ISBN: 978-981-18-7950-0
2023 the 13th International Workshop on Computer Science and Engineering (WCSE 2023)
doi: 10.18178/wcse.2023.06.019

Non-local Dense RepPoints for Instance Segmentation in Remote Sensing

Xinmin Xiang [1+]

[1] College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China

Abstract. Recent advances in deep convolutional neural networks have significantly improved instance segmentation. In large-scale remote sensing images, however, the high density, arbitrary shapes and orientation, large aspect ratios, and huge scale variation of the objects pose significant challenges to general instance segmentation algorithms. In this paper, we propose a new framework, called non-local dense RepPoints, to improve the performance of instance segmentation in remote sensing images. First, we propose a hierarchical non-local block that iteratively integrates global information, and our method enables the model to accurately represent the relationship between two locations. Second, we enhance Dense RepPoints by designing an efficient dynamic vector to more efficiently model the objects by a large number of adaptive points. We conduct experiments on the iSAID dataset and compare our method with several commonly-used state-of-the-art networks. The experimental results demonstrate that our proposed approach can achieve promising results.

Keywords: Remote Sensing, Instance Segmentation, Non-local, Dense Reppoints

1. Introduction

Due to the rapid development of optical remote sensing technology, modern satellites are able to acquire remote sensing images with very high resolution (VHR). Instance segmentation aims to recognize the category labels of individual objects and localize them using pixel-level masks. It plays an essential role in many applications in the field of remote sensing, such as urban management [1] [2] [3], land planning [4] [5], and land cover classification [6] [7] [8]. However, instance segmentation in remote sensing scenes is more challenging than natural scenes due to the following characteristics of remote sensing scenes:

1) Objects in geospatial space tend to appear randomly distributed in the image, with arbitrary shapes and orientations, as well as occlusions. Consequently, non-local information interaction is crucial in remote sensing scenarios.

2) The background ratio in remote sensing scenes is extremely high, and the background is more complex than natural scenes. It causes a severe drop in accuracy due to excessive intra-class variance.

3) Due to the high density and large-scale variation of the object, it brings a multi-scale challenge to existing methods.

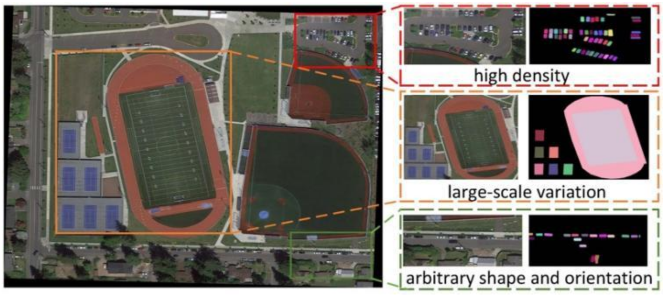

Fig. 1: Three challenging aspects of instance segmentation in remote sensing.

As a result of the development of deep convolutional neural networks [9] [10] [11], many deep learning-based approaches have been proposed to improve the accuracy and speed of instance segmentation. Some powerful network structures, such as Mask R-CNN [12], SOLO [13], and Dense RepPoints [14], have been

[+] Corresponding author. Tel.: + 86 18356409925
E-mail address: 2101090209@hhu.edu.cn

widely used in natural scene. However, instance segmentation in remote sensing is still challenging in three aspects (as shown in Fig. 1): (1) Objects occur in a high density and an assortment of shapes and orientation. (2) The background in the VHR remote sensing image is much more complex. (3) The scale of remote sensing images is much larger, and the object exhibits large-scale variation.

In the recently proposed Dense RepPoints method, objects are represented by a large number of adaptive points, which effectively improves the detection of small objects and simplifies the formulation of instance segmentation. However, it requires a large number of points (i.e., 729 points) to represent the object, which is evidently inappropriate for remote sensing images. In addition, the method fails to characterize non-local relationships, resulting in insufficient interactions between locations.

In this work, we improve Dense RepPoints from two aspects: non-local relations and adaptive representative points. First, we design a hierarchical non-local block to model global context, which can enhance the model's ability to capture long-range dependencies and model the relationships of related objects over long distances iteratively. Second, we propose an effective dynamic vector for adaptive point regulation. By adding an auxiliary offset, we perform a finer correction of representative points, thereby enabling our model to accurately predict objects with fewer points. Non-local Dense RepPoints achieves not only state-of-the-art performance, but also superior adaptability and stability in the extensive experiments on the iSAID dataset. Our goal is to develop a framework for instance segmentation in remote sensing that is stronger, faster, and more robust.

The main contributions of our work are as follows:

(1) We propose Non-local Dense RepPoints to improve the efficacy of instance segmentation in remote sensing. It promotes semantic interaction in various locations and adds an auxiliary offset to improves prediction accuracy for dense objects, making remote sensing more applicable.

(2) We design a hierarchical non-local block to capture pixel-wise global information and derive more robust relationships between locations in a more efficient and effective manner.

(3) We propose an efficient dynamic vector to regulate the adaptive points with finer corrections for representative points, thereby improving the prediction quality and enabling our model to accurately predict objects by fewer points.

(4) We obtain the state-of-the-art performance on iSAID datasets. Extensive experiments demonstrate that our model has the superior adaptability and stability.

2. Related Work

2.1. Framework of Instance Segmentation

Instance segmentation is challenging because it requires both instance-level and pixel-level predictions. The existing methods can be categorized into two groups. Two-stage instance segmentation usually expresses this task as a detect-and-then-segment paradigm. Typically, they first detect the bounding boxes and then segment within the region of each bounding box. Most of the two-stage work builds on Faster R-CNN [15], such as Mask R-CNN, by adding an additional mask branch and employing RoI-Align rather than RoI-Pooling to improve performance. PANet [16] introduces bottom-up path augmentation, adaptive feature pooling, and fully connected fusion to improve the accuracy. In summary, the aforementioned frameworks consist of two steps: detecting and then segmenting the object in the box. They can attain cutting-edge performance but are frequently slow.

One-stage methods seek to dealing with instance segmentation directly without requiring box detection or embedding learning as a prerequisite. Deep Watershed Transform predicts the energy map for every pixel and groups them using the watershed algorithm. Polarmask formulates the problem as the prediction of an instance's profile based on classification of the instance's center and dense distance regression in polar coordinates. Dense RepPoints utilizes a large set of points to describe an instance. In these methods, each pixel generates auxiliary information, which is then used by a clustering algorithm to group object instances based on their information.

2.2. Non-local Context Modeling

There has been a significant amount of work aimed at incorporating contextual information into existing deep learning models. One type of strategies tends to aggregate context information by stacking many local operations, such as the convolution operator [17] or the recurrent operator [18]. These models are usually defined on features with local neighborhoods or specific computational functions, which are limited in modeling non-local context relationships by stacking local receptive fields and short-range context. Pixel-level interaction based on attention mechanisms is also a common strategy. Non-local neural networks [19] utilize the self-attention that enables a single feature from any location to perceive features from all other locations. DANet [20] proposes spatially-wise and channel-wise attention modules to enhance feature representation. However, the redundant computations in these methods lead to prohibitive noise, which hinders its application, particularly in some intensive prediction tasks.

2.3. Deformable Convolution

In traditional convolutional layers, the kernel is fixed and independent of the input, i.e., the weights are identical for each image and each location of the image. To bring more flexibility to traditional convolution, deformable convolution [21] is a powerful and efficient mechanism that can handle sparse spatial locations. It learns the sampling locations dynamically by predicting the offsets for each image location. RepPoints [22] is inspired by it, and Dense RepPoints brings the dynamic mechanism into instance segmentation. We enhance it further to make it more appropriate for remote sensing images.

2.4. Remote Sensing Image Understanding

Remote sensing image understanding is one of the hottest topics in the computer vision community. Remote sensing scenes are more complex because they consist of more multi-scale objects and a larger proportion of small objects in the imagery. There have been plenty of recent works to solve the aforementioned problems. For geospatial instance/object detection, Wang et al. [23] proposes single-shot detection for multi-scale objects, and Chen et al. [24] introduces a pipeline of hybrid supervision for geospatial instance segmentation. Fully-weighted HGNN [25] captures both short- and long-range dependencies in spatial features by hypergraph. In large-scale instance segmentation, HMANet [26] captures global context from the perspectives of space, channel, and category. Sun et al. [27] concentrates on improving geospatial object segmentation in complex scenes using a semi-supervised method.

3. Methodology

We present the details of the proposed Non-local Dense RepPoints in this section.

3.1. A Revisit to Dense RepPoints

Vanilla RepPoints represents an object with a few representative points ($n = 9$). A few points are sufficient for object detection because the category and bounding box of an object can be fit with few points. Nevertheless, instance segmentation provides annotations at the pixel level for objects requiring precise estimation for fine-grained geometric localization. Consequently, a larger volume of point sets and a vector of attributes associated with each representative point are required to describe the object:

$$\boldsymbol{\mathcal{R}} = \{(\boldsymbol{x_i} + \Delta\boldsymbol{x_i}\,,\ \boldsymbol{y_i} + \Delta\boldsymbol{y_i}\,, \boldsymbol{a_i})\}_{\boldsymbol{i=1}}^{\boldsymbol{n}}, \tag{1}$$

where a_i is the attribute vector associated with the i-th point. x_i and y_i denote an initialized location. Δx_i and Δy_i are learnable offsets, and n is the number of points.

The feature of each point $\mathcal{F}(p)$ is extracted from the feature map $\mathcal{F}$ by bilinear interpolation, and the feature of a point set $\mathcal{F}(p)$ can be defined as the concatenation of all adaptive representative points of $\mathcal{R}$:

$$\boldsymbol{\mathcal{F}}(\boldsymbol{\mathcal{R}}) = \mathbf{concat}\big(\boldsymbol{\mathcal{F}}(\boldsymbol{p_1}), \ldots, \boldsymbol{\mathcal{F}}(\boldsymbol{p_n})\big), \tag{2}$$

which is employed to recognize the category of the point set. The object segment of a point set can be obtained by a conversion function.

In instance segmentation, the attribute map can be defined as the foreground score of a point set. Therefore, Dense RepPoints proposes a point-level classification loss L_{cls}^{p} and a point-level localization loss L_{loc}^{p}, in addition to the box-level classification and localization terms L^b_{cls} and L^b_{loc}. The objective function can be expressed as:

$$\boldsymbol{L}_{\mathbf{det}} = \boldsymbol{L}^{\boldsymbol{b}}_{\mathbf{cls}} + \boldsymbol{L}^{\boldsymbol{b}}_{\mathbf{loc}}, \tag{3}$$

$$\boldsymbol{L}_{\mathbf{mask}} = \boldsymbol{L}^{\boldsymbol{p}}_{\mathbf{cls}} + \boldsymbol{L}^{\boldsymbol{p}}_{\mathbf{loc}}, \tag{4}$$

$$\boldsymbol{L} = \boldsymbol{L}_{\mathbf{det}} + \boldsymbol{L}_{\mathbf{mask}}, \tag{5}$$

where $L^p_{\mathbf{cls}}$ is responsible for predicting the representative point foreground score, and $L^p_{\mathbf{loc}}$ is for learning point localization.

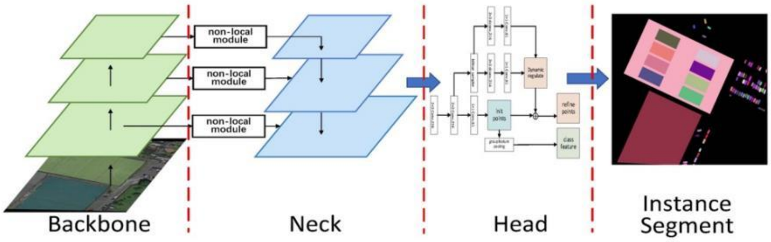

Fig. 2: Architecture of our Non-local Dense RepPoints.

3.2. Non-local Dense RepPoints

As shown in Fig. 2, we propose the Non-local Dense RepPoints architecture for more precise instance segmentation tasks in remote sensing scenes. The model is divided into three parts: Backbone for extracting image feature maps, Non-local module + Feature pyramid networks (FPN) [28] for non-local feature interactions at multiple scales, and Head for recognition and classification.

In this paper, we adopt ResNet as Backbone to extract the feature maps of the images. Since our main innovations are Hierarchical Non-local Block and Dynamic regulate for Dense RepPoints, in the following, we will mainly introduce these methods.

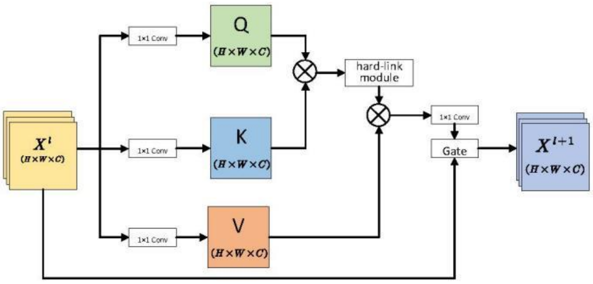

Fig. 3: Illustration of an non-local block.

3.3. Hierarchical Non-local Block

Traditional global feature interactions often adopt global feature interactions based on self-attention, and such methods have obtained excellent model performance and effects in natural scenes. However, due to the complex and high background ratio in remote sensing scenes, the traditional global feature interaction often introduces additional background noise in the interaction process. To tackle this problem, we propose the Hierarchical Non-local Block, which improves the model's focus on the effective information by filtering the noise of the feature vector after the interaction by the hard-link module. Meanwhile, the Hierarchical Non-local Block contains a trainable gating mechanism that can filter the background noise twice by adjusting the weights of the original input and global interaction features, as shown in Fig. 3.

We design a hard-link module that uses a threshold θ to filter correlated noise that is too small and normalizes it by the softmax function. The hard-link module is calculated as follows:

$$\boldsymbol{\mathcal{W}} = \mathbf{softmax}(\mathrm{W}) \begin{cases} \mathbf{W}_{\boldsymbol{i},\boldsymbol{j}} = \mathbf{W}_{\boldsymbol{i},\boldsymbol{j}}, \mathbf{W}_{\mathbf{i},\mathbf{j}} > \boldsymbol{\theta}, \\ \mathbf{W}_{\boldsymbol{i},\boldsymbol{j}} = \mathbf{0}, \mathbf{W}_{\boldsymbol{i},\boldsymbol{j}} < \boldsymbol{\theta}, \end{cases} \tag{6}$$

where W is calculated by matrix multiplication of Q and K.

Given that the original feature map X_{src} and the feature map X_g after global interaction will carry different noise information, we propose a gating mechanism to suppress noise interference in order to reduce noise interference. The formula of the gating mechanism is as follows:

$$\gamma = \sigma\left(W_\gamma X_{\mathrm{src}} + U_\gamma X_g + b_\gamma\right), \tag{7}$$

$$X = \gamma \odot X_{\mathrm{src}} + (1 - \gamma) \odot X_g, \tag{8}$$

where W_γ and U_γ are learnable weight matrices. b_γ is the bias, and γ is sigmoid function. X represents features after interaction.

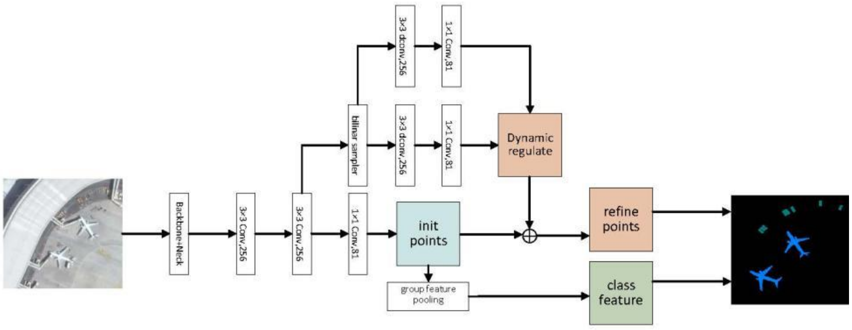

Fig. 4: Illustration of dynamic regulation method.

3.4. Dynamic Regulate for Dense RepPoints

Dense RepPoints adopt the method of initializing representative points from the center point and then refining them to generate object contours. This approach achieves excellent results in natural scenes, however, in remote sensing images, the model cannot locate the object contour directly and effectively due to the objects with arbitrary shapes and orientations, as well as occlusions. In addition, since Dense RepPoints only uses a single refinement, it is necessary to initialize a large number of representative points, and the method suffers from an inordinate amount of redundancy due to the high density of objects in the remote sensing scene.

To solve the above problem, we propose the dynamic regulation method, which improves the precision of model prediction of object contours by secondary regulation of feature nodes. The structure is shown in Fig. 4. Specifically, we add an additional dynamic convolutional layer to learn the Dynamic regulation weight in order to optimize the refind offset. By employing additional Dynamic regulation, the secondary correction of refind offset is achieved, allowing the method to be more applicable to remote sensing scenes and to extract the target contour more accurately. This method also reduces the number of initialization points, thus reducing the computational effort and making it applicable to high-density scenes.

4. Experiments

4.1. Dataset and Metrics

iSAID [29] contains 2,806 aerial images with high-quality annotations, which is the largest dataset for instance segmentation in aerial images. It contains 655,451 instance annotations for 15 categories varying greatly in scale, orientation, and aspect ratio. The spatial resolution of images ranges from 800 to 13,000 in width. The training set includes 1411 aerial images, the validation set contains 458 aerial images and the test set includes 937 aerial images. The biggest challenge with the iSAID dataset is the extreme foreground-background imbalance problem, where the foreground ratio is significantly smaller than the background ratio. In addition, the background is much more complex, which leads to significant false positives due to large intra-class differences. As the official evaluation server of iSAID is still being improved and annotations for the test set are not available, this paper evaluates the method of this paper on the validation set.

We use the standard COCO metrics [30]: AP (averaged over IoU threshold), AP_{50}, AP_{75}, AP_{S}, AP_{M}, AP_{L}, respectively.

4.2. Training Details

We used PyTorch to implement our model. All experiments were conducted on 4 Nvidia GeForce RTX 3090 Ti GPUs. The model was trained with stochastic gradient descent (SGD) optimizer. If not specifically specified, the model was trained for 24 rounds (i.e., two training cycles). In addition, the initial learning rate of the model is 0.006, which decays to 10% of the original at the 16th and 20th rounds, respectively. The weight decay and momentum parameters were set to 10^{-4} and 0.9, respectively. All ablation studies were conducted on the validation set using ResNet-50 and FPN as the underlying backbone network.

Table 1: Instance segmentation mask AP (%) on iSAID val dataset. ∗ indicates training with scale augmentation.

Method	Backbone	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage:								
Mask R-CNN	ResNet-50-FPN	24	34.2	57.9	38.1	21.7	39.4	46.9
Mask Scoring R-CNN	ResNet-50-FPN	24	36.4	58.2	38.7	25.1	41.4	48.6
One-stage:	ResNet-50-FPN	24	28.7	49.9	27.4	14.9	32.4	40.1
YOLACT	ResNet-50-FPN	24	27.7	48.6	25.6	14.5	34.9	38.8
PolarMask	ResNet-50-FPN	24	32.5	53.2	32.4	18.1	36.8	42.6
Dense RepPoints	ResNet-50-FPN	24	33.9	55.2	33.8	19.6	37.5	44.3
Non-local Dense RepPoints	ResNet-50-FPN	24	36.8[**+2.9**]	58.3	39.6	25.5	42.3	48.9
Non-local Dense RepPoints	ResNet-101-FPN	48	37.9[**+4.0**]	59.2	42.6	26.1	43.5	50.3
Non-local Dense RepPoints*	ResNet-101-FPN	48	38.5[**+4.6**]	59.9	42.8	27.2	44.7	50.1

Table 2: Ablation studies of number of points on iSAID validation dataset.

Number of Points	AP
9	34.9
25	35.7
81	36.8
121	36.5

4.3. Quantitative Analysis

To demonstrate the effectiveness of our Non-local Dense RepPoints in instance segmentation, we compare our network with several strong baseline methods, including both two-stage methods (Mask R-CNN, Mask Scoring R-CNN) and one-stage methods (YOLACT, PolarMask, SOLOv2, and Dense RepPoints). Table 1 shows the results on the validation set of the iSAID dataset, where “*” denotes the training results enhanced by multi-scale. We can see Non-local Dense RepPoints achieves the best results for instance segmentation, outperforming other two-stage models. With the ResNet-101 pre-trained model, Non-local Dense RepPoints can achieve 37.9 mAP after 48 rounds of training, outperforming all other methods; by increasing the training scale of the model via multi-scale augmentation, the method can achieve 38.5 mAP, yielding the best performance among all models. It demonstrates that our proposed model can accurately locate and detect objects in aerial images.

Table 3: Ablation studies of non-local module in different feature maps on iSAID validation dataset.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Dense RepPoints:						
FPN	33.9	55.2	33.8	19.6	37.5	44.3
+2	35.0	57.4	37.5	22.9	39.1	46.6
+3	34.8	56.8	**38.1**	22.5	38.8	47.3
+4	35.2	57.1	37.9	23.6	39.5	47.1
+234 (non-local module)	**35.5**	**57.8**	37.9	**24.6**	**39.8**	**47.6**

We attribute the improvement to the following aspects: (1) Our model can capture short- and long-range dependencies to iteratively model relations between related objects; (2) Dynamic regulation for Dense RepPoints can effectively adjust the position information of the predicted points to more accurately distribute them on the target contour, thus improving the prediction performance of the object position.

Analysis on Variant Models. We conduct ablation experiments by varying the number of points, and the results are shown in Table 2. We find that the model's performance is gradually improved as the number of points rises, however, when the number of points reaches a certain threshold (tested as 81 in our experiments), the performance stops increasing. We further compare our method with Dense RepPoints, and we find that our methods using 81 points can even outperform Dense RepPoints with 729 points.

We also conduct ablation studies by applying the non-local module on feature maps at different scales, and the results are shown in Table 3. We find that the inclusion of the module in all scale leads to performance improvements, with the module being the most significant at the highest level, where a 1.3 mAP improvement can be achieved. This is due to the fact that this integration strategy integrates global information from the upper layer to all subsequent layers. In addition, we apply the non-local module to multiple stages, augmenting the feature graph with multi-scale contextual information for more efficient global information interaction.

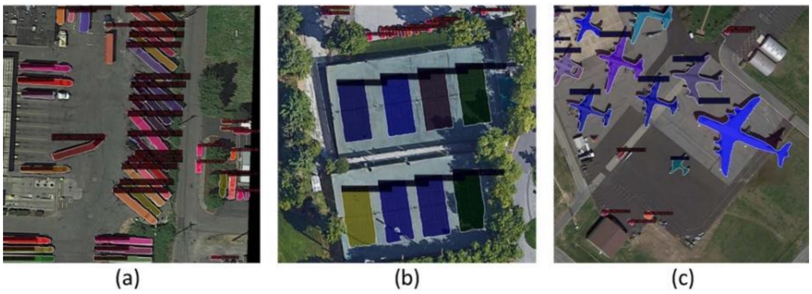

Fig. 5: Three challenging scenarios.

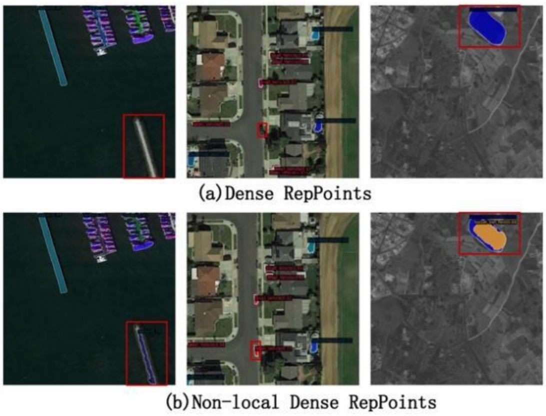

Fig. 6: Comparison between our Non-local Dense RepPoints and the classical method.

4.4. Qualitative Analysis

Fig. 5 visualizes three classic and challenging scenarios: (a) arbitrary shapes and orientations; (b) a complex background; and (c) large-scale transformations. Our method shows excellent results in these complex scenes. The comprehensive learning and interaction of global spatial features and the secondary optimization of the target's contour points allow the model to effectively identify and decode targets in scenes with high density, targets with occlusion or noise, and large-scale transformations. Notable also is the fact that the method still produces satisfactory results for objects subject to large-scale transformation conditions.

Fig. 6 shows the visualization of our method and other classical methods in the same scene. By comparison, we find that Non-local Dense RepPoints achieves the best visualization results, in terms of object recognition rate and object contour.

5. Conclusions and Future Work

In this paper, we explored the instance segmentation in remote sensing. Based on the Dense RepPoints model, we proposed Non-local Dense RepPoints that are more applicable to remote sensing scenes. It enhances the Dense RepPoints in two aspects: non-local relations and adaptive representative points. First, we designed a hierarchical non-local block to model the global context, which improves the ability of the model to capture short- and long-range dependencies to iteratively model relations of related objects. Second, we proposed an effective dynamic validation method to regulate the adaptive points. By including a supplementary offset, we apply a finer correction to the representative points, which also allows our model to accurately predict objects with fewer points.

We conducted extensive experiments on iSAID dataset. As expected, the experimental results validate the efficiency of Non-local Dense RepPoints with the state-of-the-art performance and the superior adaptability and stability. We believe that our model will significantly inspire scholars in remote sensing and encourage more research on instance segmentation to apply to real-life situations. In the future, we hope to address the omissions in current work and expand our model to more remote sensing image interpretation tasks, such as rotating object detection and large-scale semantic segmentation.

References

[1] S. Azri, U. Ujang, F. A. Castro, A. A. Rahman, and D. Mioc. Classified and clustered data constellation: An efficient approach of 3d urban data management. ISPRS Journal of Photogrammetry and Remote Sensing, 113:30–42, 2016.

[2] S. Lumnitz, T. Devisscher, J. R. Mayaud, V. Radic, N. C. Coops, and V. C. Griess. Mapping trees along urban street networks with deep learning and street-level imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 175:144 – 157, 2021.

[3] W. LU, P. WANG, R. NIU, H. YU, K. FU, X. SUN, Y. TIAN. From single- to multi-modal remote sensing imagery interpretation: a survey and taxonomy. SCIENCE CHINA Information Sciences,66(4):140301 – , 2023. 1

[4] R. Cong, Y. Zhang, L. Fang, J. Li, Y. Zhao, and S. Kwong. Rrnet: Relational reasoning net-work with parallel multiscale attention for salient object detection in optical remote sensing images. IEEE T GEOSCI REMOTE, 60:1 – 11, 2021.

[5] J. He, X. Li, P. Liu, X. Wu, J. Zhang, D. Zhang, X. Liu, and Y. Yao. Accurate estimation of the proportion of mixed land use at the street-block level by integrating high spatial resolution images and geospatial big data. IEEE T GEOSCI REMOTE, 59(8):6357 – 6370, 2020

[6] T. Hermosilla, M. A. Wulder, J. C. White, and N. C. Coops. Land cover classification in an eraof big and open data: Optimizing localized implementation and training data selection to improve mapping outcomes. Remote Sensing of Environment, 268:112780, 2022.

[7] X. Li, C. Sun, H. Meng, X. Ma, G. Huang, and X. Xu. A novel efficient method for landcover classification in fragmented agricultural landscapes us-ing sentinel satellite imagery. Remote Sensing, 14(9):2045,2022.

[8] Y. Li, Y. Zhou, Y. Zhang, L. Zhong, J. Wang, and J. Chen. Dkdfn: Domain knowledge-guided deep collaborative fusion network for multimodal unitemporal remote sensing land cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 186:170 – 189, 2022.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. CVPR, pages 770 – 778, 2016.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural net-works. Communications of the ACM, 60(6):84 – 90, 2017.

[11] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. CVPR, pages 2961 – 2969, 2017.

[13] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li. Solo: Segmenting objects by locations. ECCV. Springer, 2020.

[14] Z. Yang, Y. Xu, H. Xue, Z. Zhang, R. Urtasun, L. Wang, S. Lin, and H. Hu. Dense rep-points: Representing visual objects with dense point sets. ECCV, pages 227 – 244. Springer, 2020. 1.

[15] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. NeurIPS, 2015.

[16] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path aggregation network for instance segmentation. CVPR, pages 8759 – 8768, 2018.

[17] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwrit-ten zip code recognition. Neural computation,1(4):541 – 551,1989.

[18] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8):1735 – 1780, 1997.

[19] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7794 – 7803, 2018. 2.

[20] H. Xue, C. Liu, F. Wan, J. Jiao, X. Ji, and Q. Ye. Danet: Divergent activation for weakly supervised object localization. ICCV, pages 6589 – 6598, 2019. 2.

[21] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks.CVPR, pages 764 – 773, 2017.

[22] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin. Reppoints: Point set representation for object detection. ICCV, pages 9657 – 9666, 2019. 2.

[23] P. Wang, X. Sun, W. Diao, and K. Fu. Fmssd: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing,58(5):3377 – 3390, 2019. 2.

[24] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan. Vehicle detection in satellite images by hybrid deep convolutional neural networks. IEEE Geoscience and remote sensing letters, 11(10):1797–1801, 2014.

[25] Y. Tian, X. Sun, R. Niu, H. Yu, Z. Zhu, P. Wang, and K. Fu. Fully-weighted HGNN: Learning efficient non-local relations with hypergraph in aerial imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 191:263 – 276, 2022. 2.

[26] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu. Hybrid multiple attention network for semantic segmentation in aerial images. IEEE T GEOSCI REMOTE: 1-18.

[27] X. Sun, A. Shi, H. Huang, and H. Mayer. Basˆ{4} net: Boundary-aware semi-supervised semantic segmentation network for very high resolution remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13:5398 – 5413, 2020.

[28] T.-Yi Lin, P. Dollɑ́r, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. CVPR, pages 2117 – 2125, 2017.

[29] S. W. Zamir, A. Arora, A. Gupta, S. Khan, G. Sun, F. S. Khan, F. Zhu, L. Shao, G. Xia, and X. Bai. ISAID: A large-scale dataset for instance segmentation in aerial images. ICCV Workshops, pages 28 – 37, 2019. 4.

[30] T.-Yi Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollɑ́r, and C. L. Zitnick. Microsoft coco: Common objects in context. ECCV, pages 740 – 755. Springer, 2014.