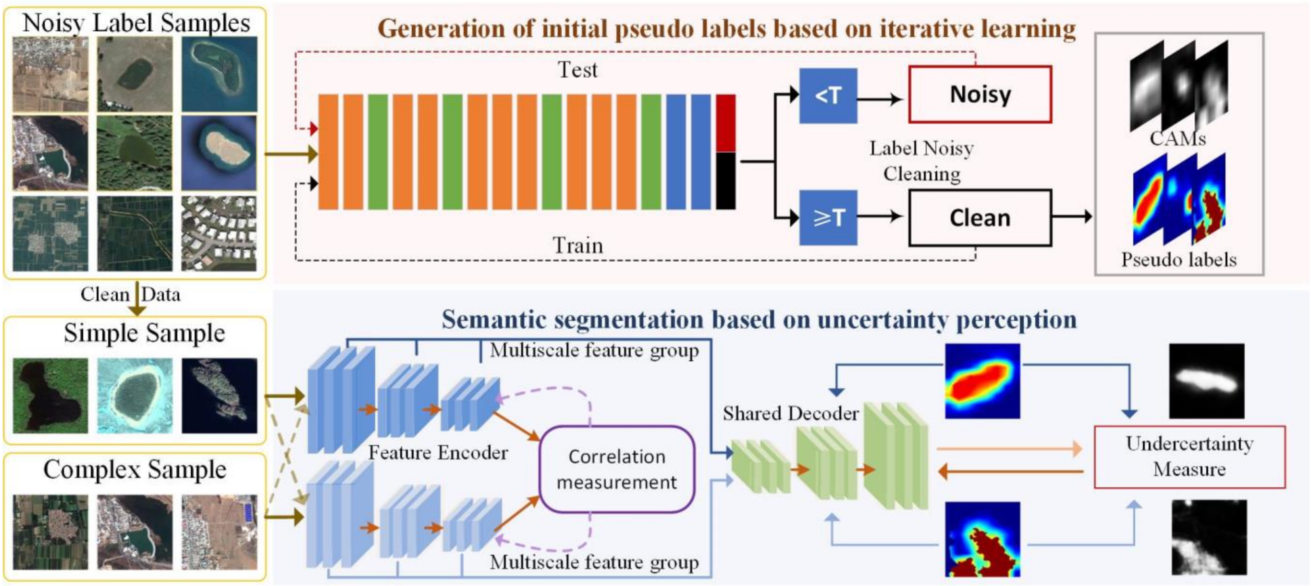

Fig.1. The framework of the proposed model.

2. METHODOLOGY

The overall flowchart of our model is shown in Fig. 1. As shown in Fig. 1, the proposed model mainly contains two parts: scene classification based on iterative learning and semantic segmentation based on uncertainty perception. The main task of the first part is to generate an initial pseudo label and noise cleaning. We are using the superpixel segmentation method to maintain edge details of remote sensing images. The second part utilizes initial pseudo labels for training semantic segmentation tasks. By joint training simple and complex datasets, uncertainty perception is introduced to achieve pixel-level semantic segmentation of remote sensing images.

2.1. Generation of initial pseudo labels based on iterative learning

The main objective of this part is to complete two tasks: the first is to implement label noise cleaning, and the second is to achieve the leap from image-level labels to pixel-level labels. To achieve these two goals, we first constructed a CNN-based scenario classification network and trained it using a noisy label dataset. After the training is complete, we feed the training dataset into the network for testing to obtain a probability distribution of categories. Then we use the robustness of deep neural networks to perform initial label noise cleaning. That is, we refer to data with a class probability greater than the threshold as clean data.

For clean data, we use the Grad-CAM [13] to obtain the initial saliency map for training data. Crop the data in the high confidence area to expand the training data, and input the data to be classified for testing—continuous iterations to eventually achieve noise cleaning on the labels.

For the computation of initial pixel-level pseudo labels, we introduce superpixels to achieve boundary preservation due to the rich details of remote sensing images. Directly calculating the superpixel average may magnify the influence of background noise, as shown in the third row of Figure 2. Therefore, we have designed an uncertain pseudo label generation method.

For a training image X, $\{sp_n\}$ represents superpixels in X, $\bigcup_n sp_n = X$. $Smap$ represents the initial saliency map of image X. S_i is the saliency value of a pixel in $Smap$. Set T_1, T_2 as the threshold. The initial pseudo label Y is calculated as follows:

$$Y(y_i) = \begin{cases} 1, & y_i \in sp_n, 1/n(\sum_{sp_n} S_i) \geq T_1 \\ 0, & y_i \in sp_n, 1/n(\sum_{sp_n} S_i) < T_2 \\ S_i, & otherwise \end{cases} \quad (1)$$

2.2. Semantic segmentation for remote sensing images based on uncertainty perception

The main purpose of this section is to achieve accurate semantic segmentation of remote sensing images. The pixel-level pseudo labels obtained in the previous section have low precision and significant background noise. Additionally, the previous process is complex and requires post-processing steps, resulting in low efficiency for semantic segmentation.

Therefore, we propose a remote sensing image semantic segmentation model based on uncertainty perception, which is trained using the pseudo labels generated in the previous step. Inspired by Li et al. [19], we consider the significant differences in features of different remote sensing objects: in remote sensing images, some scenes are simple, with

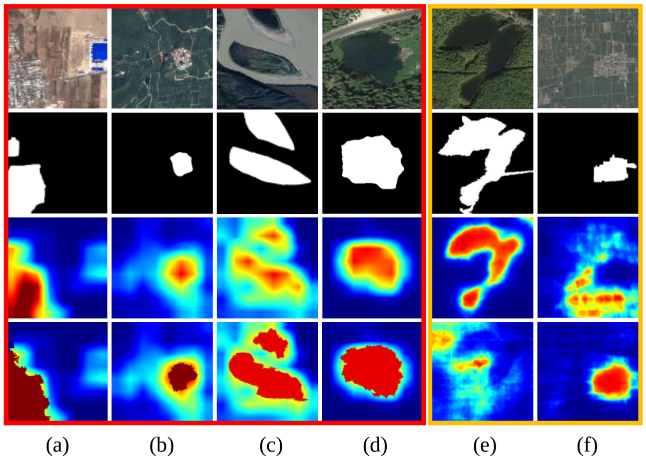

Fig.2 The red box: pseudo pixel-level labels on four datasets. From top to bottom: original remote sensing images, ground-truth, initial saliency maps and pseudo pixel-level labels. The yellow box: the detection results of simple and complex datasets. The first line is the original image, the second line is ground truth, the third line is the segmentation result of the training model on a simple dataset, and the fourth line is the segmentation result of the training model on a complex dataset.

monotonous texture, such as lakes, islands, etc.; but some scenes have complex textures and irregular edge shapes, such as residential areas, etc. And we utilize the uncertainty of different network predictions to improve the quality of segmentation results. So, we divide the dataset into two groups, one called the simple dataset D_s and the other the complex dataset D_c.

This model consists of three parts: a feature encoder, a shared decoder, and an uncertainty analysis module. The feature encoder is a deep neural network based on CNN, such as a residual network, that utilizes intermediate convolutional layers to output multi-scale feature vectors. $F_s = (f_s^1, f_s^2, \cdots, f_s^k)$ is a feature vector of the image in the simple dataset and $F_c = (f_c^1, f_c^2, \cdots, f_c^k)$ is a feature vector of the image in the complex dataset. Then perform data cross input to obtain F_c^s and F_s^c. Wherein, F_c^s represents the simple data features extracted from the encoder trained on a complex dataset. We use the cosine as the loss function to measure the similarity between the two sets of features to train this encoder:

$$L_{similarity} = \cos(F_s^c, F_c^c) = \frac{F_s^c \cdot F_c^c}{\left\|F_s^c\right\| \times \left\|F_c^c\right\|} \qquad (2)$$

Then, build a shared decoder. The decoder consists of two structures, one is top-down feature extraction, and the other is multi-scale feature fusion. Input the feature code to obtain the final segmentation result.

In addition, due to the insufficient accuracy of the initial pixel-level label, we introduce confidence maps C as the reliability of pixels to assist in the calculation of the cross-entropy loss function L_{ce}.

$$C = 2*\left|Y-0.5\right| \qquad (3)$$

$$L_{ce} = C\cdot(Y\cdot\log(P(\text{pre=1}))+(1-Y)\cdot\log(P(\text{pre=0}))) \qquad (4)$$

In addition, it also introduces boundary-IoU loss L_{iou}:

$$L_{str} = 1-\frac{w*(Pre*Y)+1}{w*(Pre+Y)-w*(Pre*Y)+1} \qquad (5)$$

3. EXPERIMENTS

In this section, we conduct our experiments on four remote sensing datasets: two residential area datasets captured by the GeoEye-1 and Google Earth satellite, and lake and island datasets are selected from some classification datasets, both captured by the Google Earth satellite.

To verify the effectiveness of our method, we compared it with several other models in visual comparison and quantitative analysis.

Among these comparison methods, MFF [20] and SACH [21] are traditional visual saliency models designed specifically for remote sensing images. Grad-CAM [13] and Ablation-CAM [15] are typical methods for spanning image-level labels to pixel-level results, which are calculated based on the first part of our method. PSPNet [22] and U-Net [23] are two classic CNN based fully supervised semantic segmentation models. PSL [24] is a weakly supervised semantic segmentation model based on image-level annotation, constructed for remote sensing images.

To be fair, the training effects of cleaned datasets (10% label noise) are shown in 3.1 and 3.2, and the effectiveness of the label noise cleaning function will be shown in 3.3.

3.1. Visual comparison

In this section, we present the semantic segmentation results of different comparison methods and our proposed method, visually demonstrating the advantages and disadvantages of the methods.

From Fig. 3, it can be seen that the traditional visual saliency method, MFF, has achieved very detailed segmentation results, but cannot distinguish between the target and other prominent objects, such as roads. There are many fragments and holes in the segmentation results. The SACH detection results contain too much background interference and cannot segment the target area when the target is not significant. From Fig. 3 (c) and (d), it can be seen that these two unsupervised single image analysis models may also segment opposite regions and cannot perform semantic recognition. Compared to traditional methods, the detection results of Grad-CAM and Ablation-CAM based on weak supervision can eliminate more significant background interference, and there are no holes in the results. But they all have very blurry edges, and the segmentation results are not accurate enough.

Compared with the above methods, the fully supervised methods PSPNet and U-Net have higher accuracy, but due to the high label error used for training, these models do not have good self-correction ability. The weak supervision

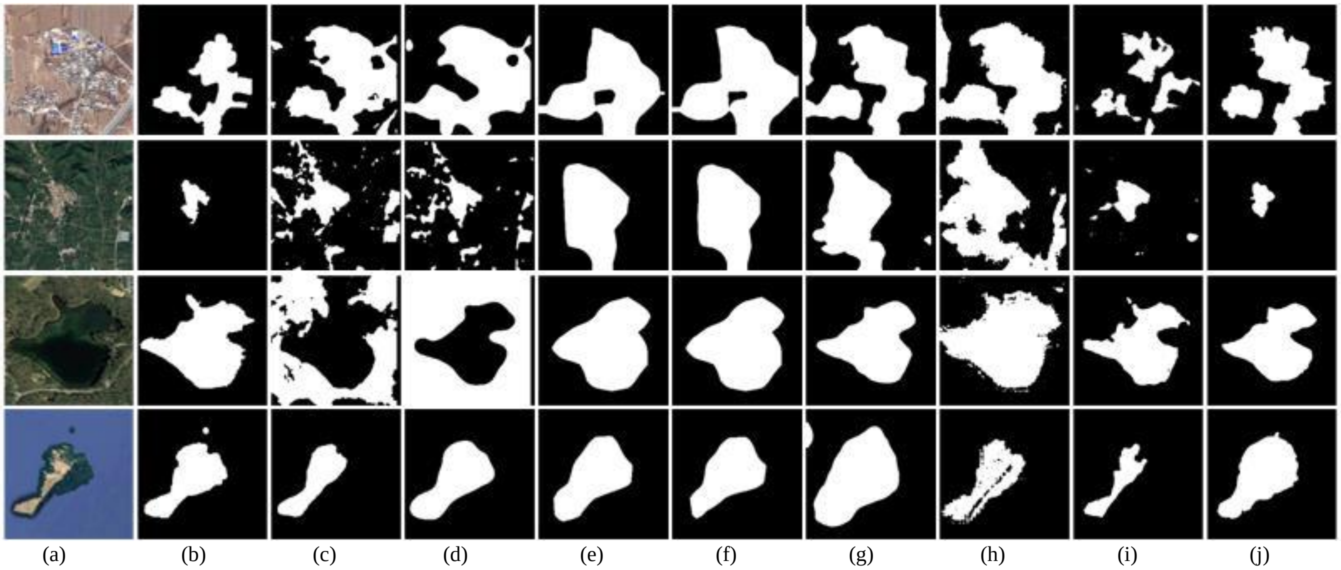

Fig.3. Semantic segmentation on three datasets. The top two rows: residential area, the third row: lake and the fourth row: island. (a) Original images, (b) Ground-Truth, (c) MFF, (d) SACH, (e) Grad-CAM, (f) Ablation-CAM, (g) PSPNet, (h) U-Net, (i) PSL, (j) Ours.

method PSL performs better. However, its recall value is relatively low, especially on island and lake datasets. Relatively speaking, our method has obtained more accurate results.

3.2. Quantitative analysis

In this section, we evaluate the 8 methods in terms of Precision, Recall, F-Measure, Pixel Accuracy (PA), Mean Intersection over Union (MIoU), and Frequency Weighted Intersection over Union (FWIoU).

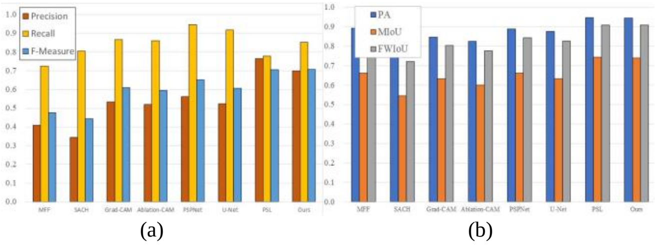

Fig.4. (a) Precision, Recall, and F-Measure values and (b) PA, MIoU, and FWIoU values of the semantic segmentation results. The methods for each image from left to right are as follows: MFF, SACH, Grad-CAM, Ablation-CAM, PSPNet, U-Net, PSL, Ours.

For the semantic segmentation result, we use the Precision, Recall, F-Measure (PRF value), PA, MIoU, and FWIoU (PMF value) to evaluate the segmentation accuracy. Wherein, the F-Measure value is calculated from the weighted sum of Precision and Recall. Figs. 4(a) and 4(b) show the comparison of the PRF and PMF histograms respectively. The effectiveness of our method can be seen from the comparison results. It can be seen from Fig.4 that the proposed method is slightly smaller than PSL in terms of Precision and PA, but higher than PSL in Recall. It is higher than all the comparison methods in F-Measure.

3.3. Ablation experiments

We next conduct an ablation experiment on the residential area test dataset, and the corresponding results are shown in Table I. We compared the initial saliency maps (ISM), pseudo pixel-level labels (PPL), and training with ISM (TISM). In addition, to verify the effectiveness of denoising, we use the dataset with 10% and 30% noise labels for training (Noisy10, Noisy30). In the table, we can see the effectiveness of our method at each stage.

Table I. PA, MIoU, and FWIoU values of ablation experiments.

Methods	PA	MIoU	FWIoU
ISM	0.8478	0.6318	0.8033
PPL	0.9027	0.6567	0.8643
TISM	09300	0.7248	0.8791
Noisy10	0.9169	0.6957	0.8567
Noisy30	0.9110	0.6800	0.8493
Ours	**0.9456**	**0.7389**	**0.9089**

4. CONCLUSION

In conclusion, the proposed semantic segmentation method for remote sensing images based on uncertainty perception has effectively addressed the issues of image-level annotation and label noise. The iterative label cleaning method improves the quality of the labeled data, and the two-stage strategy eliminates the need for post-processing steps during testing. The complementary uncertainty perception model enhances dataset feature utilization, resulting in more accurate semantic segmentation. Future improvements can be made by exploring more advanced algorithms to enhance the performance of the semantic segmentation model and by incorporating additional datasets to evaluate the generalizability of the method.

5. REFERENCES

[1] R. Yang, F. Pu, Z. Xu, C. Ding and X. Xu, "DA2Net: Distraction-Attention-Driven Adversarial Network for Robust Remote Sensing Image Scene Classification," IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2022.

[2] J. Zhou, R. Zhang, W. Zhao, S. Shen and N. Wang, "APS-Net: An Adaptive Point Set Network for Optical Remote-Sensing Object Detection," IEEE Geoscience and Remote Sensing Letters, vol. 20, pp. 1-5, 2023.

[3] A. Ma, J. Wang, Y. Zhong and Z. Zheng, "FactSeg: Foreground Activation-Driven Small Object Semantic Segmentation in Large-Scale Remote Sensing Imagery," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-16, 2022.

[4] C. Zheng, C. Hu, Y. Chen and J. Li, "A Self-Learning-Update CNN Model for Semantic Segmentation of Remote Sensing Images," IEEE Geoscience and Remote Sensing Letters, vol. 20, pp. 1-5, 2023.

[5] R. Wang, Y. Hao, L. Hu, J. Chen, M. Chen and D. Wu, "Self-Supervised Learning with Data-Efficient Supervised Fine-Tuning for Crowd Counting," IEEE Transactions on Multimedia, vol. 25, pp. 1538-1546, 2023.

[6] L. Chen, X. Dou, J. Peng, W. Li, B. Sun and H. Li, "EFCNet: Ensemble Full Convolutional Network for Semantic Segmentation of High-Resolution Remote Sensing Images," IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2022.

[7] J. Fan, Z. Zhang, C. Song et al., "Learning Integral Objects with Intra-Class Discriminator for Weakly-Supervised Semantic Segmentation," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4282-4291.

[8] Y. Liu, X. Kang, Y. Huang, et al., "Unsupervised Domain Adaptation Semantic Segmentation for Remote-Sensing Images via Covariance Attention," IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2022.

[9] B. Zhang, Y. Zhang, Y. Li et al., "Semi-supervised Deep Learning via Transformation Consistency Regularization for Remote Sensing Image Semantic Segmentation," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 5782-5796, 2023.

[10] Y. Wei and S. Ji, "Scribble-Based Weakly Supervised Deep Learning for Road Surface Extraction from Remote Sensing Images," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-12, 2022.

[11] Y. Hua, D. Marcos, L. Mou, X. Zhu and D. Tuia, "Semantic Segmentation of Remote Sensing Images with Sparse Annotations," IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2022.

[12] X. Zhang, W. Yu, X. Ma and X. Kang, "Weakly Supervised Local–Global Anchor Guidance Network for Landslide Extraction with Image-Level Annotations," IEEE Geoscience and Remote Sensing Letters, vol. 20, pp. 1-5, 2023.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization", IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618-626.

[14] P. T. Jiang, C. B. Zhang, Q. Hou, M. M. Cheng and Y. Wei, "LayerCAM: Exploring Hierarchical Class Activation Maps for Localization," IEEE Transactions on Image Processing, vol. 30, pp. 5875-5888, 2021.

[15] S. Desai and H. G. Ramaswamy, "Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization," 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 972-980.

[16] Z. Li, X. Zhang, P. Xiao and Z. Zheng, "On the Effectiveness of Weakly Supervised Semantic Segmentation for Building Extraction from High-Resolution Remote Sensing Imagery," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 3266-3281, 2021.

[17] Y. Liu and L. Zhang, "Weakly Supervised Region of Interest Extraction Based on Uncertainty-Aware Self-Refinement Learning for Remote Sensing Images," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-16, 2022.

[18] L. Zhang, J. Ma, X. Lv and D. Chen, "Hierarchical Weakly Supervised Learning for Residential Area Semantic Segmentation in Remote Sensing Images," IEEE Geoscience and Remote Sensing Letters, vol. 17, no. 1, pp. 117-121, 2020.

[19] A. Li, J. Zhang, Y. Lv, B. Liu, T. Zhang and Y. Dai, "Uncertainty-aware Joint Salient Object and Camouflaged Object Detection," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10066-10076.

[20] L. Zhang, K. Yang and H. Li, "Regions of Interest Detection in Panchromatic Remote Sensing Images Based on Multiscale Feature Fusion," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 7, no. 12, pp. 4704-4716, 2014.

[21] L. Zhang and A. Li, "Region-of-Interest Extraction Based on Saliency Analysis of Co-Occurrence Histogram in High Spatial Resolution Remote Sensing Images," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 8, no. 5, pp. 2111-2124, 2015.

[22] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, "Pyramid Scene Parsing Network," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6230-6239.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation, " Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234-241.

[24] L. Zhang and J. Ma, "Salient Object Detection Based on Progressively Supervised Learning for Remote Sensing Images," IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 11, pp. 9682-9696, 2021.