Deep Supervision-Based Simple Attention Network

Subjects: Remote Sensing

Contributor: Wenxu Shi , Qingyan Meng , Linlin Zhang , Maofan Zhao , Chen Su , Tamás Jancsó

Semantic segmentation for remote sensing images (RSIs) plays an important role in many applications, such as urban planning, environmental protection, agricultural valuation, and military reconnaissance. With the boom in remote sensing technology, numerous RSIs are generated; this is difficult for current complex networks to handle. Efficient networks are the key to solving this challenge.

convolutional neural network (CNN) deep supervision lightweight model

1. Introduction

Remote sensing is a crucial technical tool for large-scale observations of the Earth's surface. With the rapid development of Earth observation and remote sensing imaging technology, remote sensing has entered the era of big data [1]. Big data qualities for remote sensing primarily involve three Vs: volume, velocity, and variety of data [2]. Every day, a massive volume of remote sensing data must be handled in the era of big data for remote sensing. Furthermore, increasingly diverse remote sensing data are playing important roles in several fields. Due to advances in imaging technology, very high-resolution (VHR) imagery has shown considerable potential in remote sensing images (RSIs) interpretation and has been the focus of semantic segmentation.

Semantic segmentation is a critical task in computer vision, and its special application to remote sensing is RSI interpretation. It requires pixelwise parsing of the input image to retrieve the predefined categories to which the elements belong. Semantic segmentation has broad and vital applications in a variety of fields. This is especially true in the realm of remote sensing, where subjects such as integrated land use and land cover mapping [3][4], town change detection [5][6], urban functional areas [7], building footprints [8], impervious surfaces [9], and water body [10] extraction. The majority of these applications and methodologies are based on VHR images and are constrained by the two issues listed below. (1) Information modeling with little detail. In comparison to prior low-resolution images, VHR images give unequal spatial and semantic information volume gains. The significant improvement in spatial resolution allows for the observation of previously unseen features. However, vital detail information is mixed in with a vast volume of redundant information, providing additional obstacles for information extraction. (2) Inefficient processing. On the data processing front, high-resolution imagery implies that the amount of data to be processed per unit of observation area for interpretation is rising dramatically, posing a considerable challenge for hardware and algorithms.

Researchers have proposed numerous ways to overcome the difficulties of semantic segmentation for VHR images in the age of big data. Deep learning algorithms are the primary techniques for semantic segmentation at the moment. Unlike classic machine learning algorithms based on prior knowledge and predetermined rules, deep learning algorithms are data-driven algorithms that perform poorly with tiny data samples but may be utilized to great advantage in the era of big data. Deep learning-based convolutional neural networks (CNNs) outperform classic machine learning methods in terms of performance. Fully convolutional networks (FCNs) [11] have been utilized to obtain outstanding results in the semantic segmentation of RSIs. Following study, numerous model variants based on the FCN architecture have been developed, making substantial advances in various aspects. UNet [12], which is based on an encoder-decoder architecture, enhances the FCN's capacity to represent the multiscale features of images through contraction paths and expansion paths for achieving high-precision road [13] and coastline recognition [14] in RSIs. The DeepLabv3 series [15][16] utilize parallelized atrous spatial pyramid pooling (ASPP) with varying ratios to expand the models' reception fields while obtaining multiscale features; these models are widely used in RSI semantic segmentation, cloud detection [17], etc. However, because to the poor inference speeds of these models and the high hardware needs placed on deployed devices, these approaches find it difficult to overcome the aforementioned two problems. **Figure 1** depicts the problem of building segmentation models that take both efficiency and performance into account.

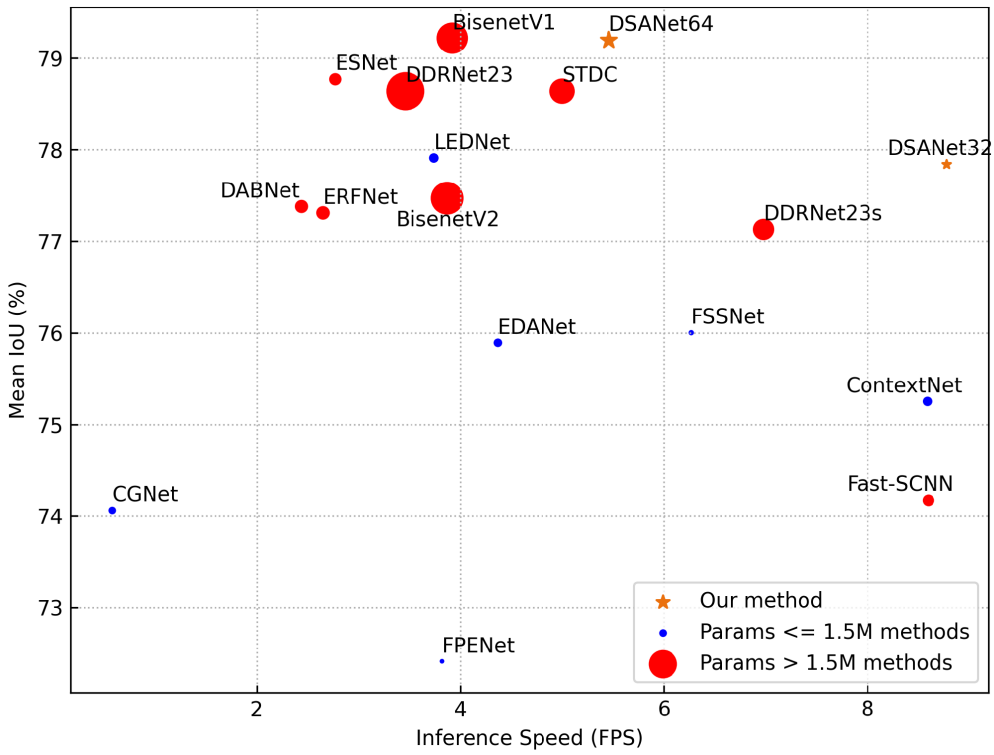

Figure 1. Speed-accuracy tradeoff yielded by different semantic segmentation methods on the ISPRS Potsdam dataset with a size of 6000 × 6000 pixels using an RTX 3090 GPU. Orange points: different versions of the proposed method. Red points: lightweight methods with more than 1.5 M parameters. Blue points: lightweight methods with less than 1.5 M parameters. The proposed methods achieve the best speed-accuracy tradeoffs. It is worth noting that that the sizes of the corresponding points of the methods are positively correlated with their parameters.

In addition to investigating model segmentation performance, another approach is to optimize the efficiency and accelerate the inference speed of the utilized model. A conceivable way to accomplish lightweight model building is to reduce the number of model channels and add an attention mechanism to compensate for the loss in model performance [18]. In addition to incorporating an attention module, the introduction of a deep supervision [19] module can also enhance the segmentation performance of the model. By actively monitoring the body and edge characteristics of the object of interest, a lightweight semantic segmentation network was suggested to maximize the overall consistency and object details of semantic segmentation results [20]. Loss functions expressly designed for the semantic segmentation task can speed up the learning process of the resultant model for fundamental spatial information such as borders [21] and spatial correlations [22], as evidenced by higher performance with the same amount of training epochs. These lightweight networks struggle to capture the rich, detailed aspects of VHR images with fewer parameters, reducing accuracy significantly.

2. Efficient Network Designs

Researchers are discovering that network design is becoming increasingly crucial as the Visual Geometry Group network (VGGNet) [23], the residual network (ResNet) [24], and DenseNet [25] models continue to be suggested. Because semantic segmentation is a dense prediction task, related models tend to have more parameters and slower inference speeds, which is harmful to model deployment and severely limits their application possibilities. An efficient network design paradigm lends itself well to the creation of efficient segmentation networks. By extensively replacing the 3 × 3 convolution in the model with a 1 × 1 convolution and reducing the number of channels in the 3 × 3 convolution, SqueezeNet [26] achieves comparable classification accuracy to AlexNet [27] with 2% of the total parameters. The MobileNet series [28][29][30] has steadily introduced new techniques to deep separable networks such as inverted residuals and neural architecture search (NAS). By integrating group convolution and channel shuffling operations and employing four recommendations, the ShuffleNet series [31][32] achieves a balance between accuracy and parameter number. 1. Equal channel widths minimize the memory access cost (MAC). 2. Excessive group convolution increases the MAC. 3. Network fragmentation reduces the degree of parallelism. 4. Elementwise operations are nonnegligible. Several outstanding and efficient semantic segmentation models have been presented as a result of these exploratory efforts on efficient network construction.

3. Efficient Semantic Segmentation Methods

Efficient semantic segmentation models strive for a balance between accuracy and speed, with considerable inference speed benefits at a low accuracy cost. They represent a significant development in the field of semantic segmentation in terms of efficiency, and they have created many good works based on the collaborative efforts of scholars. The two dominant approaches point the way to achieving high-accuracy and efficient semantic segmentation. 1. Light-weight backbones. ENet [33], a representative of earlier efficient segmentation models, greatly reduces the number of required parameters and floating point operations (FLOPs) by employing an asymmetric encoder-decoder structure and factorizing filters. Subsequent work has focused on asymmetric

networks, with the goal of improving model performance by using deeply separable convolutions [34], dilated convolutions [35], factorized convolutions [36][37], dense connections [38], skip connections [39], pyramidal pooling [40] and channel splitting and shuffling [41]. The Fast-shallow CNN (SCNN) [42] adopts shared shallow network paths to encode details while learning contexts at low resolutions, saving computing costs. STDCNet [38] utilizes a lightweight backbone network from DenseNet with layer concatenation. Dual-resolution branch networks [43], exemplified by the bilateral segmentation network (BiSeNet) series [44][45], provide effective segmentation by modifying extraction branches for spatial and semantic information independently. 2. Feature aggregation. The deep feature aggregation network (DFANet) [46] recommends two deep branches where several bilateral fusions are conducted. By steering upper-level feature upsampling using low-level features, SFNet [47] achieves higher-resolution restoration and cross-layer feature aggregation. DDRNet [48] advises two deep branches between which multiple bilateral fusions are performed.

4. Information Enhancement Modules

The information in computer vision tasks can be divided into spatial and semantic information, both of which contribute significantly to accurate segmentation. (1) Enhancing spatial information. Typically, the shallow layer of the encoder may better describe spatial information. Ensuring that a branch has a high resolution preserves spatial information to the greatest extent possible. STDCNet adopts the Laplacian kernel of the pyramid hierarchy as an auxiliary loss function, which expedites the process of learning spatial edge features. Researchers suggest that the quantifications and statistics of spatial texture aspects are likewise of great significance due to quantization and counting operators. (2) Enhancing semantic information. PSPNet [49] adopts pyramid pooling to enhance the observed multiscale semantic features. The DeepLab series [15][16][50][51] utilizes parallel atrous convolutions with varying dilation rates; this approach is called ASPP, which can encode multiscale semantic information more effectively. DANet [52] models long-range dependencies in the channels and positions of sematic features using a dual self-attention module. OCRNet [53] explicitly turns the pixel classification problem into an object area classification problem, computes the relationship between each pixel and each object region, and augments the representation of each pixel with an object-contextual representation.

5. Attention Mechanisms

The selected attention mechanism is a crucial component of model design and is a key module for improving model performance. It is a descriptive weighting of the relationship between a particular attribute (from a small pixel value to an entire channel) and the data, so that it can be chosen to suppress or amplify that attribute at a particular location in order to achieve a selective representation of a particular feature for the model. The outstanding early approach is the squeeze-and-excitation network (SENet) [54], which squeezes the features on each channel by global maximum pooling and uses a fully connected layer to encode the features into a low-dimensional space before performing decoding. This makes the SENet an excellent attention module without imposing many additional parameters or a large computational burden on the subject network. The SENet's concept of squeezing and extracting channels and examining spatial attention inspired further research. Important

follow-ups include the block attention module (BAM) [55] and convolutional BAM (CBAM) [56]. A BAM includes a two-branch parallel attention computation paradigm, with channel attention branches that adhere to the SENet's approach. Spatial features are squeezed in the channel dimension by a 1 × 1 convolution, key spatial features are extracted using a 3 × 3 convolution, and finally, a pixelwise summation operation is performed for both attention weights. A CBAM selects a multistep attention paradigm that combines channel attention and spatial attention simultaneously. The combination of spatial attention with gated mechanisms is another way to utilize attention mechanisms [57]. Unlike the idea of feature compression and extraction in the above work, self-attention [58] is a pixel-level attention mechanism. The computational complexity and resource needs of this method are an order of magnitude more than those of the preceding approaches, despite the fact that its performance is superior. Transformers [59], which outperform CNNs in many tasks, are excellent models based on self-attention; however, researchers are still designing optimizations for visual tasks such as patches [60] and hierarchical architectures [61] [62] to overcome the fatal flaw of a computationally intensive attention mechanism. Fortunately, self-attention based on queries, keys and values can be optimized from O(n2) complexity to linear complexity by changing the order of computation [63], performing approximate computation [64], and conducting low-rank singular value decomposition [65].

It is typical practice for effective semantic segmentation networks [33][44] to utilize an attention module based on the SENet or linear simplified self attention due to its computational efficiency and inference speed.

References

1. Liu, P. A survey of remote-sensing big data. Front. Environ. Sci. 2015, 3, 5.
2. Laney, D. 3D data management: Controlling data volume, velocity and variety. META Group Res. Note 2001, 6, 1.
3. der Sande, C.V.; Jong, S.D.; Roo, A.D. A segmentation and classification approach of IKONOS-2 imagery for land cover mapping to assist flood risk and flood damage assessment. Int. J. Appl. Earth Obs. Geoinf. 2003, 4, 217–229.
4. Costa, H.; Foody, G.M.; Boyd, D.S. Supervised methods of image segmentation accuracy assessment in land cover mapping. Remote Sens. Environ. 2018, 205, 338–351.
5. Im, J.; Jensen, J.; Tullis, J. Object-based change detection using correlation image analysis and image segmentation. Int. J. Remote Sens. 2008, 29, 399–423.
6. Chen, G.; Hay, G.J.; Carvalho, L.M.; Wulder, M.A. Object-based change detection. Int. J. Remote Sens. 2012, 33, 4434–4457.
7. Du, S.; Du, S.; Liu, B.; Zhang, X. Mapping large-scale and fine-grained urban functional zones from VHR images using a multi-scale semantic segmentation network and object based approach. Remote Sens. Environ. 2021, 261, 112480.

8. Wang, J.; Hu, X.; Meng, Q.; Zhang, L.; Wang, C.; Liu, X.; Zhao, M. Developing a method to extract building 3d information from GF-7 data. Remote Sens. 2021, 13, 4532.
9. Li, P.; Guo, J.; Song, B.; Xiao, X. A multilevel hierarchical image segmentation method for urban impervious surface mapping using very high resolution imagery. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2010, 4, 103–116.
10. Miao, Z.; Fu, K.; Sun, H.; Sun, X.; Yan, M. Automatic water-body segmentation from high-resolution satellite images via deep networks. IEEE Geosci. Remote Sens. Lett. 2018, 15, 602–606.
11. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
12. 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.
13. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual u-net. IEEE Geosci. Remote Sens. Lett. 2018, 15, 749–753.
14. Heidler, K.; Mou, L.; Baumhoer, C.; Dietz, A.; Zhu, X.X. Hed-unet: Combined segmentation and edge detection for monitoring the antarctic coastline. IEEE Trans. Geosci. Remote Sens. 2021, 60, 1–14.
15. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. arXiv 2017, arXiv:1706.05587.
16. 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.
17. Yang, J.; Guo, J.; Yue, H.; Liu, Z.; Hu, H.; Li, K. CDnet: CNN-based cloud detection for remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 2019, 57, 6195–6211.
18. Liu, S.; Cheng, J.; Liang, L.; Bai, H.; Dang, W. Light-weight semantic segmentation network for UAV remote sensing images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2021, 14, 8287–8296.
19. Lee, C.-Y.; Xie, S.; Gallagher, P.; Zhang, Z.; Tu, Z. Deeply-Supervised Nets. In Machine Learning Research, Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics; Lebanon, G., Vishwanathan, S.V.N., Eds.; PMLR: San Diego, CA, USA, 2015; Volume 38, pp. 562–570.

20. Deng, C.; Liang, L.; Su, Y.; He, C.; Cheng, J. Semantic segmentation for high-resolution remote sensing images by light-weight network. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 3456–3459.
21. Liu, S.; Ding, W.; Liu, C.; Liu, Y.; Wang, Y.; Li, H. ERN: Edge loss reinforced semantic segmentation network for remote sensing images. Remote Sens. 2018, 10, 1339.
22. Yuan, W.; Xu, W. Neighborloss: A loss function considering spatial correlation for semantic segmentation of remote sensing image. IEEE Access 2021, 9, 641–675.
23. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2014, arXiv:1409.1556.
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
25. Huang, G.; Liu, Z.; Maaten, L.V.D.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
26. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. arXiv 2016, arXiv:1602.07360.
27. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–8 December 2012; pp. 1097–1105.
28. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv 2017, arXiv:1704.04861.
29. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
30. Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 16–20 June 2019; pp. 1314–1324.
31. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856.

32. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. Shufflenet v2: Practical guidelines for efficient CNN architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 14–18 September 2018; pp. 116–131.
33. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. Enet: A deep neural network architecture for real-time semantic segmentation. arXiv 2016, arXiv:1606.02147.
34. Li, G.; Yun, I.; Kim, J.; Kim, J. Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. In Proceedings of the British Machine Vision Conference (BMVC), Cardiff, UK, 9–12 September 2019; pp. 186.1–186.12.
35. Lo, S.-Y.; Hang, H.-M.; Chan, S.-W.; Lin, J.-J. Efficient dense modules of asymmetric convolution for real-time semantic segmentation. In Proceedings of the ACM Multimedia Asia, Nice, France, 21–25 October 2019; pp. 1–6.
36. Romera, E.; Alvarez, J.M.; Bergasa, L.M.; Arroyo, R. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. IEEE Trans. Intell. Transp. Syst. 2017, 19, 263–272.
37. Zhang, X.; Chen, Z.; Wu, Q.J.; Cai, L.; Lu, D.; Li, X. Fast semantic segmentation for scene perception. IEEE Trans. Industr. Inform. 2018, 15, 1183–1192.
38. Fan, M.; Lai, S.; Huang, J.; Wei, X.; Chai, Z.; Luo, J.; Wei, X. Rethinking bisenet for real-time semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Montreal, QC, Canada, 11–17 October 2021; pp. 9716–9725.
39. Chaurasia, A.; Culurciello, E. Linknet: Exploiting encoder representations for efficient semantic segmentation. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), Venice, Italy, 22–29 October 2017; pp. 1–4.
40. Liu, M.; Yin, H. Feature pyramid encoding network for real-time semantic segmentation. arXiv 2019, arXiv:1909.08599.
41. Wang, Y.; Zhou, Q.; Liu, J.; Xiong, J.; Gao, G.; Wu, X.; Latecki, L.J. Lednet: A lightweight encoder-decoder network for real-time semantic segmentation. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22—25 September 2019; pp. 1860–1864.
42. Poudel, R.P.; Liwicki, S.; Cipolla, R. Fast-scnn: Fast semantic segmentation network. arXiv 2019, arXiv:1902.04502.
43. Poudel, R.P.; Bonde, U.; Liwicki, S.; Zach, C. Contextnet: Exploring context and detail for semantic segmentation in real-time. arXiv 2018, arXiv:1805.04554.
44. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 325–341.

45. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. Int. J. Comput. Vis. 2021, 129, 3051–3068.
46. Li, H.; Xiong, P.; Fan, H.; Sun, J. Dfanet: Deep feature aggregation for real-time semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9522–9531.
47. Li, X.; You, A.; Zhu, Z.; Zhao, H.; Yang, M.; Yang, K.; Tan, S.; Tong, Y. Semantic flow for fast and accurate scene parsing. In European Conference on Computer Vision; Springer: Berlin/Heidelberg, Germany, 2020; pp. 775–793.
48. Hong, Y.; Pan, H.; Sun, W.; Jia, Y. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. arXiv 2021, arXiv:2101.06085.
49. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
50. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv 2014, arXiv:1412.7062.
51. Chen, L.C.; Papreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 40, 834–848.
52. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
53. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. In European Conference on Computer Vision; Springer: Berlin/Heidelberg, Germany, 2020; pp. 173–190.
54. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
55. Park, J.; Woo, S.; Lee, J.-Y.; Kweon, I.S. Bam: Bottleneck attention module. arXiv 2018, arXiv:1807.06514.
56. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
57. Meng, Q.; Zhao, M.; Zhang, L.; Shi, W.; Su, C.; Bruzzone, L. Multilayer feature fusion network with spatial attention and gated mechanism for remote sensing scene classification. IEEE Geosci.

Remote Sens. Lett. 2022, 19, 1–5.

58. Ambartsoumian, A.; Popowich, F. Self-attention: A better building block for sentiment analysis neural network classifiers. arXiv 2018, arXiv:1812.07860.
59. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Bench, CA, USA, 4–9 December 2017; pp. 5998–6008.
60. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv 2020, arXiv:2010.11929.
61. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
62. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 568–578.
63. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Wang, L.; Atkinson, P.M. Abcnet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. ISPRS J. Photogramm. Remote Sens. 2021, 181, 84–98.
64. Wang, S.; Li, B.Z.; Khabsa, M.; Fang, H.; Ma, H. Linformer: Self-attention with linear complexity. arXiv 2020, arXiv:2006.04768.
65. Guo, M.-H.; Liu, Z.-N.; Mu, T.-J.; Hu, S.-M. Beyond self-attention: External attention using two linear layers for visual tasks. arXiv 2021, arXiv:2105.02358.