An Adaptive Spatial Network for UAV Image Real-Time Semantic Segmentation

Qian Wu, Jiayu Song, Yanbo Luo, Hao Li, Qin Tian, Qi Wang, Jinglong Gao, Zhuoran Jia

An Adaptive Spatial Network for UAV Image Real-time Semantic Segmentation *

Qian Wu[1], JiaYu Song[1], YanBo Luo[1], Hao Li[1], Qin Tian[1], Qi Wang[1], Jinglong Gao[1], and Zhuoran Jia[1]

Institute of Information and Navigation, Air Force Engineering University, Xi′an, Shaanxi Province 710038, China
zhuangzaierhao@163.com;321585573@qq.com;lybobo0515@163.com;
lihao13lb@gmail.com;15191432029@139.com;
570557058@qq.com;jzr0508@qq.com;1183430308@qq.com

Abstract. Unmanned aerial vehicle(UAV) aerial image interpretation plays an important role in the military and civilian files. The latest semantic segmentation methods are based on deep learning with different structure to encoder spatial feature. However, they are larger networks which are not effective for UAV with limited resources. Thus, a real-time adaptive spatial structure semantic segmentation network, ASRNet, is proposed for UAV aerial image. Firstly, ASRNet is based on an encoder-decoder structure with a module called local structure feature descriptor in the middle. Secondly, the descriptor utilizes features at different abstraction levels from both the encoder and decoder to describe different target with higher spatial resolution adaptively. Lastly, the local structure feature descriptor enables a better gradient flow from deeper layers to shallower layers by adding short paths for the back-propagation. The experiments validate the effectiveness of the proposed method from the accuracy and computation time.

Keywords: UAV Aerial Image · Real-time Semantic Segmentation · Encoder-Decoder · Adaptive Feature Descriptor.

1 Introduction

With the development of UAV technology, image interpretation on the UAV platform is of increasing importance in the application of urban scene observation, reconnaissance and navigation[1, 2]. Semantic segmentation is an essential task for the UAV image interpretation. Recently, the state-of-the-art semantic segmentation methods are based on deep neural network with multi-layer nonlinear function, which maps the input image to the semantic label output[3]. Despite these advances of deep learning methods, it remains a challenging task for UAV image.

* Supported by the Natural Science Basic Research Plan in ShaanXi Province of China under Grant 2022JQ-0344.

Most aerial images obtained by UAV are with oblique views which have a much larger land coverage. The images in oblique view have very large spatial resolution variation across the entire image. As shown in 1, the semantic segmentation of a small car in the aerial image is better handled in higher resolution where finer details can be observed, such as wheels. For larger objects like roads and buildings, it is better to have more global context with low resolution to recognize the objects since their whole shapes can be observed for semantic segmentation. For the boundary between different objects, such as vegetation and road, the detailed local descriptor with high resolution should be adopted to distinguish different category. When designing the deep neural networks, there is usually a performance trade-off for objects in different scales.

Fig. 1. Example of images with different scales objects, the wheels in red is in small scale, the vegetation in green in large object, the boundary between vegetation and road is in small scale.

1.1 Related Work

The end-to-end semantic segmentation of deep learning method is introduced by Long[4] with the seminal fully convolutional network (FCN). In order to handle the complex spatial structure with different resolution targets, various network structures have been designed.

1) Context-based models: To capture the contextual information at multiple scales, DeepLabV2[6] and DeeplabV3[7]] exploit multiple parallelatrous convolutions with different dilation rates, while PSPNet[9] performs multi-scale spatial pooling operations. Although these methods encode rich contextual information, they can not capture boundary details effectively due to strided convolution or pooling operations[8].

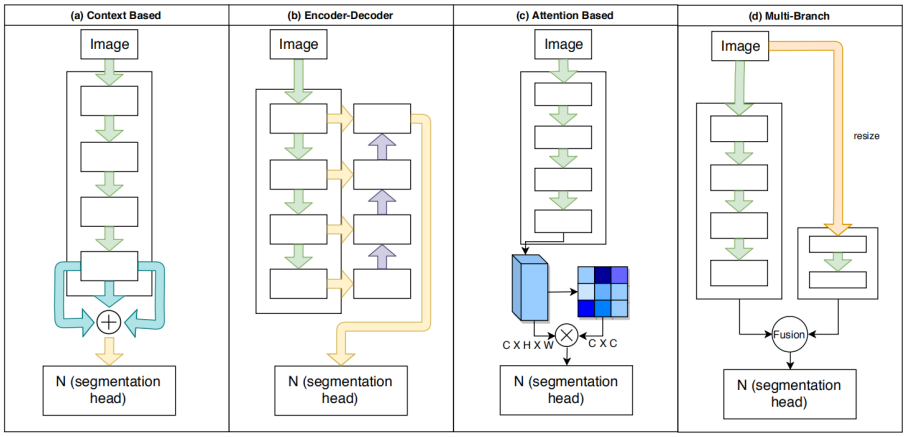

Fig. 2. Illustration of different semantic segmentation methods[20].

2) Encoder-decoder structure: Encoder extracts global contextual information and decoder recovers the spatial information. To meet the real-time requirements of CNN tasks in mobile devices, a lightweight network is proposed proposed, named LEDNet, which adopts asymmetric encoder-decoder architecture. Encoder Network is a network similar to FCN, and the decoder is an attention pyramid[5]. Sun et al. augments the high-resolution representation by aggregating the (up-sampled) representations from all the parallel convolutions rather than only the representation from the high-resolution convolution[19]. However, implementation of dilated convolution at higher dilation rates is computationally intensive making them unsuitable for real-time applications.

3) Attention-based models: Attention mechanisms, which help networks to focus on relevant information and ignore the irrelevant information. Wang et al. [13] formalized self-attention by calculating the correlation matrix between each spatial point in the feature maps in video sequences. To capture contextual information, LEDnet[5], DaNet[12] and OCNet[14] apply a self-attention mechanism. PSANet[15] learns to aggregate contextual information for each individual position via a predicted attention map. Attention models, however, generally require expensive computation cost.

4) Multi-Branch models:The deeper branches extract the contextual information by enlarging receptive fields or shallower branches retain the spatial structure, which makes them suitable for run time efficient implementations[16, 17]. However, they are mostly applicable to the relatively simpler datasets with fewer number of classes. On the other end, HRNet[18] proposed a model with fully connected links between output maps of different resolutions. However, without reduction of spatial dimensions of features, the computational overhead is very high and makes the model no longer feasible for real-time usage.

Building on these observations, we propose a real-time general purpose semantic segmentation architecture that obtains deep features with high resolution

resulting in improved accuracy and lower latency in a single branch encoder-decoder network.

1.2 Motivation

Semantic segmentation, which associates each pixel to the object class it belongs to, is a computationally expensive task in computer vision. Fast semantic segmentation is broadly applied to several real-time applications including autonomous driving, medical imaging and robotics. However, accurate CNN-based semantic segmentation requires larger neural networks which are therefore not suitable for UAV as they are cumbersome and require substantial resources.

Down-sampling operations, such as pooling and convolutions with stride greater than one, can help decrease the latency of deeper neural networks, however they result in decreased pixel-level accuracy due to the lower resolutions at deeper levels. Many recent approaches employ either encoder-decoder structure, a two or multi-branch architecture or dilated convolutions to recover spatial information. While these real-time architectures perform appropriately on simple datasets, their performance is sub-optimal for complex datasets possessing more variability in terms of classes, sizes, and shapes.

Thus, there is a significant interest in designing CNN architectures that can perform well on UAV datasets and, at the same time, are mobile enough to be of practical use in real-time applications of UAV aerial images.

1.3 Contributions

In this paper, A real-time adaptive spatial structure semantic segmentation network, ASRNet, is proposed that performs well on complex scenarios. ASRNet is based on an asymmetric encoder-decoder structure with a new module called local feature descriptor in the middle. The descriptor utilizes features at different abstraction levels from both the encoder and decoder to improve the feature refinement at a given level allowing the network to preserve deeper level features with higher spatial resolution. Furthermore, the descriptor enables a better gradient flow from deeper layers to shallower layers by adding short paths for the back-propagation. Since training an average deep learning model has a considerable carbon footprint, we reduce the training time by 60% with negligible effect on performance by applying progressive resizing for training.

The contributions are summarized as follows: We propose ASRNet as a real-time semantic segmentation architecture that obtains deep features with high resolution resulting in improved accuracy and lower latency in a single branch network. It performs competitively in complex environments. We introduce an adaptive local descriptor module to capture multiple levels of abstraction to help in boundary refinement of segments. Besides, progressive resizing technique is adopted during the training which leads to 60% reduction in training time and the environmental impact. We combat aliasing effect in label map on lower resolutions by employing a modified label relaxation.

The remainder of the paper is organized as following. In section II, the proposed adaptive spatial network for real-time semantic segmentation(ASRNet) is shown in detail. The analysis of parameters and the experimental results on one UAV image are described in section III to validate the real-time and effectiveness. In the end, the conclusions and future work are drawn in section IV.

2 Proposed Approach

ASRNet is based on a light-weight encoder-decoder structure for fast and efficient inference. It comprises of three components: an encoder which extracts high-level semantic features, a light asymmetric decoder, and an local feature descriptor which links different stages of encoder and decoder. The encoder decreases the resolution and increases the number of feature maps. The decoder reconstructs the lost spatial information. The local feature descriptor combines the information to preserve and refine the information between multiple levels.

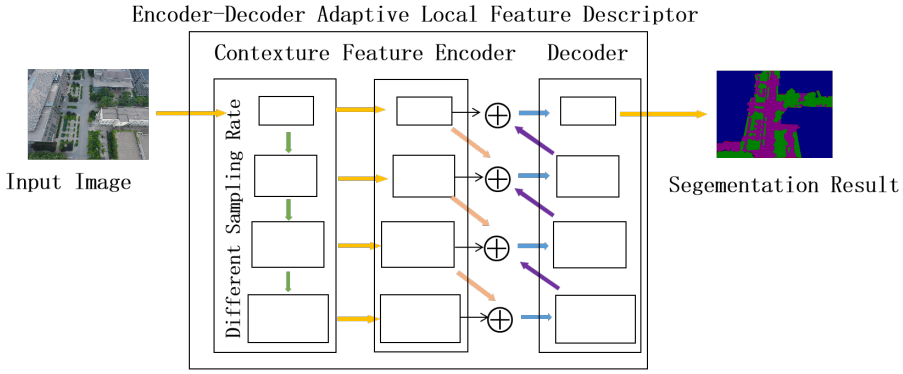

Fig. 3. Framework of the proposed ASRnet with the local feature descriptor "Encoder" and label relaxation: "Decoder".

2.1 Structure of ASRnet

ASRnet architecture is illustrated in 3. All the tensors have the same spatial resolution with the number of channels in the same row. Four level outputs are extracted from the encoder at different spatial resolutions $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$ and $\frac{1}{32}$ with 256, 512, 1024 and 2048 channels, respectively. The number of channels are reduced by a factor of four using 1×1 convolutions followed by batch norm and ReLU activation function at each level. These outputs are then passed through a decoder structure with descriptor in the middle. Finally, segmentation output is extracted from the largest resolution via 1×1 convolution to match the number

of channels to segmentation categories. The object function of the three levels is as below:

$$x_s^a = D(T(x_{s-1}^a)) + T(x_s^e) + U(x_{s+1}^d) \tag{1}$$

where a, e, and d denote descriptor, encoder, and decoder respectively. Besides, s represents the spatial level in the semantic segmentation network. $D(\cdot)$ and $U(\cdot)$ are down-sampling and up-sampling functions. Down-sampling is carried out by convolution with stride 2 and up-sampling is carried out by de-convolution with stride 2 matching spatial resolution as well as the number of channels in the current level. $T(\cdot)$ is a transfer function that reduces the number of output channels from an encoder block and transfers them to the descriptor:

$$T(x_s^e) = \sigma(w_s^a \otimes x_s^e + b_s^a) \tag{2}$$

where w and b are the weight matrix and bias vector, $\otimes$denotes the convolution operation, and σ is the activation function. The decoder contains a modified basic residual block, F, where we use shared weights within the block. The decoder function is as follows:

$$x_s^d = F(x_s^m; {}_s^d) \tag{3}$$

2.2 Adaptive Local Feature Descriptor

Stepwise resizing is a commonly used technique to reduce training time for classification. The image size is small at the beginning of the training, and then gradually increased until the final stage of the training using the original image size. However, applying progressive resizing in semantic segmentation is more challenging because it needs to be applied to the image and its corresponding label mapping. Bilinear or bicubic interpolation cannot be applied to label maps because they exist in integer space, and these methods will result in floating point values for tags. Besides, nearest neighbor interpolation for resizing introduces noise into label maps near object boundaries due to aliasing. Thus, Inspired by Zhu et al.[9], an optimized variant of the label relaxation method named local feature descriptor is proposed, as shown in Fig. 4, in order to reduce the influence of boundary artifacts in the gradual adjustment of the label map adaptively.

In the cross entropy loss function, the negative logarithmic likelihood of soft-maximum probability is used for a given label. In contrast, label relaxation is a loss function where the negative logarithmic likelihood of soft-maximum probability for a given label as well as for adjacent pixel labels is maximized. This is established by taking the sum of the soft maximum probabilities mentioned earlier before applying negative log likelihood. We identify boundary pixels as those with multiple unique labels in the window centered around the kernel size K. The loss at a given boundary pixel is calculated as follows, where N is the boundary label set:

$$\mathcal{L} = -\log \sum_{C \in N} P(C) \tag{4}$$

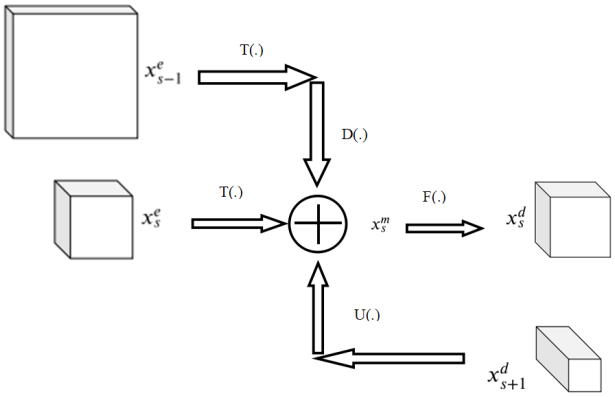

Fig. 4. Illustration of the local feature descriptor.

To apply label relaxation effectively, a hot label from the label map is created, followed by a max-pooling operation with Stride 1. This will effectively expand each single hot label channel and transform it into multi-hot labels along the boundary, thus realizing optimal selection of boundary pixels and their corresponding labels. Border pixels are usually in the minority. The loss function applies only to boundary pixels, and the normal cross entropy loss applies to the remaining pixels.

3 Experiments

The experiments are conducted on Urban Drone Dataset(UDD)[21] as the complex dataset. UDD are collected at Peking University, Huludao city, Henan University and Cangzhou city, which includes six categories: Vegetation, Building, Road, Vehicle, Roof and Other. All the results are with the average of 5 experiments.

The ASRnet are implemented based on PyTorch framework[22]. For training, a polynomial learning rate policy is employed where the initial learning rate is multiplied by $(1 - iter/totaliter)^{0.9}$ after each iteration. The learning rate is set to 1×10^{-3}. Momentum and weight decay coefficients are set to 0.9 and 1×10^{-4}, respectively.

The performance evaluation indexes used in this part are the ratio of pixels correctly classified named pixel accuracy(PA), the mean of PA named mean pixel accuracy(MPA) and the mean intersection over union(mIOU). These three indexes are computed with the following equations:

$$PA = \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{i=0}^{k}} \sum_{j=0}^{k} p_{ij} \tag{5}$$

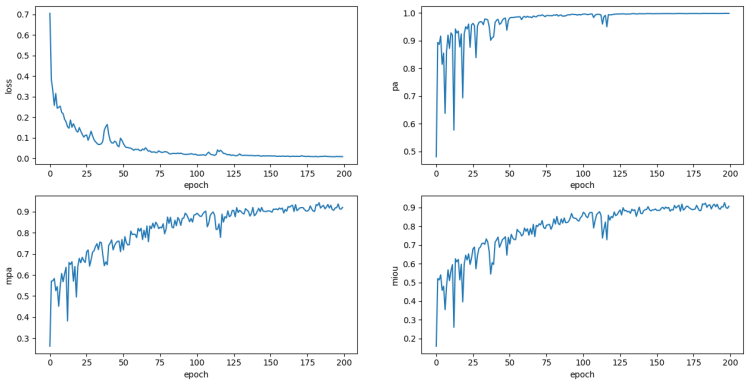

Fig. 5. Indexes of segmentation results of methods ASRnet, LEDnet, Segnet, Unet and DeeplabV3.

$$MPA = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{i=0}^{k}}\sum_{j=0}^{k}p_{ij} \tag{6}$$

$$mIOU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \tag{7}$$

First of all, the convergence of the algorithm is evaluated on the dataset UDD as shown in Fig. 5, four curves corresponding to the loss and three performance indexes are given. It can be seen that the loss is decreasing with each iteration, while the three performance indexes of PA MPA and mIOU are increasing with each iteration. The general trend shows that the proposed ASRnet can converge in a much faster time.

Two real-time semantic segmentation methods LEDnet[5], Unet[10] and two non real-time methods Segnet[19], DeeplabV3[11] are compared with the proposed method to validate the effectiveness from the perspective of accuracy and time. These two methods are both based on encoder-decoder structure with different local feature descriptor. The segmentation result images are given in Fig. 6 for the compared methods. From the visual results in Fig. 7, it can be seen some pixels of road are classified wrongly which results from the illumination intensity in the first image. Because the detail local feature descriptor, the narrow road can be segmented from the background which can not be realized by the compared methods. Besides, in the last image there are great differences in shape and size of different object. ASRnet can segment the small cars more accurate than the compared methods due to the adapt local feature descriptor.

From the segmentation index values given in Fig. 7 with PA, MPA and mIOU of the compared methods. The index PA and mIOU of the proposed ASRnet is the best among the compared methods, higher almost 5% to 15%, as for

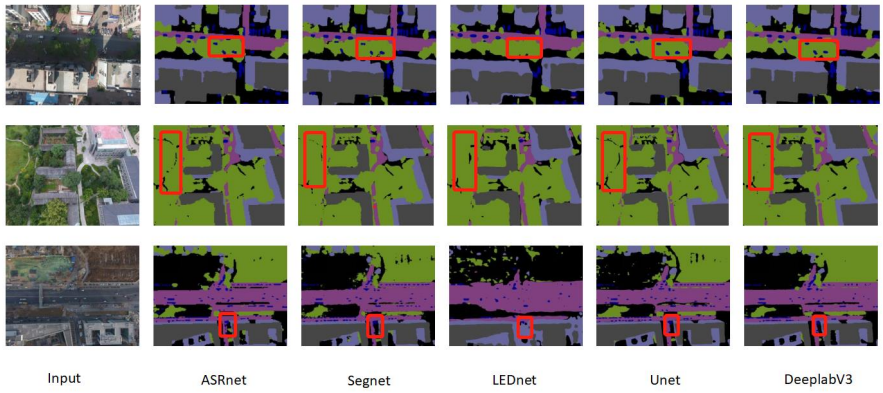

Fig. 6. Segmentation results on data set UDD with different methods including: (a) is the aerial image of UAV; (b) - (f) are the segmentation results with methods ASRnet, LEDnet, Segnet, Unet and DeeplabV3, respectively.

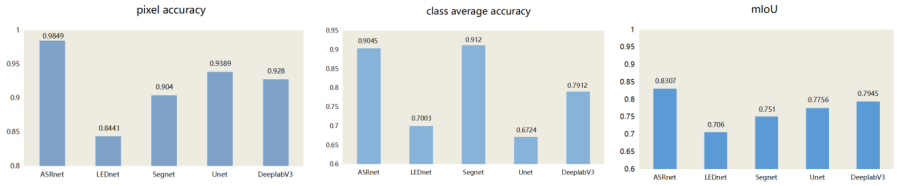

Fig. 7. Indexes of segmentation results of methods ASRnet, LEDnet, Segnet, Unet and DeeplabV3.

the adaptive descriptor of local feature. A detailed descriptor is adopted for the boundary to improve the segmentation result with complex texture. While a larger local feature descriptor is adopted for the smooth region to reduce the computation complexity. Thus, it results into the improvement of overall accuracy value. Besides, the MPA value of the proposed method is similar to the best one of Segnet which results from the much more computation time.

As for the computation time, the results of the five compared methods are given in Fig. 8. It can be seen the time of real-time methods ASRnet, LEDnet and DeeplabV3 all are 3 times less than that of Segnet. Although the computation time of LEDnet is less than the proposed method, the segmentation accuracy value shown in Fig. 8 of LEDnet is much less than Segnet and ASRnet.

Therefore, by comparing the semantic segmentation networks designed for specific datasets, ASRnet is a real-time semantic segmentation model that performs competitively both on semantic segmentation results and computation time.

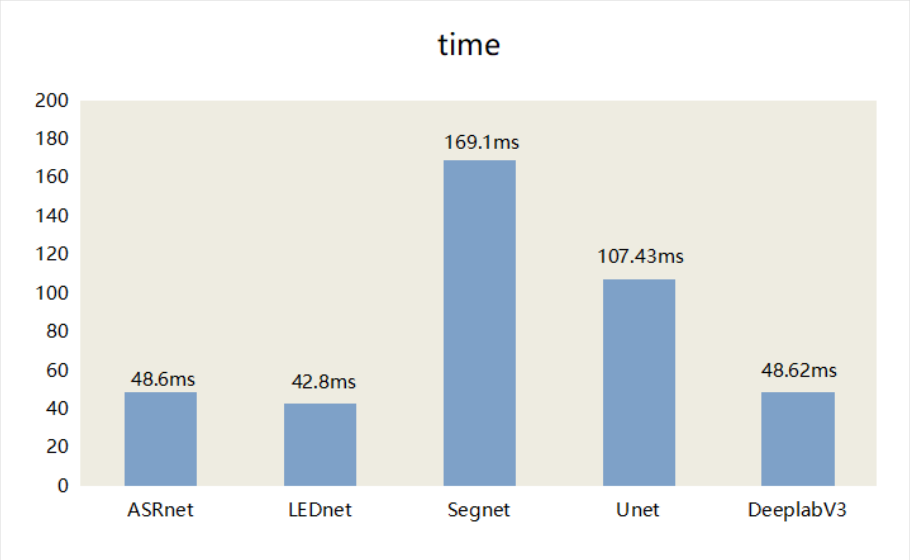

Fig. 8. Computation time of the compared methods.

4 Conclusion

In this paper, an adaptive spatial structure learning network is proposed for UAV aerial image semantic segmentation. It incorporates a local feature descriptor that aggregates features from different abstraction levels and coordinates with encoder-decoder framework. This model is conceptually simple yet effective to achieve efficient inference speed and accuracy on resource constrained devices UAV in complex environment. By employing an optimized progressive resizing training scheme, the training time on dataset UDD is less 3 times than the state-of-the-art non real-time method. Overall, the proposed ASRnet can generate semantic segmentation results in real-time and comparable accuracy. This optimal balance of speed and accuracy makes our model suitable for real-time applications of UAV aerial images where the environment is highly dynamic due to the presence of high variability in real world scenarios.

Acknowledgements Thanks the open datasets UDD of UAV to validate the proposed semantic segmentation method.

References

1. Demir, I., Koperski, K. and Lindenbaum, D. et al.: DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images, IEEE,2018,
Author, F.: Article title. Journal **2**(5), 99–110 (2016)

2. Ivancsits C , Lee M.: Visual navigation system for small unmanned aerial vehicles[J]. Sensor Review, 2013, 33(3):267-291.
3. Ye L, Vosselman G, Xia G S, et al.: Bidirectional Multi-scale Attention Networks for Semantic Segmentation of Oblique UAV Imagery[J]. 2021.
4. Jonathan Long, Evan Shelhamer, and Trevor Darrell.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431âĂŞ3440, 2015.
5. Yu Wang, Quan Zhou, Jia Liu,et.al.: Lednet: A Lightweight Encoder-Decoder Network for Real-time Semantic Segmentation. IEEE International Conference on Image Processing, October 25-28, 2020
6. Chen L C , Papandreou G , Kokkinos I , et al. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs[J]. Computer Science, 2014(4):357-361.
7. Chen L C , Papandreou G , Kokkinos I , et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4):834-848.
8. Chen L C , Zhu Y , Papandreou G , et al. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation[J]. Springer, Cham, 2018.
9. Zhu Y , Sapra K , Reda F A , et al. Improving Semantic Segmentation via Video Propagation and Label Relaxation[J]. 2018.
10. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation[J]. Springer International Publishing, pp.234âĂŞ241, 2015.
11. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4):834âĂŞ848, 2017.
12. Jun Fu, Jing Liu, Yuhang Wang, et al. Stacked deconvolutional network for semantic segmentation. IEEE Transactions on Image Processing, 2019.
13. Xiaolong Wang, Ross Girshick, Abhinav Gupta, et al. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.7794âĂŞ7803, 2018.
14. Yuan Y, Huang L, Guo J, et al. OCNet: Object Context for Semantic Segmentation[J]. International Journal of Computer Vision, 2021:1-24.
15. Hengshuang Zhao, Yi Zhang, Shu Liu, et al. Psanet: Point-wise spatial attention network for scene parsing. In Proceedings of the European Conference on Computer Vision (ECCV), pages 267âĂŞ283, 2018.
16. Yu C, Wang J, Peng C, et al. BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation[J]. Springer, Cham, 2018.
17. Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, et al. Icnet for real-time semantic segmentation on high-resolution images. In Proceedings of the European Conference on Computer Vision (ECCV), pages 405âĂŞ420, 2018.
18. Sun K, Zhao Y, Jiang B, et al. High-Resolution Representations for Labeling Pixels and Regions. 2019.
19. Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12):2481âĂŞ2495, 2017.
20. Arani E , Marzban S , Pata A , et al. RGPNet: A Real-Time General Purpose Semantic Segmentation[J]. 2019.

21. Chen Y , Wang Y , Lu P , et al.: Large-Scale Structure from Motion with Semantic Constraints of Aerial Images. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV), pp. 347–359. Springer(2018).
22. Paszke A , Gross S , Chintala S , et al. Automatic differentiation in PyTorch. In: Conference and Workshop on Neural Information Processing Systems(NeurIPS) Workshop 2017.