PHASE LEARNING BASED ON INTERACTIVE PERCEPTION FOR LIMITED-SAMPLE RESIDENTIAL AREA SEMANTIC SEGMENTATION

Xinran Lyu and Libao Zhang[*]

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

ABSTRACT

Due to the rich details of residential areas and the characteristics of remote sensing image sharpness vulnerable to haze, it will not only consume a lot of labor costs but also be very difficult to produce a large-scale dataset with strong labels. Therefore, the limited-sample dataset has become a hotspot in recent years. To address this issue, we proposed a semantic segmentation method for residential areas by phase learning. The main task of the first stage is to generate a joint saliency map by reducing the interference of haze noise through the feature comparison similarity sorting algorithm and combine them to generate initial pixel-level pseudo labels for the next stage of training. In the second stage, we proposed to construct a group feature interactive perception module to achieve image group semantic co-segmentation. Comprehensive evaluations with 2 datasets and the comparison with 7 methods validate the superiority of the proposed model.

Index Terms— Remote sensing, phase learning, weak annotation, interactive perception, contrast similarity ranking.

1. INTRODUCTION

With the rapid development of remote sensing technology, the huge quantity, necessary domain expert knowledge, and susceptibility to weather interference pose difficulties for accurate manual annotation of remote sensing images.

Therefore, although a large number of fully supervised semantic segmentation methods [1-3] have been proposed with good processing results, there is still a significant gap between them and practical applications due to their reliance on massive and accurately annotated remote sensing data.

The labels of the training dataset used in weakly supervised methods are often not accurate enough, such as image-level labels. Compared to pixel-level labels, such labels greatly save manpower and resources. How to obtain more reliable pseudo labels has become a key issue in current weakly supervised methods.

The mainstream method of achieving image-level labels to pixel-level segmentation relies on class activation maps (CAM). Wang et al. [4] proposed a farmland segmentation method based on weakly supervised learning, which introduces CAM to extract the intermediate layer features in the U-Net [5]. In addition, there are many improvement methods for CAM [6-8] widely used in weakly supervised semantic segmentation, such as gradient-weighted class activation mapping (Grad-CAM) [6], Layer-CAM [7], etc.

After obtaining pixel-level labels, weakly supervised methods often construct a model for error correction to obtain the final segmentation result [9-11]. Zhang et al. [12] proposed a hierarchical weakly supervised learning (HWSL) model to obtain residential areas by computing the gradient with intermediate convolutional layers. Wang et al. [13] proposed an iterative framework to refine the target region.

Although the deep learning model trained with a weakly labeled dataset effectively reduces the labor cost, however, the existing weakly supervised methods often face problems such as low computational accuracy, complex model structure, and low segmentation efficiency.

In addition, remote sensing images are also susceptible to weather conditions such as haze [14-16], resulting in blurry details, low contrast, and loss of important information in the image. For such problems, the general solution is to first remove fog from the image, and then perform interpretation work. The existing image dehaze methods can be roughly divided into three types: image enhancement-based methods [17], atmospheric scattering model-based methods [18], and deep learning-based methods [19-20]. For the presence of haze in the image, these methods have certain dehaze effects, but they will greatly increase the complexity of the model.

In this paper, we proposed a phase learning method based on interactive perception for limited-sample residential area semantic segmentation (PLIP). Wherein, limited-sample datasets refer to image-level annotated datasets containing a small portion of noisy images. The contributions of this article can be summarized as follows.

1) For foggy images in a limited-sample dataset, we proposed a feature comparison similarity sorting algorithm to reduce the influence of haze noise on pixel-level pseudo label generation.

This work was supported in part by the National Natural Science Foundation of China under Grant 41771407 and Grant 61571050, in part by the Beijing Natural Science Foundation under Grant 4222046. (*Corresponding author: Libao Zhang*.)

The authors are with the School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China. (e-mail: libaozhang@bnu.edu.cn)

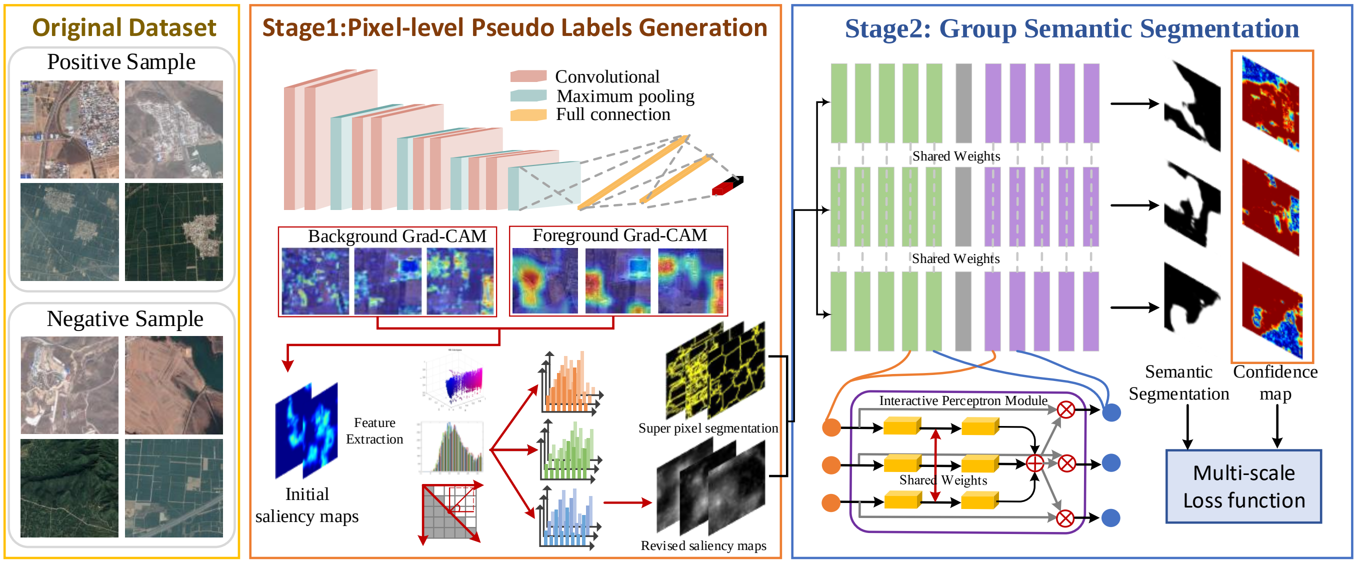


Fig.1. Framework of the PLIP model.

2) For inaccurate annotations in a limited-sample dataset, we proposed a new phase learning method to achieve semantic segmentation of residential areas without the need for post-processing steps during testing.

3) In response to the rich detailed information of remote sensing images and the common features of image groups, we designed similarity sorting and interactive perception modules in two phases.

2. METHODOLOGY

The overall framework of PLIP is shown in Fig. 1. The proposed model mainly contains two parts: pixel-level pseudo labels generation and final residential area semantic segmentation. In the first part, we used class feature awareness and proposed a feature similarity ranking method to generate and correct initial pixel-level pseudo labels. In the second part, we proposed to construct an interactive perceptron module in the structure of the codec to achieve image group semantic co-segmentation.

2.1. Pixel-level pseudo label generation based on class feature awareness and similarity ranking

At this stage, we proposed to generate pixel-level pseudo labels of positive samples required for the next stage of training, while minimizing fog noise interference by feature similarity ranking.

To achieve this goal, we will divide this stage into two modules. The first module is to generate initial pseudo labels using Grad-CAM, and the second module is to correct pseudo labels using similarity sorting.

Firstly, we trained a CNN based binary classification network. Then, the middle convolutional layer is used to calculate CAM, and the initial saliency map $S_{initial}$ is obtained after fusion.

The initial pixel-level pseudo labels can be obtained by directly binary segmentation of the initial saliency maps, as shown in the third row of Fig. 2. However, the initial pixel-level labels contain a large amount of background interference, which can hurt the subsequent semantic segmentation results. To eliminate background interference, further refinement is needed based on the commonality of low-level visual features.

For common feature similarity within image groups, we combine the initial saliency maps with the local feature similarity ranking list to achieve correction for the initial pseudo labels. In addition, we perform superpixel segmentation on the images.

Considering that hue is a representation of the reflection and radiation energy of ground objects in the image, as well as the rich texture information of remote sensing images, combined with the problem of reduced contrast in haze images, we proposed a multi-dimensional feature composed of enhanced hue contrast, spectral contrast, and texture information of superpixels in image I .

Then, we calculate the superpixel distance among images through formula (3) to sort the superpixel similarity.

$$D(P_k^{I_{M1}}, P_t^{I_{M2}}) = \sum_i \phi_i (\sqrt{(Feature_k^{I_{M1}} - Feature_k^{I_{M2}})^2}) \quad (1)$$

Wherein, $P_k^{I_{M1}}$ is the k th superpixel in image I^{M1} . $Feature_k^{I_{M1}}$ is the comprehensive feature vector of $P_k^{I_{M1}}$, and ϕ_i is the weight of different features.

The farther the superpixel feature distance is, the lower the similarity is. We rank from near to far according to the feature distance of superpixels to get the similarity ranking list D^I of the image group.

The weight is calculated according to the feature distance between superpixels. We only select the superpixel value of the top M in the similarity ranking list for calculation.

$$w_m = \frac{1}{M}\sum_{m=1\cdots M}\exp\left(-D_m^I/(\sum_{m=1\cdots M}D_m^I)\right)^{-1} \qquad (2)$$

Finally, the superpixel saliency values with high similarity are fused to obtain the final superpixel joint saliency value $S(P_k^I)$.

$$S_{revise}(P_k^I) = \frac{1}{M}\sum_{m=1\cdots M} w_m \cdot S_{initial}(P_m) \qquad (3)$$

Wherein, S_{revise} is the revised saliency map after haze suppression, which will be used as the pixel-level pseudo label in the second step. The revised saliency maps are shown in the fourth row of Fig. 2. From this figure, it can be seen that the revised saliency maps can eliminate significant background interference, as shown in Fig.2 (a), (b), and (e). In addition, for images with fog (c) and (f), the revised saliency map retains more details to reduce fog interference.

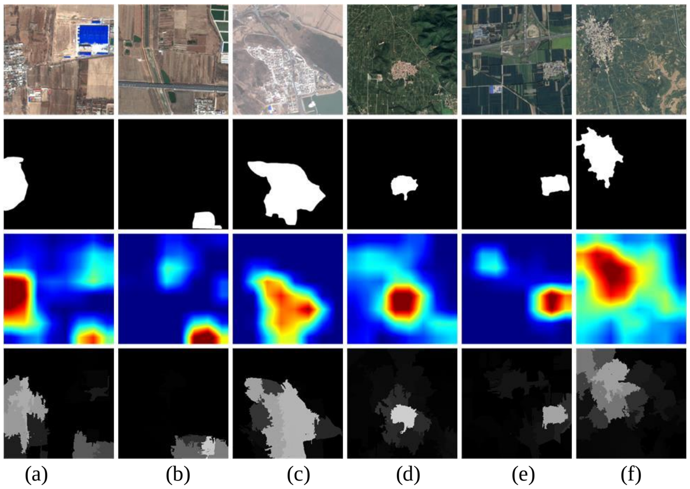


Fig.2. Initial saliency maps and confidence maps on two datasets. From top to bottom: the original remote sensing image, the ground-truth maps, the initial saliency maps, and the revised saliency maps.

2.2. Semantic co-segmentation based on construct feature interactive perception for image group

In this section, we received inspiration from Qin et al. [21], and constructed a multi-branch network that consists of several codecs. Each codec contains multiple residual modules. The residual module contains three parts: local features extraction, symmetric encoder-decoder, and multi-scale feature fusion. That is, first, the convolutional layer is used to transform the input feature map to achieve local feature extraction. Then, the extracted local feature maps are input into the symmetric encoder-decoder structure to learn multi-scale contextual information. The extraction of multi-scale information reduces the loss of detail caused by large-scale direct upsampling. Finally, integrate local features and multi-scale features.

Setting K encoders in a symmetric encoder-decoder structure can obtain K feature maps of different scales. Then, we mapped them to the size of the input image and fused them to get the result y. After building a multiple branches network, we proposed a feature interaction perception module to help us enhance the semantic information.

Due to the need to identify common semantic information among multiple inputs, attention weights should be the same in feature interaction perception.

Assuming that the network has N branches. We respectively extract the feature maps $f_1(x)$, $f_2(x)$ … $f_N(x)$ of the encoder output in the multi-branch networks. After the global average pooling, they are transferred to the full connection layer to obtain weight vectors Att_1, Att_2 … Att_N.

$$Att_i = \sigma(W^T * AvgPool(\sum_i f_i(x)) + b), i = 1, 2, \cdots, N \qquad (4)$$

The final feature map $f_i^{'}(x)$ is obtained by channel-level multiplication of Att_j and $f_i(x)$.

$$f_i^{'}(x) = Att_j \odot f_i(x), i, j \in [1, 2, \cdots, N], i \neq j \qquad (5)$$

In addition, because there are some problems in the modified pixel-level labels, such as background interference and targets are not detected, we introduce the confidence map as the pixel weight to help calculate the cross-entropy loss function. For S_{revise}, the pixel with a value approaching 1 or 0 has a high confidence level, and approaching 0.5 has a low level. We calculate the confidence map C as follows:

$$C = 2*\left|S_{revised} - 0.5\right| \qquad (6)$$

Finally, we introduced a weighted binary cross-entropy loss function to train the network. $S_{revised}$ is used as a pseudo label GT_i after binarization.

$$Loss^k = \sum_{U\in\mathbf{U}}\sum_{i=1}^{|U|} C_i^k \cdot (GT_i^k \cdot \log(P(\mathrm{y}_i^k = 1)) + (1 - GT_i^k)\cdot\log(P(\mathrm{y}_i^k = 0))), \quad k = 1\cdots K \qquad (7)$$

3. EXPERIMENTS

We evaluate these methods on the GeoEye-1 dataset and Google Earth dataset. For the GeoEye-1 dataset, we randomly selected 30% of the images for hazing processing to simulate cloud and fog interference in remote sensing images. For Google Earth datasets, due to the inherent cloud and fog interference in their screenshots, they can be considered partially foggy datasets.

We compare our proposed method with 7 state-of-the-art methods. MFF [22], and SACH [23] are traditional unsupervised methods, designed to detect the salient object for remote sensing images. Grad-CAM [6] and Layer-CAM [7] are weakly supervised methods with image-level annotations. FCN [3] and U-Net [5] are two fully supervised semantic segmentation networks. PSL [24] is proposed for residential area semantic segmentation with weak annotation, which is based on hierarchical learning. The residential area segmentation results are compared by visual comparison and quantitative analysis.

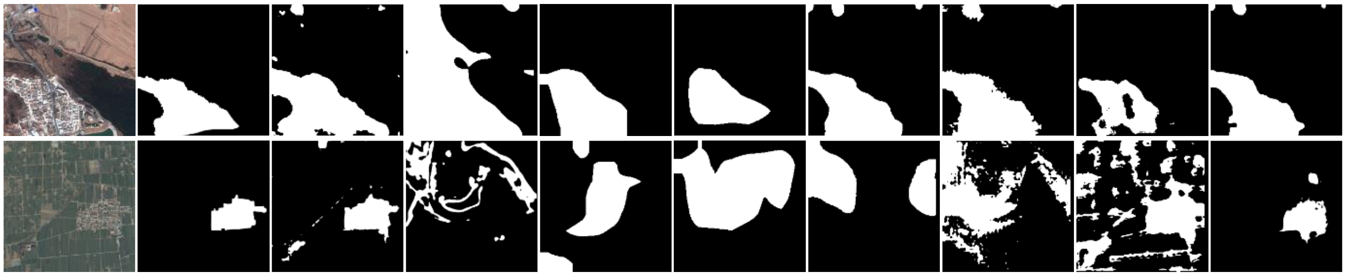
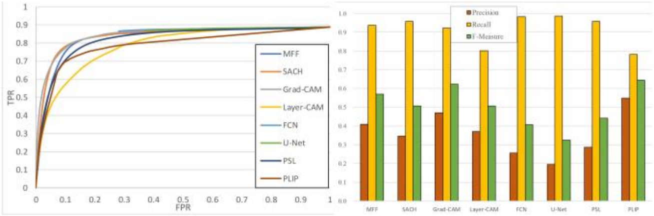

(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig.3. Residential area semantic segmentation on GeoEye-1 dataset. (a) Original images, (b) Ground-Truth, (c) MFF, (d) SACH, (e) Grad-CAM, (f) Layer-CAM, (g) FCN, (h) U-Net, (i) PSL, (j) PLIP.

3.1. Visual comparison

Fig. 3 shows the residential areas semantic segmentation results of different methods respectively. From Fig. 3, we can see that most methods have the problem of false detection. In comparison, our method can use the correlation between images, to adjust the saliency value of the region, to achieve a suppression effect.

In addition, MFF and SACH algorithms are unsupervised traditional methods. They are specially designed for the saliency analysis of remote sensing images, whose results have holes and fragments in the salient area. The detection results of the SACH method for the first row of images contain a large amount of background. The Grad-CAM and Layer-CAM algorithms based on weakly supervised learning have unclear boundaries and incomplete object extraction, although the saliency value in the detection area is relatively uniform. For the second and the fourth line, no target image, the four methods all have relatively serious error detection. Relatively speaking, the results of FCN, U-Net, and PSL are more accurate and have more precise boundaries. However, FCN and U-Net contain a lot of background interference, and more target areas are lost in PSL results. Compared to others, our method has obtained more accurate results.



(a) (b)

Fig.4. ROC curves and Precision, Recall, and F-Measure values of the saliency maps and residential area segmentation results of two datasets.

3.2. Quantitative analysis

We evaluate the 8 methods in terms of a Receiver Operator Characteristic (ROC) curve, Precision, Recall, and F-Measure. For a given saliency map, a binary map is obtained by segmenting it with varying quantization thresholds and then compared with the ground truth to compute the false positive rate (FPR), and true positive rate (TPR) for an image. FPR and TPR are then depicted in the ROC curve. For the semantic segmentation result, we use the Precision, Recall, and F-Measure. 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 ROC curves and PRF histograms respectively.

3.3. Ablation experiments

We next conduct an ablation experiment on the test dataset, and the corresponding results on Pixel Accuracy (PA), Mean Intersection over Union (MIoU), and Frequency Weighted Intersection over Union (FWIoU) to evaluate the accuracy are shown in Table I. The data in this table is the average of the results of two datasets. We compared the initial pixel-level labels, revised saliency maps, one-way model, multi-branch model training with initial saliency maps (MBMIS), and the entire model PLIP.

In the table, we can see that the MBMIS model verified the results of suppressing haze noise without using similarity ranking. The effectiveness of each step of our model was verified through ablation experiments.

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

Methods	PA	MIoU	FWIoU
Initial saliency map	0.8597	0.6441	0.8125
Revised saliency map	0.8979	0.6785	0.8440
One-way model	0.8544	0.6287	0.8049
MBMIS	0.8030	0.5551	0.7410
PLIP	0.8980	0.6657	0.8551

4. CONCLUSION

In this paper, a new residential area semantic segmentation method based on an image-level annotation dataset is proposed. First, we constructed a classification network and proposed an initial pixel-level label calculation method based on class feature awareness. At the same time, to reduce the interference of haze noise, we propose the use of the contrast feature similarity ranking method. Then, to achieve image group semantic co-segmentation and maintain the edge information, we proposed to construct a feature cross perception module. In the future, we intend to propose a new weak annotation method for multiple classes of semantic segmentation in remote sensing images to further reduce the dependence on human labor.

5. REFERENCES

[1] X. Pan, J. Zhao and J. Xu, "A Scene Images Diversity Improvement Generative Adversarial Network for Remote Sensing Image Scene Classification," IEEE Geoscience and Remote Sensing Letters, vol. 17, no. 10, pp. 1692-1696, 2020.

[2] B. Bai, W. Fu, T. Lu and S. Li, "Edge-Guided Recurrent Convolutional Neural Network for Multitemporal Remote Sensing Image Building Change Detection," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-13, 2022.

[3] E. Shelhamer, J. Long and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 640-651, 2017.

[4] S. Wang, W. Chen, S.M. Xie, G. Azzari, and D.B. Lobell, "Weakly Supervised Deep Learning for Segmentation of Remote Sensing Imagery," Remote Sensing, vol. 12, no. 2, pp. 207, 2020.

[5] 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.

[6] 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.

[7] 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.

[8] 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

[9] 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.

[10] 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.

[11] X. Lyu and L. Zhang, "Progressive Refinement Learning Based on Feature Cross Perception for Residential Areas Semantic Segmentation," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1-5.

[12] 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.

[13] X. Wang, S. You, X. Li and H. Ma, "Weakly-Supervised Semantic Segmentation by Iteratively Mining Common Object Features," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1354-1362.

[14] Y. Zheng, J. Su, S. Zhang, M. Tao and L. Wang, "Dehaze-AGGAN: Unpaired Remote Sensing Image Dehazing Using Enhanced Attention-Guide Generative Adversarial Networks," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-13, 2022.

[15] J. Nie, W. Wei, L. Zhang, J. Yuan, Z. Wang and H. Li, "Contrastive Haze-Aware Learning for Dynamic Remote Sensing Image Dehazing," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-11, 2022.

[16] C. Li, H. Yu, S. Zhou et al., "Efficient Dehazing Method for Outdoor and Remote Sensing Images," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 4516-4528, 2023.

[17] J. Wang, K. Lu, J. Xue, N. He and L. Shao, "Single Image Dehazing Based on the Physical Model and MSRCR Algorithm," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 9, pp. 2190-2199, 2018.

[18] M. Zhu, B. He and Q. Wu, "Single Image Dehazing Based on Dark Channel Prior and Energy Minimization," in IEEE Signal Processing Letters, vol. 25, no. 2, pp. 174-178, 2018.

[19] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "MAXIM: Multi-Axis MLP for Image Processing," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5759-5770.

[20] B. H. Park, S. Chattopadhyay and J. Burgin, "Haze Mitigation in High-Resolution Satellite Imagery Using Enhanced Style-Transfer Neural Network and Normalization Across Multiple GPUs," IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2021, pp. 2827-2830.

[21] X. Qin, Z. Zhang, C. Huang, et al., "U2-Net: Going Deeper with Nested U-Structure for Salient Object Detection," Pattern Recognition, vol. 106, pp. 107404, 2020.

[22] 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.

[23] 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.

[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.