remote sensing

MDPI

Communication

Gaussian Dynamic Convolution for Semantic Segmentation in Remote Sensing Images

Mingzhe Feng [1], Xin Sun [1,2,*], Junyu Dong [1] and Haoran Zhao [3]

[1] College of Information Science and Engineering, Haide College and Institute of Advanced Ocean Study, Ocean University of China, Qingdao 266100, China
[2] The Department of Aerospace and Geodesy, Technical University of Munich, 80333 Munich, Germany
[3] School of Information and Control Engineering, Qingdao University of Technology, Qingdao 266520, China
* Correspondence: sunxin1984@ieee.org

Abstract: Different scales of the objects pose a great challenge for the segmentation of remote sensing images of special scenes. This paper focuses on the problem of large-scale variations of the target objects via a dynamical receptive field of the deep network. We construct a Gaussian dynamic convolution network by introducing a dynamic convolution layer to enhance remote sensing image understanding. Moreover, we propose a new Gaussian pyramid pooling (GPP) for multi-scale object segmentation. The proposed network can expand the size of the receptive field and improve its efficiency in aggregating contextual information. Experiments verify that our method outperforms the popular semantic segmentation methods on large remote sensing image datasets, including iSAID and LoveDA. Moreover, we conduct experiments to demonstrate that the Gaussian dynamic convolution works more effectively on remote sensing images than other convolutional layers.

Keywords: remote sensing; semantic segmentation; deep learning; convolutional neural networks

Citation: Feng, M.; Sun, X.; Dong, J.; Zhao, H. Gaussian Dynamic Convolution for Semantic Segmentation in Remote Sensing Images. *Remote Sens.* **2022**, *14*, 5736. https://doi.org/10.3390/rs14225736

Academic Editor: Claudio Piciarelli

Received: 25 October 2022
Accepted: 9 November 2022
Published: 13 November 2022

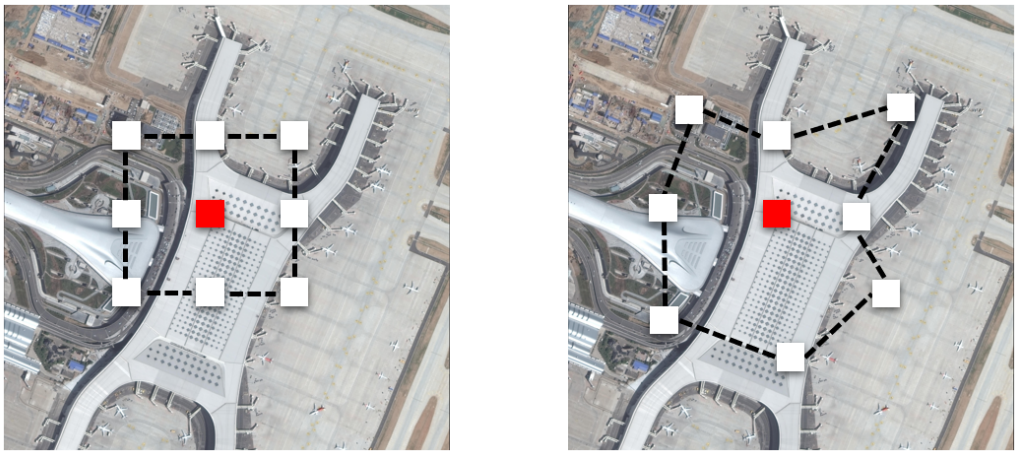
1. Introduction

With aerospace remote sensing technology, high-resolution images of remote sensing are popularly used in urban management, weather prediction, agricultural development, and environmental protection. The processing and analysis of remote sensing images with deep network models have become a popular research topic nowadays. As a special semantic segmentation task on remote sensing images, it aims to detect the specific location and predict the class of objects of interest from high spatial resolution images. A unique semantic label will be assigned to each pixel in the region where these objects are located.

Remote sensing images are often taken from an overhead perspective via the observation platform. Therefore, the images are generally characterized by large-scale variations, which makes the detection and location of targets more difficult. Moreover, the same category of objects may have irregular shapes, such as airports. This makes the task of segmenting the target objects in high-resolution remote sensing images more challenging than that in natural scenes.

In recent years, semantic segmentation methods that work well in the natural scenes have been directly applied to the remote sensing images [1], such as U-Net [2], PSP-Net [3], Deeplabv3+ [4], SDFCNv2 [5]. Unfortunately, the performances are often difficult to be satisfied. We argue that the objects on the high-resolution remote sensing images are variable in scale which makes them more difficult to be recognized than in natural scenes. Researchers have realized that the variant convolution with large receptive fields can alleviate the problems of large-scale variations and irregular shapes of the objects on the remote sensing images, such as dilated convolution [6] and deformable convolution [7]. However, the dilated convolution is limited by the dilation factors and can only collect regular-shaped features, while the target object features in remote sensing images are often irregular in shape. To handle this problem, this paper will pay attention

to the dynamical receptive field of the deep network. We introduce a Gaussian dynamic convolution (GDConv) kernel which dynamically increases its receptive field by sampling different offsets via Gaussian distributions. It can effectively enhance the ability to extract features at different scales on remote sensing images. Furthermore, we construct a Gaussian dynamic convolutional network (GDCN) for remote sensing image segmentation, which makes the receptive field for extracting contextual information more flexible and richer in the extracted features. As shown in Figure 1, we use an example of the airport to illustrate the difference between GDConv and generally dilated convolution. The dilated convolution used in Figure 1a increases the receptive field; however, it only provides features at a fixed scale. In contrast, our GDConv can randomly select the center of the convolution kernel. It gives an additional offset to the convolution weight vector in each direction to expand the receptive field, which extracts features of different scales and shapes, as shown in Figure 1b.



(**a**) (**b**)

Figure 1. (**a**) The fixed receptive field of the dilated convolution; (**b**) The dynamic receptive field of the Gaussian dynamic convolution.

The main contributions of this paper are as follows:

(1) We introduce the GDConv layer to the field of semantic segmentation for high-resolution remote sensing images. It can dynamically adjust the size of the receptive field to make the extracted multi-scale features rich and vivid.
(2) We construct a Gaussian pyramid pooling (GPP) module and a Gaussian dynamic convolutional network (GDCN) to obtain high accuracy of the multi-scale object segmentation.
(3) The experiments show that our method obtains convincing results on both remote sensing datasets iSAID [8] and LoveDA [9], which indicates that the GDConv works well on the problems of large-scale changes and difficult shapes of the objects on the remote sensing images.

The rest of this work will be organized as follows. In Section 2, related works are reviewed. In Section 3, we propose the Gaussian Dynamic Convolution Network. Furthermore, we present a detailed comparison in Section 4. Finally, we conclude this paper in Section 5.

2. Related Work

2.1. General Semantic Segmentation

Semantic segmentation, as one fundamental task in the area of computer vision and image processing, has a wide range of application prospects, and it can be said that accurate recognition cannot be produced without proper segmentation. Traditional semantic

segmentation methods mainly extract low-level features of images for segmentation by hand-made feature descriptors, which are very tedious and too dependent on the a priori knowledge of relevant experts, and often only two types of semantic segmentation can be achieved, and multiple operations have to be performed if the object being segmented has more than one target, which brings very great inconvenience to the segmentation task at that time. This situation did not end until the advent of the machine learning era.

With the prosperity of deep convolutional neural networks (DCNN) [10], semantic segmentation has entered the era of deep learning. In 2015, Long et al. proposed an enabling end-to-end and pixel-to-pixel approach FCN [11], which first induced deep learning to semantic segmentation. Many subsequent approaches have adopted the "encoder–decoder" structure, such as U-Net [2], SegNet [12], RefineNet [13], and PSPNet [3]. The HRnet [14] reduced the semantic information lost during upsampling and downsampling by employing branches of different resolutions as parallel structures and maintaining the information interaction between different branches. It allows the whole structure to maintain high resolution from the beginning to the end. Sun et al. [15] proposed Gaussian dynamic convolution, which can dynamically select sampling regions based on the offset of the Gaussian distribution, and they further constructed a lightweight network to handle the complex single-image segmentation task. Lv et al. [16] introduced the embedded attention module to generate feature bases and continuously update them using contextual information to ensure accuracy while greatly reducing the computation and improving segmentation efficiency.

Although these methods mentioned above are widely used in the segmentation task of natural scenes and have achieved remarkable results, they still have the problems of insufficient quality and accuracy in high-resolution remote sensing images. This is mainly because the objects on the remote sensing images show their large-scale variation compared with natural scenes. Moreover, the objects are densely arranged and the boundaries are blurred. Therefore, special segmentation methods in high-resolution remote sensing images are required to handle the problem of the large variety of objects.

2.2. Semantic Segmentation in Remote Sensing

In the field of remote sensing, semantic segmentation is also widely used [17,18], including urban planning [19], vehicle detection [20], climate prediction [21] and ocean vortex tracking [22]. For different remote sensing scenes and their characteristics such as large changes in object scale, dense arrangement of objects, and too small target objects, researchers also modified the generic semantic segmentation network accordingly to better fit the characteristics of the remote sensing images. For example, SSCDnet [23] uses semi-supervised learning of domain adaptive features to solve the data distribution drift problem of remote sensing images. Mou et al. [24] added spatial relations and channel information to convolutional networks to improve the competitiveness of the network. FarSeg [25] correlates target objects with background features through geospatial scenes and enhances the recognition of foreground objects with contextual information, which greatly reduces the false positives due to excessive similarity between target objects and background. Factseg [26] enhances the feature recognition of small targets by designing foreground activation drives. The framework consists of a double branch, partly used to suppress the large-scale background and partly used to activate the features of small objects. Some recent methods [5,27,28] try to incorporate modules that are effective in the field of general segmentation into remote sensing image segmentation networks, such as the well-known transformer or attention mechanisms, which are effective in improving the accuracy of the networks to some extent. However, these methods mainly target special application scenarios and are not effective in solving problems of semantic segmentation for high-resolution remote sensing images, such as multi-scale variation of objects and loss of foreground details.

3. Method

3.1. Gaussian Dynamic Convolution

The Gaussian dynamic convolution was first proposed to efficiently deal with the segmentation problem for the single image [15]. This work will introduce the Gaussian dynamic convolution to construct convolutional networks to handle the problem of large scales of objects in high-resolution remote sensing images. Here we first give an example to illustrate the principle of GDConv implementation.

For an ordinary 3×3 convolutional kernel, a feature map is obtained by sliding on the input layer. Assuming that the coordinates of the center of the convolutional kernel are $d = (x, y)$, the coordinates of the positions in the remaining eight directions can be represented by d plus the set of orientations s and the offset Δ:

$$\begin{gathered} \bar{d}_i = d + \Delta_i \odot s_i \\ s_i \in \{\langle -1, -1\rangle, < -1, 0 >, \cdots, < 1, 1 >\}, \Delta =< 1, 1 > \end{gathered} \tag{1}$$

For example, the feature vector in the lower left corner can be expressed as:

$$\langle x, y\rangle + \langle -1, -1\rangle \odot \langle 1, 1\rangle = \langle x - 1, y - 1\rangle. \tag{2}$$

We illustrate a toy example of GDConv in Figure 2. We first fix the center weights and then give a random offset Δ to each direction, the offset is obtained from a two-dimensional Gaussian distribution with standard deviation $\sum$. The Gaussian distribution equation is:

$$\text{Gaussian}(0, \Sigma) = \frac{\sqrt{2}}{\Sigma\sqrt{\pi}} \exp\left(-\frac{x^2}{2\Sigma^2}\right). \tag{3}$$

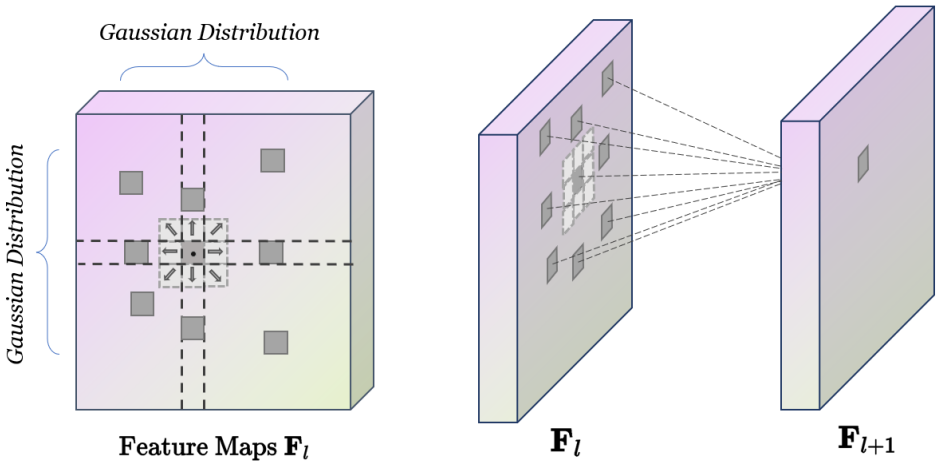


Figure 2. Illustration of the Gaussian dynamic convolution with kernel size 3.

We can see that the dynamic receptive field breaks the shape limitation. Different from the deformable convolution of which the offset needs to be learned from the previous feature map by an additional convolution layer, the GDConv is able to change the offset dynamically.

3.2. Gaussian Dynamic Convolution Network

To further illustrate the advantages of GDConv in the field of image segmentation for remote sensing, we construct a novel Gaussian Dynamic Convolution Network (GDCN) as shown in Figure 3. The input image is first passed through a deep convolutional neural network for feature extraction to obtain both the low-level and high-level features. Then, the high-level features are subjected to our Gaussian pyramid pooling (GPP) module for mining the multi-scale contextual information content. The GPP module is shown in

Figure 4. It is worth mentioning that it is quite different from the Gaussian dynamic pyramid pooling [15]. The feature pyramid module mentioned in [15] is composed of multiple dilated convolutions with different dilation factors and Gaussian dynamic convolutions with fixed offsets, which does not break the shape limitation. Our GPP module is composed of Gaussian dynamic convolution and dilated convolution. The Gaussian dynamic convolution ensures features of different scales and natural shapes of objects. Therefore, our GDCN can better extract the multi-scale features in high-resolution remote sensing images.

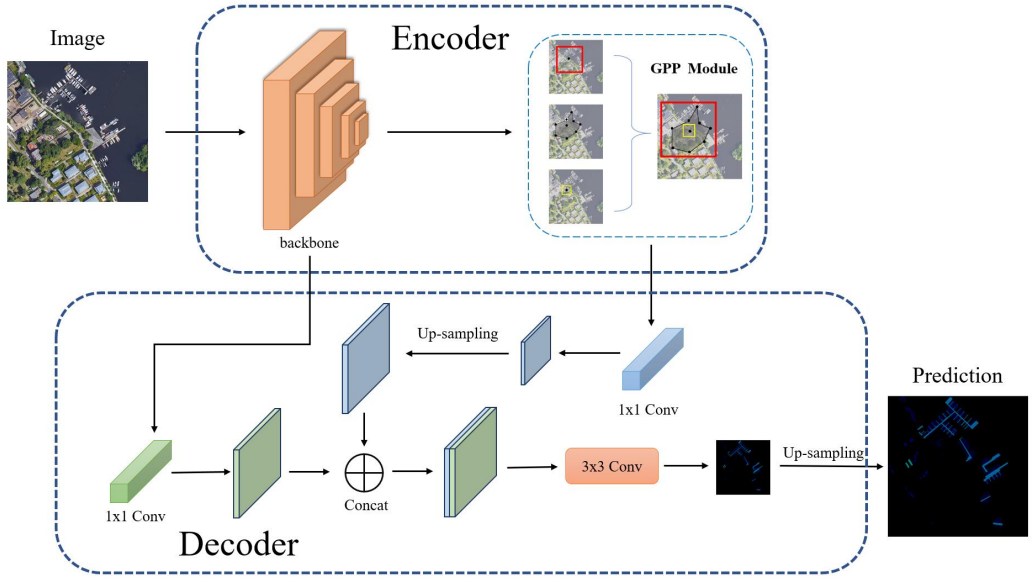


Figure 3. Illustration of the Gaussian Dynamic Convolutional Network.

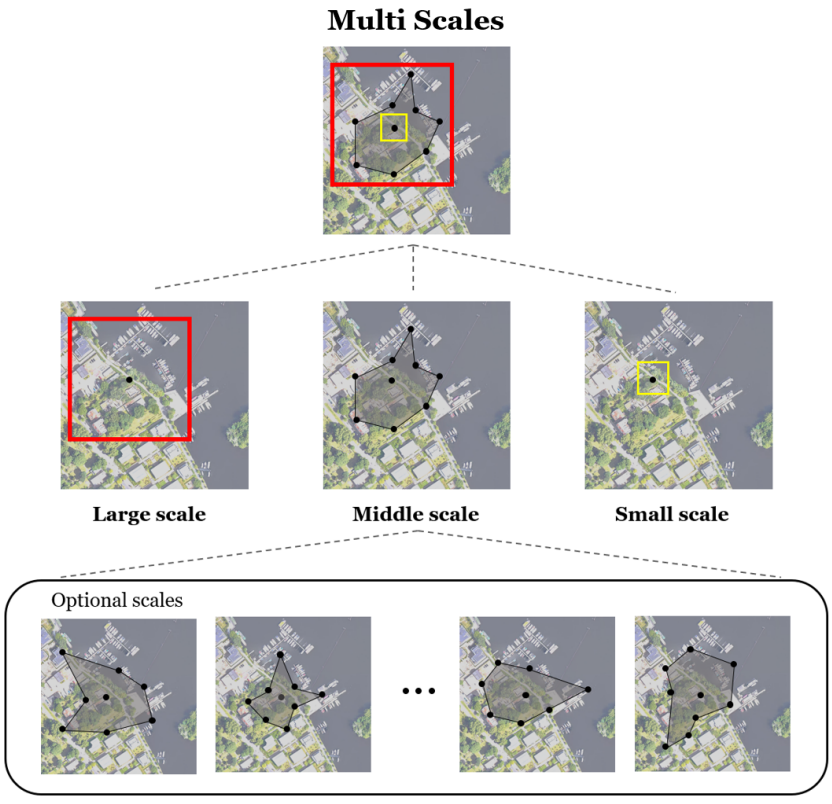


Figure 4. Illustration of the Gaussian Pyramid Pooling.

Many segmentation tasks use dilated convolution with different dilation rates to fuse information at different scales. As shown in Figure 4, the large red box represents an atrous convolution with a large dilation factor used to collect large-scale information in the image, while the small yellow square indicates a small dilation rate used to collect small scale information in the image. In order to fuse the mesoscale information, people often use a layer of dilated convolution with a moderate dilation rate or multiple layers of atrous convolution with increasing dilation factor in order, such as Atrous Spatial Pyramid Pooling (ASPP) module. We do a different job. We use GDConv to capture the information at different scales. GDConv can break the limitation of the expansion factor by randomly generating Gaussian offsets to extract information at various scales instead of limiting it to extracting intermediate scales. This is the reason why our GPP module can produce richer feature maps.

We concatenate the feature layers of different scales obtained by the GPP module, and then adjust the number of channels via an 1×1 convolution layer to obtain the blue feature layer as shown in Figure 3. This feature layer is first subjected to $4\times$ upsampling and fused with the low-level feature layer having the same spatial resolution. We downsample the low-level features with another 1×1 convolutional layer before fusion, because low-level features often contain a large number of channels. The fused feature layers are then passed through several 3×3 convolutional layers for feature extraction and finally subjected to $4\times$ upsampling to restore the feature layer to the same size as the original image, which gives the final prediction result.

4. Experiments

In this section, we will conduct extensive experiments to verify the effectiveness of the proposed method on iSAID and LoveDA datasets. We implement our method with Pytorch on NVIDIA 2080Ti GPUs. The code is available at https://github.com/ouc-ocean-group/, accessed on 24 October 2022.

4.1. Dataset and Metric

The iSAID is a large dataset for the high-resolution remote sensing image segmentation task. Its original images are derived from the large-scale aerial dataset DOTA, which is captured by aircraft, UAVs, and satellites from different platforms. The dataset contains 2806 images and 655,451 object instances in 15 categories. The categories are defined as: storage tank (ST), baseball diamond (BD), tennis court (TC), basketball court (BC), ground field track (GTF), large vehicle (LV), small vehicle (SV), helicopter (HC), swimming pool (SP), roundabout (RA), soccer ball field (SBF). Since the work in this paper involves only semantic segmentation, only the semantic mask annotations of this dataset are used.

The LoveDA was released by the team from Wuhan University and contains 5987 high-resolution images of remote sensing from the cities of Nanjing, Changzhou, and Wuhan. The dataset is divided into two parts: urban and rural, and contains seven categories.

We adopt mean intersection over union (mIoU), commonly used in semantic segmentation, as the main evaluation metric to demonstrate the performance of our method.

4.2. Experimental Setting

Implementation Details To demonstrate the performance of GDCN better, we take ResNet-50 as the backbone and pretrain it on ImageNet. The feature layers generated by the backbone will be input into the GPP module, which consists of two dilated convolutional layers and two Gaussian dynamic convolutional layers. The dilation factors of the two dilated convolutions are, respectively, set to 1 and 18 and the base offsets of the two Gaussian dynamic convolutions are set to 6 and 12, with $\sum$ set to 2. We use a "poly" learning rate strategy for iteration, where the initial learning rate is 0.007, and the learning rate is calculated by $1 - (\frac{iter}{max_iter})^{power}$ (power = 0.9) for each subsequent step. Our weight decay is set to 0.0001 and the momentum is set to 0.9.

We crop the images to a fixed size of 800 × 800 for the iSAID dataset, and for the Love DA dataset, we randomly crop the images to 512 × 512. Moreover, we implement some common data enhancement strategies during training. For instance, we randomly scale the images in the range of $[0.5, 2.0]$. Furthermore, the random horizontal flip operation is performed to process the input image. For the batch size, we set it to 8 on the iSAID dataset and 16 on the LoveDA dataset. We choose the SGD optimizer to complete the network optimization.

4.3. Performance

We compared our method with several successful semantic segmentation methods, including DeepLabv3+ [4], PSPNet [3], DCN [7], FarSeg [25], and FactSeg [26]. The results are shown in Tables 1 and 2. Among them, Deeplab v3+ employs dilated convolution to build the Atrous Spatial Pyramid Pooling module. Dilated convolution is also employed by PSPNet. The problem is that it does not take full advantage of the dilated convolution to expand the receptive field range, and it does not perform well in dealing with multi-scale problems. The Deformable Convolutional Network (DCN) uses deformable convolution for feature extraction, but the disadvantage is that it introduces too much useless contextual information. Moreover, the methods of FarSeg and FactSeg are currently the most effective methods. Therefore, we choose these four methods as our comparison methods.

Table 1. The results (mIoU) of object segmentation on the iSAID val set. The best results are illustrated in bold in each column.

Method	mIoU(%)	Ship	ST	BD	TC	BC	GTF	Bridge
Deeplab v3+ [4]	59.33	59.02	55.15	75.94	84.18	58.52	59.24	32.11
PSPNet [3]	60.25	65.2	52.1	75.7	85.57	61.12	60.15	32.46
DCN [7]	60.12	61.24	54.69	72.88	82.96	55.32	55.46	34.58
FarSeg [25]	63.71	65.38	61.80	77.73	86.35	62.08	56.70	36.70
FactSeg [26]	63.79	68.34	56.83	78.36	**88.91**	64.89	54.60	36.34
ours	**64.22**	**69.33**	**62.2**	**78.48**	85.59	**65.26**	**60.28**	**37.23**
Method	**LV**	**SV**	**HC**	**SP**	**RA**	**SBF**	**Plane**	**Harbor**
Deeplabv3+ [4]	54.54	33.79	31.14	44.24	67.51	73.78	75.70	45.76
PSPNet [3]	58.03	42.96	40.89	46.78	68.6	71.9	79.5	54.26
DCN [7]	56.25	39.25	33.36	47.77	69.81	70.33	76.21	49.85
FarSeg [25]	60.59	46.34	35.82	51.21	71.35	72.53	82.03	53.91
FactSeg [26]	**62.65**	49.53	42.72	51.47	69.42	**73.55**	84.13	55.74
ours	61.67	**49.85**	**42.85**	**52.01**	**69.65**	73.25	**84.29**	**56.71**

Table 2. Smantic segmentation results achieved on the test set of LoveDA.

Method	mIoU(%)
DeepLab v3+ [4]	47.62
PSPNet [3]	48.31
DCN [7]	47.92
FarSeg [25]	48.69
FactSeg [26]	48.94
ours	**49.75**

On the iSAID dataset, our method achieves the mIoU of 64.22, which is a 4.89 improvement over DeepLabv3+ and a 3.97 improvement over PSPNet. This is because Gaussian dynamic convolution breaks the limitation of the dilation factors and can further collect irregularly shaped features. Therefore, it achieves better performance on remote sensing datasets. In classes with a large number of irregular instances, such as small vehicle (SV), helicopter (HC), and Plane, our method is significantly better than deeplabv3+ and PSPNet.

Moreover, our results work well compared with the recent methods, as seen in Table 1. The results improve 0.43 mIoU over the well-known FarSeg and 0.51 mIoU over FactSeg.

Among the 15 categories provided by the iSAID dataset, we have the best results in 12 categories such as BC, ST, GTF, HC, Ship, Bridge, BD, SV, SP, RA, Plane, and Harbor. Only on three categories, TC, LV, and SBF, our results are not as good as the previous method. In the category of SBF, our method is almost the same as FactSeg with only 0.3 lower. Further analysis of the results shows that our results are 1 higher than the best method for the category with a lot of irregular boundaries such as harbor, which shows that our method has good results for handling irregular shapes in remote sensing images. In addition, for categories containing a large number of small and densely distributed objects, such as small vehicle (SV) and ship, our results are 0.3 and 1 higher than the best results. This indicates that our method can provide a better solution for the problem of densely arranged target objects in remote sensing images. However, there are also some categories with degraded performance because of the small number of objects, such as tennis courts (TC) and large vehicles (LV).

On the Love DA dataset, with the backbone of ResNet-50, our method achieves the mIoU of 49.75, which is an improvement over deeplab v3+ by 2.13 and 1.06 over FarSeg. The above multiple sets of experimental data can prove that our algorithm is very effective in dealing with the multi-scale problem on the remote sensing images. The reason is that the proposed GDConv is optimized accordingly for the characteristics of remote sensing images and performs better than the general segmentation methods. The visualization results on the iSAID dataset are shown in Figure 5. We can see that on regular round and square objects, the results of our method are sharper than other methods, where the edges are much smoother. On finer objects, our method is even sharper.

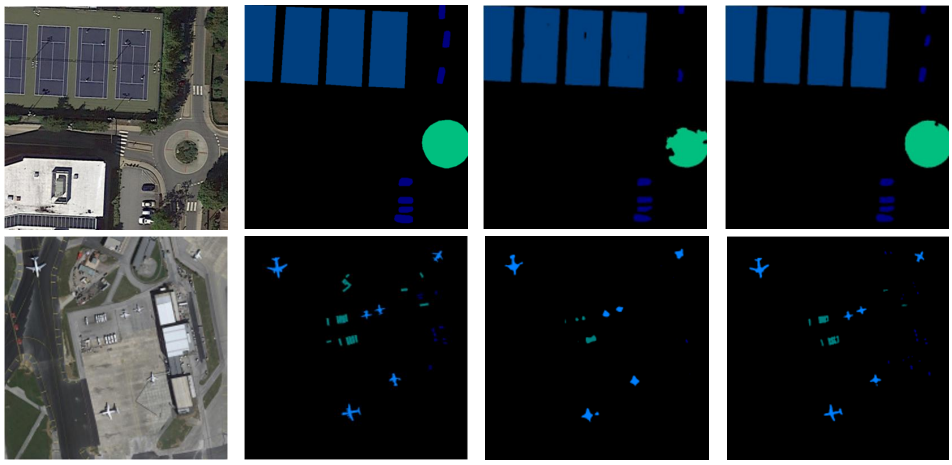


Figure 5. Visualization results on the iSAID val set.

To better illustrate our approach, we select two typical examples, as shown in Figure 6. In the first row, we can see that the size of the target plane is small and has an obviously different color between the fuselage and wings. The two wings of the plane are completely disappeared by Deeplab v3+, and only the fuselage part is detected. Thus, it misclassifies the object into the small vehicle class. Meanwhile, our method can completely label the whole object, so that the target can be correctly identified as the plane class. This is because the Gaussian dynamic convolution breaks the limitation of the dilation factors and can randomly produce arbitrary irregularly shaped receptive fields. It has the natural advantage of distinguishing remote sensing image target objects and effectively alleviates the false detection caused by the small size and dense arrangement of remote sensing image target objects. This example fully illustrates the superiority of our method to reduce the probability of false detection and improve the accuracy of segmentation. In the second row, there is a wild space next to the target area, and it is very similar to the soccer ball field class, which makes it misclassify the area into the soccer ball field class by both Deeplab

v3+ and our method. It shows that our method also has some shortcomings to be improved. In the next step, we will continue to optimize our algorithm, for example, deepening the potential of Gaussian dynamic convolution to explore better possibilities in the field of remote sensing image segmentation.

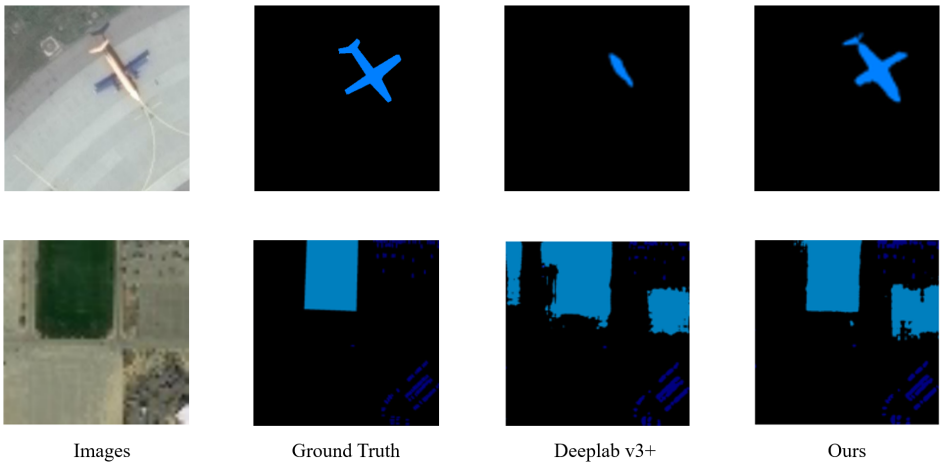


Figure 6. More visualization results on the iSAID val set.

4.4. Ablation Study

We explore the impact of different types and numbers of convolutional layers in the pyramid pooling module on the segmentation results in Table 3. We first set the number of intermediate convolutional layers to 1, and compare the results obtained using dilated convolution (with a dilation factor of 6), deformable convolution, and Gaussian dynamic convolution (to ensure fairness, the base offset of GDConv is also set to 6 and $\sum$ to 2). Our GDConv module is going to improve 1.3 mIoU over the dilated convolution and 4.1 mIoU over the deformable convolution. Then, we set the number of intermediate convolution layers to 2, where the dilation factors of the two dilated convolutions are set to 6 and 12 (reverting to ASPP modules). To ensure fairness, the base offsets of GDConv are also set to 6 and 12, and $\sum$ remains 2. The results show that our GPP structure is 2.2 mIoU better than the original ASPP module and 5.2 mIoU better than the two deformable convolutions, which fully demonstrates the superiority of Gaussian dynamic convolution in handling remote sensing image segmentation tasks.

Table 3. Segmentation performance between GDConv and other convolutions on iSAID val set.

Methods	mIoU(%)
1 × Dilated Conv (dilation = 6)	58.4
1 × Deformable Conv	56.6
1 × GDConv (base = 6, $\sum$ = 2)	60.7
2 × Dilated Conv (dilation = 6, 12)	61.3
2 × Deformable Conv	59.3
2 × GDConv (base = 6, 12, $\sum$ = 2)	64.2

In Table 4, a set of experiments was performed in order to investigate how the standard deviation $\sum$ would affect the final segmentation results. The value of the standard deviation $\sum$ indicates the size of the range that affects the receptive field. From Tables 3 and 4, it can be seen that Our GDConv produces better results than the two compared convolutions for different values of $\sum$. In particular, the best results are obtained when $\sum$ is taken as 2.0. We can see that either large or small $\sum$ causes performance degradation. The reason is that the large $\sum$ may fuse unnecessary noise information, and the small $\sum$ leads to insufficient sampling area to fully acquire the desired features.

Table 4. Segmentation performance between different standard deviation $\sum$ on iSAID val set.

$\sum$	1.0	1.5	2.0	2.5
mIoU (%)	63.5	64.4	64.9	64.2

5. Conclusions

The current popular semantic segmentation methods are insufficient to deal with the problem of large-scale variations of target objects in high-resolution remote sensing images. To alleviate this phenomenon, we introduce the Gaussian dynamic convolution to the semantic segmentation task on remote sensing images. This method can dynamically increase its receptive field. Furthermore, we construct GDCN based on Gaussian dynamic convolution to enhance remote sensing image understanding. The experimental results show that the GDCN improves the performance on remote sensing image segmentation tasks more than most of the previous common methods.

Funding: This work was supported in part by National Natural Science Foundation of China under Project Nos. 61971388 and Alexander von Humboldt Foundation.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, M.; Hu, X.; Zhao, L.; Lv, Y.; Luo, M.; Pang, S. Learning dual multi-scale manifold ranking for semantic segmentation of high-resolution images. *Remote Sens.* **2017**, *9*, 500. [CrossRef]
2. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
3. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
4. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
5. Chen, G.; Tan, X.; Guo, B.; Zhu, K.; Liao, P.; Wang, T.; Wang, Q.; Zhang, X. SDFCNv2: An Improved FCN Framework for Remote Sensing Images Semantic Segmentation. *Remote Sens.* **2021**, *13*, 4902. [CrossRef]
6. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
7. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
8. Waqas Zamir, S.; Arora, A.; Gupta, A.; Khan, S.; Sun, G.; Shahbaz Khan, F.; Zhu, F.; Shao, L.; Xia, G.S.; Bai, X. isaid: A large-scale dataset for instance segmentation in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seoul, Korea, 27–28 October 2019; pp. 28–37.
9. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. *arXiv* **2021**, arXiv:2110.08733.
10. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
11. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
12. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
13. Lin, G.; Milan, A.; Shen, C.; Reid, I. RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
14. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
15. Sun, X.; Chen, C.; Wang, X.; Dong, J.; Zhou, H.; Chen, S. Gaussian dynamic convolution for efficient single-image segmentation. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 2937–2948. [CrossRef]
16. Lv, Q.; Feng, M.; Sun, X.; Dong, J.; Chen, C.; Zhang, Y. Embedded Attention Network for Semantic Segmentation. *IEEE Robot. Autom. Lett.* **2021**, *7*, 326–333. [CrossRef]
17. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]

18. Johnson, B.A.; Ma, L. Image segmentation and object-based image analysis for environmental monitoring: Recent areas of interest, researchers' views on the future priorities. *Remote Sens.* **2020**, *11*, 1772. [CrossRef]
19. Li, Q.; Zorzi, S.; Shi, Y.; Fraundorfer, F.; Zhu, X.X. RegGAN: An End-to-End Network for Building Footprint Generation with Boundary Regularization. *Remote Sens.* **2022**, *14*, 1835. [CrossRef]
20. Chen, C.; Zhang, Y.; Lv, Q.; Wei, S.; Wang, X.; Sun, X.; Dong, J. Rrnet: A hybrid detector for object detection in drone-captured images. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019.
21. Nitze, I.; Heidler, K.; Barth, S.; Grosse, G. Developing and Testing a Deep Learning Approach for Mapping Retrogressive Thaw Slumps. *Remote Sens.* **2021**, *13*, 4294. [CrossRef]
22. Sun, X.; Zhang, M.; Dong, J.; Lguensat, R.; Yang, Y.; Lu, X. A deep framework for eddy detection and tracking from satellite sea surface height data. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 7224–7234. [CrossRef]
23. Guo, J.; Xu, Q.; Zeng, Y.; Liu, Z.; Zhu, X. Semi-Supervised Cloud Detection in Satellite Images by Considering the Domain Shift Problem. *Remote Sens.* **2022**, *14*, 2641. [CrossRef]
24. Mou, L.; Hua, Y.; Zhu, X.X. Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7557–7569. [CrossRef]
25. Zheng, Z.; Zhong, Y.; Wang, J.; Ma, A. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4096–4105.
26. Ma, A.; Wang, J.; Zhong, Y.; Zheng, Z. FactSeg: Foreground Activation-Driven Small Object Semantic Segmentation in Large-Scale Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–16. [CrossRef]
27. Yang, F.; Yuan, X.; Ran, J.; Shu, W.; Zhao, Y.; Qin, A.; Gao, C. Accurate Instance Segmentation for Remote Sensing Images via Adaptive and Dynamic Feature Learning. *Remote Sens.* **2021**, *13*, 4774. [CrossRef]
28. Wang, L.; Zhang, C.; Li, R.; Duan, C.; Meng, X.; Atkinson, P.M. Scale-aware neural network for semantic segmentation of multi-resolution remote sensing images. *Remote Sens.* **2021**, *13*, 5015. [CrossRef]