Multimedia Tools and Applications (2022) 81:29283–29304
https://doi.org/10.1007/s11042-022-12842-y

Supervised semantic segmentation based on deep learning: a survey

Yuguo Zhou[1] · Yanbo Ren[1] · Erya Xu[1] · Shiliang Liu[1] · Lijian Zhou[1]

Received: 19 February 2021 / Revised: 8 February 2022 / Accepted: 9 March 2022 /
Published online: 2 April 2022

Abstract

Recently, many semantic segmentation methods based on fully supervised learning are leading the way in the computer vision field. In particular, deep neural networks headed by convolutional neural networks can effectively solve many challenging semantic segmentation tasks. To realize more refined semantic image segmentation, this paper studies the semantic segmentation task with a novel perspective, in which three key issues affecting the segmentation effect are considered. Firstly, it is hard to predict the classification results accurately in the high-resolution map from the reduced feature map since the scales are different between them. Secondly, the multi-scale characteristics of the target and the complexity of the background make it difficult to extract semantic features. Thirdly, the problem of intra-class differences and inter-class similarities can lead to incorrect classification of the boundary. To find the solutions to the above issues based on existing methods, the inner connection between past research and ongoing research is explored in this paper. In addition, qualitative and quantitative analyses are made, which can help the researchers to establish an intuitive understanding of various methods. At last, some conclusions about the existing methods are drawn to enhance segmentation performance. Moreover, the deficiencies of existing methods are researched and criticized, and a guide for future directions is provided.

✉ Lijian Zhou
zhoulijian@qut.edu.cn

Yuguo Zhou
zhouyuguo@qut.edu.cn

Yanbo Ren
christopher0527@163.com

Erya Xu
1030061939@qq.com

Shiliang Liu
1156844855@qq.com

[1] School of information and Control Engineering, Qingdao University of Technology, Qingdao 266525 Shandong, China

Keywords Semantic segmentation · Supervised learning · Computer vision · Convolutional neural networks

1 Introduction

Object detection, image classification, and semantic segmentation are the three major research topics in computer vision. They are the basis of various complex vision tasks. From the perspective of classification, all pixels in semantic segmentation should be assigned labels. Compared with image-level image classification and region-level object detection, pixel-level semantic segmentation in the regional level classification is more challenging.

Traditional methods make use of shallow features more, such as edges, colors, spatial textures, and geometric shapes. In the statistical methods [1–4], the probability graph model is established, and the image is divided into several disjoint areas to separate the target from the background. Three kinds of segmentation methods can be categorized based on the region [5], edge detection [6], graph theory [7]. However, these methods are mainly dependent on expert systems. Expert systems rely only on artificially designed shallow features and cannot use some deep and hidden features, which leads to limited use of image features.

Due to the rapid development of deep learning, a series of deep neural networks make great achievements, in particularly, convolutional neural networks such as VGG [8], AlexNet [9], GoogLeNet [10], ResNet [11]. The emergence of these networks brings new solutions to semantic segmentation. Compared with the traditional statistical methods, the methods based on a convolutional neural network have natural advantages. It can automatically extract the semantic features instead of the biased manual feature extraction in the original method. And it is an end-to-end processing structure, in which the prediction map can be obtained directly in the output layer. Fully Convolutional Network (FCN) proposed by Long et al. [12] is the first application of convolutional neural networks in semantic segmentation.

To the best of our knowledge, there are various semantic segmentation surveys [13–15], which do a lot of work in summarizing methods, datasets and discussing future prospects. However, some of the latest research are not included. Moreover, these works are too focused on the methods to neglect the discussion of the key issues in semantic segmentation. Our work is completely different. We discuss the key issues affecting the segmentation effects, which may help researchers to define research priorities. The main goal of this work is to explore the inner connection between past research and ongoing research and provide a series of solutions to the above issues based on existing methods. Because of that, Our work is novel and useful, which is a significant contribution for academic research.

The main contributions of our work are as follows:

- Different from method-based surveys, our survey is studied with a new viewpoint, in which the key issues affecting the segmentation effects are considered first.
- To find solutions of these key issues, we provide a broad and organized survey of existing methods.
- A quantitative evaluation and comparison of current methods are made, which can help the researchers to establish an intuitive understanding of various methods.
- The aforementioned results are discussed, and some conclusions are given, in which some existing methods can be combined to enhance segmentation performance.

- By researching and criticizing the deficiencies of existing methods, a guide for future directions is provided.

The remainder of this paper is organized as follows. Section 2 discusses the key issues affecting the segmentation effects and their origins. Section 3 shows that the qualitative analysis of existing methods. Section 4 mainly contains a quantitative evaluation of various methods and brief discussion based on above results. In addition, Section 4 shows our conclusions about aggregating existing methods to enhance segmentation performance. Section 5 research the deficiencies of existing methods and provide researcher with new research directions.

2 The key issues affecting segmentation effect

Due to the complexity of semantic segmentation, it is difficult to obtain highly accurate prediction maps. The key issues affecting segmentation effect mainly come from three aspects: Resolution, Context and Boundary.

2.1 Resolution

It is necessary to obtain a high-resolution prediction image in semantic segmentation. However, in the process of pursuing semantic information, a larger sensory area of kernels is also necessary. Since some works use down-sampling to increase the sensory area of kernels, the image resolution will decrease continually. Thus, there is a scale contradiction between the reduced feature map and the high-resolution prediction map. In addition, the spatial structure and small objects information is loss in the low-resolution image. “How to solve the scale contradiction between reduced feature map the high-resolution prediction map” is an urgent problem to be solved.

2.2 Context

The extraction of context is a critical factor for semantic segmentation. On the one hand, the objects have the multi-scale characteristics due to the different shooting angles. But a single convolution kernel is difficult to adapt to multi-scale objects. On the other hand, the relationship between different objects is also hard to determine due to the diversity of backgrounds. Some objects with the context relation can promote each other’s learning. For example, if the road has been detected, then the probability of detecting cars should be higher than ones of the boats. Therefore, the context is beneficial to pixel classification. “How to fully learn context” is a challenging problem for semantic segmentation.

2.3 Boundary

Semantic segmentation focuses on the classification of pixels. However, due to the lack of features, there are some phenomena about intra-class differences and inter-class similarities. The network lacks discriminative ability in the classification of boundary pixels. The wrong segmentation of boundary pixels brings some troubles to the segmentation task. “How to improve the overall segmentation accuracy from the perspective of boundary correction" is worthy of further discussion.

2.4 Classification

To find solutions of these key issues, we provide a broad and organized survey of existing methods and use the above three factors as the main criteria. According to the specific structure, fully supervised semantic segmentation is divided into six categories. The classification is shown in Fig. 1.

3 Supervised semantic segmentation

Recently, many semantic segmentation methods based on fully supervised learning are leading the way in many computers vision. Supervised learning uses manually labeled samples for training. These samples can provide a lot of detailed information, which can help to improve training efficiency and segmentation accuracy. According to the three key issues discussed above, the current supervised semantic segmentation methods is be divided into corresponding parts to discuss: super-resolution reconstruction in segmentation, semantic enhancement and boundary correction.

3.1 Super-resolution reconstruction in segmentation

According to the difference of network depth, the features to be processed are usually divided into shallow features and deep features. Shallow features mainly include details and spatial structure information while deep features contain semantic information. To obtain a high-resolution prediction map, shallow information such as details and spatial structure are as important as the deep semantic information. To balance the utilization of shallow and deep features, there are currently two main methods. The focus of the Encoder-Decoder method is to enrich the encoding and decoding structure which can improve the information extraction and recovery capabilities. The Multi-path Fusion method focuses on the combination of shallow and deep information.

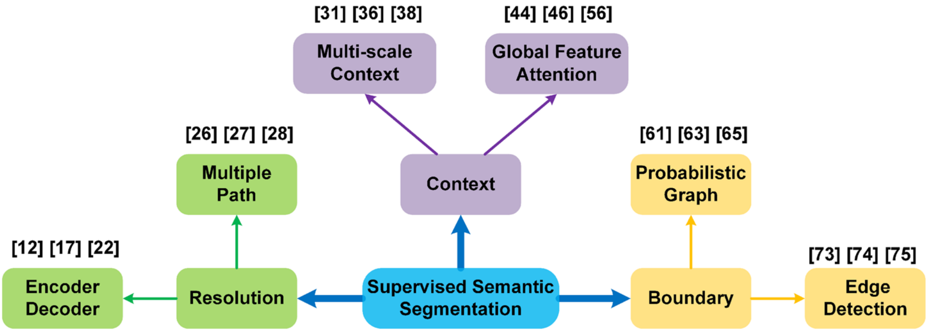

Fig. 1 Classification of supervised semantic segmentation

3.1.1 Encoder-decoder

The structure of Encoder-Decoder is shown in Fig. 2. It contains two parts. The function of the encoding part is to build a good feature extraction network. With the deepening of the network, the size of the feature map gradually decreases and the semantic features continue to increase. The purpose of the decoding part is to establish a feature map restoration model. In the decoding process, the semantic information is expressed and the details can be constantly supplemented.

The complete Encoder-Decoder structure is not achieved overnight. By replacing the last fully connected layer of VGG16 with a convolutional layer, FCN keeps the feature map as a two-dimensional structure, which preserves the spatial information. In addition, this approach also solves the problem that the size of input must be fixed. Moreover, since the nearest neighbor interpolation and bilinear interpolation are not a learnable module, FCN attempts to replace bilinear interpolation with transposed convolution [16]. Although FCN is powerful and flexible, it cannot use global context information due to its inherent spatial invariance. Similar to transposed convolution, unpooling proposed by Noh et al. [17] is equally important. Usually, the max-pooling layer is likely to cause some loss of information. Therefore, it is necessary to keep the max-pooling index in the down-sampling for the information reconstruction. SegNet [18] is one of the most classic Encoder-Decoder structures. It not only effectively combines transposed convolution and unpooling, but also adds batch normalization [19], which effectively solves the problem of gradient disappearance. To reduce computational complexity and increase operating speed, Bayesian-Seg [20] is a very valuable work. In addition, on the basis of it, the new research [21] makes improvements in three aspects. It balances the number of convolutions in each stage and adds auxiliary semantic supervision in encoding part. And it designs a dense neighborhood prediction module combining the information of 21 channels in the output part. Discriminant Feature Network (DFN) [22] holds that the current network is difficult to extract discriminative features. Therefore, border network and smoothing network are designed to deal with the down-sampling and up-sampling features respectively. The output of border network is the bounding box of the object, which is mainly to strengthen the supervision of the semantic boundary in the encoding part. Smooth network mainly uses the channel attention module to

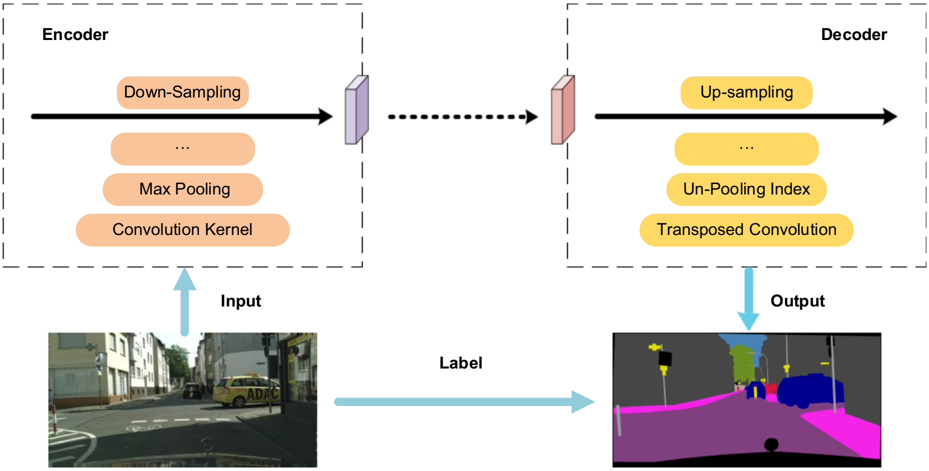

Fig. 2 Encoder-decoder structure

filter low-level information features. The main problem of these methods is still the loss of information caused by down-sampling. The up-sampling operation cannot completely restore the original information. Many approaches handling this issue are introduced in the next section.

3.1.2 Multi-path fusion

Since there is a contradiction between deep semantic and shallow detail information in feature extraction, can the information be processed separately? Therefore, the form of multi-path may be a good method. According to the difference in structure, the multi-path fusion methods can be divided into implicit fusion and explicit fusion.

Implicit fusion has many manifestations in the encoder-decoder methods, such as the skip connection structure of FCN. And the U-Net designed by Ronneberger et al. [23] also densely connects all feature maps in the down-sampling stage. RefineNet [24] uses the multi-resolution fusion method to combine the information with different resolutions. And the chain residual unit is used to get the context information. Global Convolutional Network (GCN) [25] discusses the importance of a fully connected layer for classification. A larger convolution kernel is used to achieve the effect of full connection in the feature fusion stage. This method also avoids the problem that the edge information is less than the central area information when using a small convolution kernel. HRNet proposed by Sun et al. [26] achieves the ultimate in feature fusion. It basically realizes the full connection of convolution kernel nodes. On the one hand, it always maintains the detailed information and spatial structure of the high-resolution image. On the other hand, it realizes a fusion of multi-resolution, which can enhance semantic information.

The form of explicit fusion is embodied as a dual-branch structure, and the basic framework is shown in Fig. 3. The deeper network outputs low-resolution features with semantic information, while the shallower network learns the details and spatial structure information in the high-resolution images, such as the context path and the space path of Bisenet [27]. The context path includes a complete feature extraction network, in which the feature map needs to be down-sampled to 1/32 or less of the original image. The spatial path only needs to obtain 1/8 of the original image. The final segmentation result can be obtained by merging the feature maps extracted by the two paths. Fast-SCNN [28] takes the output of the down-sampling learning module as the input of two paths. The down-sampling learning module includes a convolution and two depth separable convolutions. The advantage is that the parameters are shared in the early stage of the network.

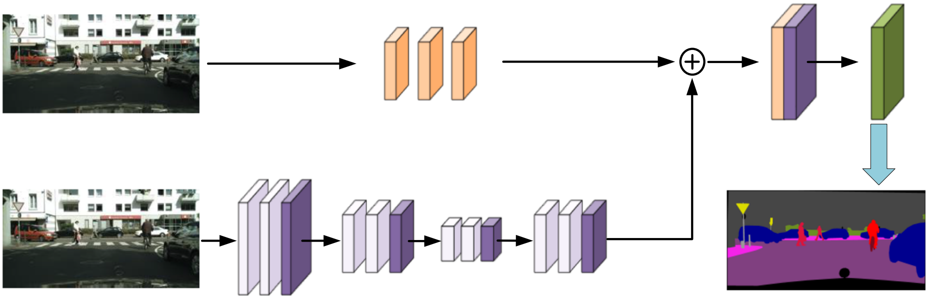

Fig. 3 Dual-path structure

3.2 Semantic enhancement

To extract context, the global receptive field is very important. There are currently two specific methods: Multiple Scale Context and Global feature attention. The receptive field of the convolution kernel is variable in Multiple Scale Context methods. Context can be obtained by fusing different sensory area of kernels. Global feature attention methods is to establish a relationship between a single pixel and the global features. In addition, Liu et al. [29] discuss the lack of global information and propose the global average pool layer.

3.2.1 Multiple scale context

Figure 4 shows several forms to increase the sensory area of kernels. In traditional methods [30], images with different resolutions are input into the network to fuse multiple scale features. By an adaptive pooling layer proposed by Zhao et al. [31], the input can be directly pooled into fixed-size feature maps. By up-sampling and concatenating the multiple scale features, the network can obtain a better segmentation result. However, the filter parameters of this network are fixed after training. It cannot match the multiple scale feature well. To adapt the network to the scale changes of the input better, DMNet [32] designed a parallel dynamic convolution module. It contains a filter generation layer and a dynamic filter layer. The filter generation layer is related to the input, and the convolution kernel is dynamically generated by the adaptive pooling method. Unlike dynamically generating convolution kernel, reference [33] proposes a way of dynamic path selection by dynamic routing. Through the gated network, the path can be adaptively selected according to the input image. The multi-size objects can be allocated to the corresponding resolution levels.

Atrous Convolution [33] introduces the concept of dilation rate on the basic convolution kernel, as shown in Fig. 4. Under the circumstance that the resolution remains unchanged, the atrous convolution kernel can obtain a larger sensory area. There is a lot of works [34–37] exploring the combination of atrous convolution and spatial pyramid pooling structure. It is called “Atrous Spatial Pooling Structure” (ASPP). By setting the different dilation rates, the segmentation requirements of multiple scale targets can be met at the same time. However, the same dilation rate causes some pixels’ information to be lost, which is called the

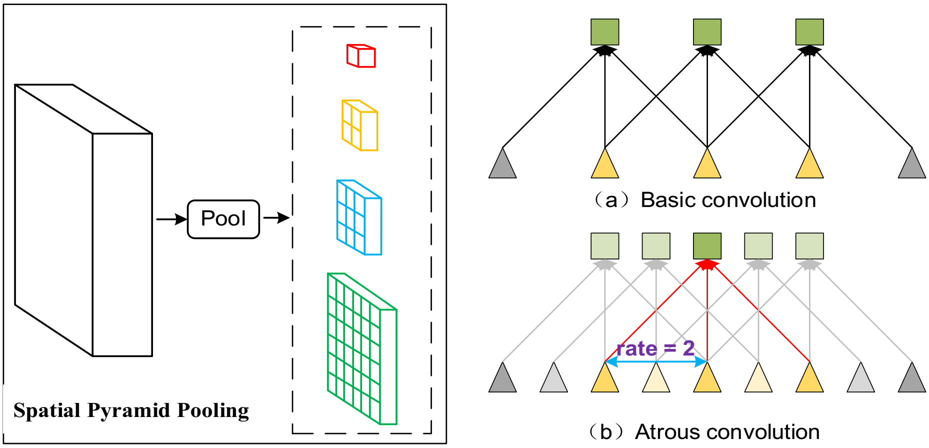

Fig. 4 The forms to increase the sensory area of kernels

“checkerboard effect”. To solve this problem, Hybrid Dilated Convolution (HDC) [38] uses specific dilation rates to form a jagged sampling structure. Deeplab V3+ [39] uses feature fusion at different stages to enrich detailed information. Another problem is that too many atrous convolutions increase computation complexity and memory footprint. Taking Resnet as an example, compared to FCN, 23 residual blocks in DilatedFCN require 4 times more resources. Therefore, Fast-FCN [40] proposes a joint pyramid upsampling module. It concatenates three feature maps before the ASPP structure and reduces the number of channels. This increases the underlying semantic information and reduces computational consumption. Auto-Deeplab [41] retains the ASPP structure of Deeplab. It follows a two-layer hierarchical structure. It not only searches the unit structure for specific calculations operation, but also searches for the network-level architecture to control spatial resolution changes. It is one of the attempts to extend neural architecture search from image classification tasks to dense pixels prediction tasks.

3.2.2 Global feature attention

In 2018, BERT [42] achieved great success in the field of natural language processing by the application of the encoder-decoder architecture based on the attention mechanism [43]. The attention mechanism is a bionic version of the human visual mechanism since human vision receives key areas with high resolution and perceives non-critical areas with low resolution. Different from the method of multi-scale context methods stacking encoder to enlarge the receptive field, it relies more on global modeling. The methods with the global attention in the network are divided into CNN-based methods and Transformer-based methods according to different network structures.

In the CNN-based methods, some works try to capture long-range dependencies through additional self-attention at the channel level [44]. Correspondingly, there are also some applications of the attention mechanism at the spatial and pixel-level [45, 46]. They establish the association of individual pixels with global pixel information through attention weights generated by matrix multiplication of queries and keys. But the matrix multiplication between queries and keys brings super high computational complexity. Some works explore how to solve this kind of problem. For example, Huang et al. [47] only compute the similarity between each pixel and elements from the same row and column. The expectation-maximization (EM) method [48] is used to generate several representative cluster points, and then the similarities are computed between the cluster point and each position. These works are the excellent solutions. Given the independence between channel-level attention and pixel-level attention, there are some works combining the above two in a parallel way [49, 50] or cascade way [51]. In addition, some works combine the attention methods with other blocks (such as context module [52], the designation of multi-scale [53], the object region representation [54]).

In the methods mentioned above, the attention block plays the role of a plug-in. Can computer vision tasks be accomplished independently only by relying on attention? Inspired by the great achievement of transformer application in natural language processing [55], many pioneering works apply the transformer-like structures to computer vision tasks and demonstrate their effectiveness in the past two years. For example, Dosovitskiy et al. [56] propose first vision Transformer that splits the image input into patches, in which the visual tasks are treated as the text tasks as well, as shown in Fig. 5. By learning some advantages of CNN, Liu et al. [57] design the operation of merging patches in Swin Transformer. The number of input

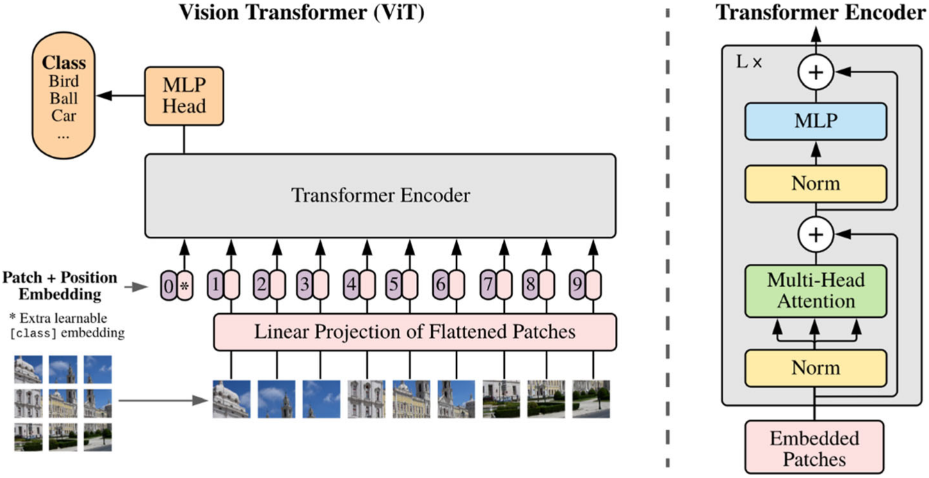

Fig. 5 Vision transformer structure based on the attention

tokens is gradually reduced, and thus the hierarchical features are extracted. To avoid the excessive calculation cost of the transformer model itself, the attention is calculated in a window, and the shift between the windows can cover the global information. When the window size is much smaller than the global size, the computation will drop significantly. And a position bias matrix is also added in calculating attention like the inductive bias in CNN. In addition, some works try to combine the vision transformer and various traditional CNN segmentation frameworks. The general operation is to replace the backbone of the CNN with the transformer and follow the input and output structure of the vision transformer. For example, SETR proposed by Zheng et al. [58] models the global context based on each layer of vision transformer and constitutes a structure of progressive upsampling and multi-level feature aggregation (a variant called SETR-MLA). TransUNet [59] is the first vision transformer for medical image segmentation, which is a hybrid model of UNet and Transformer. Segformer [60] is one of the variants of SETR. It designs the overlap patch and only uses MLP as a more lightweight decoder. In the experiment, Segformer-B5 obtains new SOTA results on ADE20K. Its mIoU is 51.8% and its parameters are 4× smaller than SETR. Furthermore, compared with DeepLabv3, Segformer have stronger robustness when testing multiple types of damage on the cityscape dataset.

3.3 Boundary correction

Due to the intra-class differences and inter-class similarities, there is still much room for improvement in the segmentation of boundary pixels. This part focuses on how to improve the overall segmentation accuracy by boundary correction.

3.3.1 Probabilistic graphical model

In the early machine learning methods, the classification problem is mainly considered from the perspective of pixel location: the label of the pixel is not only related to itself but also the label of the adjacent pixel. The methods using the correlation between pixels mainly include Markov Random Field [3] and Conditional Random Field (CRF) [4]. A potential function of the image binary component is established, which is related to the predicted value and the position of the pixel. Adjacent pixels with similar intensities are usually classified into one category. When combining CRF, MRF with deep learning, it can be used as a separate optimization processing module [61] or an end-to-end training module [62]. The workflow of fully connected CRF is shown in Fig .6. First, the input image A is processed by a Deep Convolutional Neural Network (DCNN) to obtain a rough semantic feature map C. Through a series of up-sampling methods, the feature map C is converted into a rough semantic segmentation result D. the pixels in the rough segmentation result D are modeled and optimized through the fully connected CRF (E). And the final segmentation result F is obtained. Deeplab series network, RefineNet, etc. use DenseCRF [63] method to optimize the final result. Despite the fact that usually fully connected models are inefficient, this model can be efficiently approximated via probabilistic inference. Instead of traditional CRF, GAF [64] adds Gaussian function in the network structure to optimize subsequent segmentation tasks. An end-to-end segmentation network can be formed by iterating a fixed number of Gaussian average fields. The boundary neural field proposed by Bertasius [65] improves the segmentation accuracy only by considering the use of random fields on the boundary. However, multiple iterative calculations are required in the use of probabilistic graph model methods, which leads to huge parameter storage and memory consumption.

To reduce the upsampling effect of the nearest neighbor interpolation and bilinear interpolation on the semantic boundary category, motivated by STN [66], the guided upsampling module [67] leads the sampling to the correct semantic class by generating a guide offset. In the process of upsampling, PointRend [68] separately constructs a small network for “difficult pixels” at the edge, which can improve the pixels classification accuracy by iteration.

3.3.2 Auxiliary edge detection

Since most of the classification errors are concentrated on the boundaries with different semantic categories, it is a good idea to add edge detection methods such as Sobel [69] and Canny [70] as auxiliary judgment. Due to the phenomenon of fewer edge pixels and more non-edge pixels, HED [71] and RCF [72] train a complete end-to-end edge detection network, which is significantly better than Canny and other traditional methods. The workflow of auxiliary edge detection is shown in Fig. 7. The feature map that is simply processed by the backbone is input to two branches separately. The final semantic information labels are used in the loss calculation in the semantic branch. The object contour labels are used for the loss calculation and backpropagation in the boundary branch. In the test, the final prediction map are the element-wise product of semantic branches and boundary branches results.

One branch of Gated-SCNN [73] collects detailed information such as color and texture to extract semantic features, and the other branch only supervises boundary information through gated convolution. The segmentation results can be redefined by fusing the features of the dual branches. Integrating the idea of boundary and direction, STEAL [74] and Segfix [75] directly

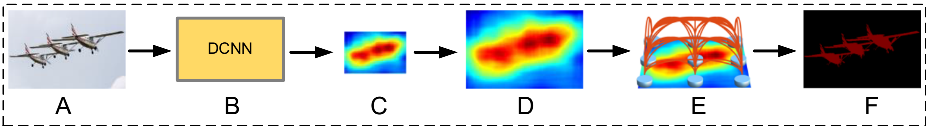

Fig. 6 Fully connected CRF

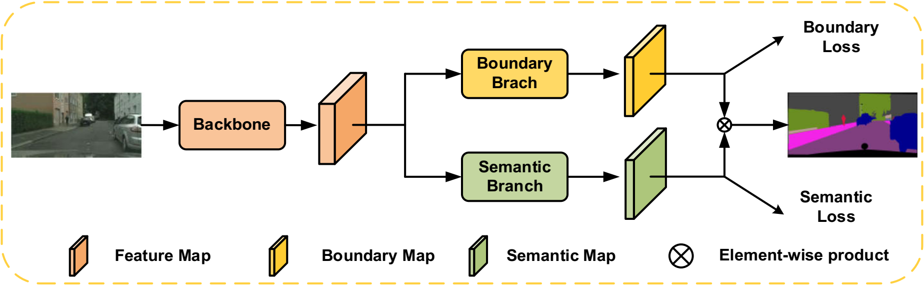

Fig. 7 Auxiliary edge detection

improve the segmentation map in network learning. The edge pixels are firstly predicted by the contour, and then the direction from the boundary pixels to the internal pixels is learned. The corresponding internal pixels are identified by moving a certain distance from the boundary pixels in this direction. The slight difference is that the former first predicts multiple independent boundaries sequentially and then estimates the direction, while the latter uses a parallel branch structure and predicts a complete boundary map and direction map.

4 Discussion

In the previous section, we study various methods from a qualitative point, but lack of quantitative evaluation. Although many authors provide enough experimental details to introduce their work, it is still difficult to compare various methods due to the differences in the datasets and metrics. For a fair quantitative evaluation, in this section we describe the popular evaluation metrics and the most representative datasets. Next, according to the classification points mentioned in the previous section, we collect the results of a series of methods on several representative datasets. At last, we summarize and draw conclusions about these results.

4.1 Evaluation metrics

Semantic segmentation is essentially a classification task. Based on the combination of true category and predicted category, examples in the binary classification task can be divided into four situations: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). The “confusion matrix” of the classification results is shown in Table 1. For simple two-category evaluation metrics, the evaluation metrics mainly include Precision (P) and Recall (R). For more complex multi-category evaluation metrics, it mainly includes pixel accuracy, mean pixel accuracy, intersection over union, mean intersection over union, frequency weighted intersection over union [76].

(1) Precision (P) is the proportion of the positive example in the prediction result, as shown in eq. (1).

$$P = \frac{TP}{(TP + FP)} \tag{1}$$

Table 1 Confusion matrix

true category	predicted category	
	True	False
True	TP	FN
False	FP	TN

(2) Recall (R) indicates the proportion of examples correctly predicted in all true examples, as shown in eq. (2).

$$R = \frac{TP}{(TP + FN)} \tag{2}$$

(3) Pixel Accuracy (PA) indicates the proportion between the number of pixels with the correct prediction category to the total number of pixels, as shown in eq. (3). N represents the number of pixels needed to be classified; T_i represents the total number of pixels of the ith category; X_{ji} represents the number of pixels whose actual type is category i and the predicted type is category j.

$$PA = \left(\sum_{i=1}^{N} X_{ii} \right) / \left(\sum_{i=1}^{N} T_i \right) \tag{3}$$

(4) Mean Pixel Accuracy (MPA): Indicates the proportion of correctly classified pixels in each class and find the average of all classes, as shown in eq. (4).

$$MPA = \left(\sum_{i=1}^{N} (X_{ii}/T_i) \right) / N \tag{4}$$

(5) Intersection over Union (IoU) is the proportion between the intersection and union of the predicted result of a certain category and the true value.

(6) Mean Intersection over Union (MIoU) indicates the result of averaging the intersection of all categories, as shown in eq. (5).

$$MIoU = \frac{\left(\sum_{i=1}^{N} \frac{X_{ii}}{T_i + \sum_{j=1}^{\mathrm{N}} \left(X_{ji} - X_{ii} \right)} \right)}{N} \tag{5}$$

(7) Frequency Weighted Intersection over Union (FWIoU) is the weighted intersection ratio of all categories, as shown in eq. (6). The weight value is set according to the frequency of each class.

$$FWIoU = \frac{\sum_{i=1}^{N} \frac{X_{ii}}{T_i + \sum_{j=1}^{N} \left(X_{ji} - X_{ii}\right)}}{\sum_{i=1}^{N} \sum_{j=1}^{N} X_{ij}} \tag{6}$$

4.2 Open datasets

Over the past years, open datasets related to deep learning semantic segmentation have emerged endlessly. We collect a series of representative test datasets. These datasets are roughly divided into autonomous driving datasets and scene analysis datasets driverless. Autonomous driving datasets include Cam Vid [77], Cityscapes [78] and KITTI [79]. And scene parsing datasets include Pascal VOC 2012 [80], Pascal Context [81], ADE20K [82]. Specific information of each dataset is shown in Table 2.

Cam Vid is the first road and driving scene parsing dataset proposed in 2009. The dataset comes from 5 video sequences of the same camera, contains 32 types of objects and 701 data samples. The original data can be divided into training set, validation set and test set according to the ratio of 7:2:1. The size of images is 960 × 720.

Focusing on image segmentation in an autonomous driving environment, cityscapes is an urban street view dataset released by Mercedes-Benz in 2015. In addition to 20,000 rough annotation labels, the datasets have a total of 5000 high-quality pixel-level label annotations. It is divided into 2975 training samples, 500 validation samples, and 1525 test samples. These samples cover 19 types of segmentation objects, and it includes ground, buildings, traffic signs in different scenes and seasons of 18 cities. The size of images is 2048 × 1024.

Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) is the largest evaluation dataset in autonomous driving scenarios. It covers real image data from various scenarios such as cities, villages, and highways. Especially there are up to 30 pedestrians and 15 vehicles in a single image. And there is a certain degree of occlusion.

The Pascal Visual Object Classes (Pascal VOC) dataset has been updated from 2005 to 2012. As of 2012, VOC 2012 dataset contains 1464 training samples, 1449 validation samples and 2913 test samples. The most universally used dataset is Pascal VOC2012 Aug. It is a collection of PASCAL VOC 2012 and SBD [83]. This dataset contains 10,582 training samples. The number of validation and test samples remains unchanged. Images vary in size. Each image contains an average of 1.4 categories and 2.3 instance targets. It involves 21 types of objects, including humans, animals, vehicles, et al.

The Pascal Context dataset defines 459 categories. It uses 4998 training samples and 5105 validation samples of VOC 2010 as data sources. Generally, the most common 59 categories are used as semantic labels, and the remaining categories are set as background.

ADE20K is a dataset launched by the Massachusetts Institute of Technology in 2017. It contains 20,210 training samples and 2000 validation samples. It covers 150 different segmentation object types in various complex scenes such as indoor and outdoor.

Table 2 Common datasets for semantic segmentation

Datasets	Categories	Training	Validation	Total	Test	Size
Cam Vid [77]	32	491	140	631	70	960 × 720
Cityscapes [78]	19	2975	1525	4500	500	2048 × 1024
KITTI [79]	10	140	–	140	112	1226 × 370
VOC2012 [80]	21	1464	1449	2913	1452	Not fixed
Pascal Context [81]	59	4998	5105	10,103	9637	Not fixed
ADE20K [82]	150	20,210	2000	22,210	–	Not fixed
SBD [83]	8	8498	2857	11,355	–	320 × 210
VOC 2012 Aug [80]	21	10,582	1449	12,031	1452	Not fixed

Compared with VOC 2012 and cityscapes, ADE20K and Pascal Context have more complex segmentation scenes, more semantic categories, and instance objects. Thus, the segmentation task on ADE20K and Pascal Context is more difficult. In addition to the above data sets, there are a several of datasets [84, 85] in the medical field, and a few of datasets [86, 87] in the remote sensing field.

4.3 Quantitative results

As we mentioned above, we have studied various methods qualitatively in the section 3. This section will evaluate the performance of each method quantitatively. According to the above classification perspectives, these results are organized into three parts: Super-resolution reconstruction in segmentation, Semantic enhancement and Boundary correction.

Since the influences from datasets, backbone, batch, output size, iterations and learning rate et al. are still incalculable, no uniform platform can be used to compare these methods. Therefore, we try our best to find a fair platform to truly reflect the difference of the core methods.

4.3.1 Super-resolution reconstruction in segmentation

It is of great significance to compare different methods on the same dataset. In this section, we select two datasets: Pascal VOC2012 and Cityscapes.

Pascal VOC2012 is one of the most popular datasets. In the early stages of segmentation task, the vast majority of methods are evaluated in this dataset. Table 3 shows the results of super-resolution reconstruction methods on this dataset. This set of results shows an obvious improvement trend from the original methods (SegNet and FCN) to the complex networks such as the winner ExFuse with 87.9% mIoU. In particular, at the stage of the mIoU from 62.5% to 82.2%, although some excellent networks can not be displayed here due to lack of data, it can still be seen that the improvement is huge. The change of backbone from VGG to ResNet make a huge contribution to the improvement of model performance.

Cityscape is the more challenging autonomous driving dataset. Table 4 shows the results of super-resolution reconstruction segmentation methods on this dataset. The improvement trend on this dataset is relatively gentle due to the gradual optimization of backbone. The result using HRNet is the top scorer with an 81.6% mIoU on this dataset.

Table 3 Results of super-resolution reconstruction methods on Pascal VOC-2012

Name	Backbone	Contribution(s)	mIoU (%)
ExFuse [21]	ResNext131	SS, DNP	87.9
DFN [22]	ResNet101	Border Network, Smooth Network	86.2
RefineNet [24]	ResNet101	RCU, MRF, CRU	84.2
GCN [25]	ResNet101	Larger convolution kernel	82.2
DeconvNet [17]	VGG-16	Full connection, Un-pooling	62.5
FCN [12]	VGG-16	Full convolution, Skip connection	62.2
Bayesian Seg [20]	VGG-16	Un-Pooling index, Dropout	60.5
SegNet [18]	VGG-16	Un-Pooling index	59.1

Although model performance has been improved, it is accompanied by an increasing convolutional layer number and an expanding memory footprint. This is a very noteworthy issue. However, many authors avoid discussing this issue. Blindly pursuing accuracy without considering resource consumption, this kind of work cannot be applied to reality and it does not have much research significance.

4.3.2 Semantic enhancement

In this section, in addition to the two datasets used above, we add two more challenging datasets for evaluation: ADE20K and Pascal Context.

Table 5 shows the mIoU of semantic enhancement methods on the two datasets. Compared with the methods belonging to the first classification, the overall results on the two datasets are higher. This shows that semantic enhancement is the core of semantic segmentation. In this case, although the difference in results between the methods is very small, it provides a variety of options for enhancing semantic.

In addition, there is an interesting phenomenon about datasets. The early methods prefer Pascal VOC2012 for testing, while the later methods use Cityscape more. Given that the mIoU of Pascal VOC2012 has reached about 88%, while Cityscape's mIoU is 84%. It has more challenges in the Cityscape dataset.

From Table 5, it can be seen that the mIoU of most methods can reach more than 80%, which basically meets dataset segmentation requirements. To adapt more complex environments, many existing methods perform the experiments on more challenging datasets. Table 6 shows the mIoU of semantic enhancement methods on ADE20K and Pascal Context. The mIoU of each method is about 50% on ADE20K and Pascal Context. The segmentation results

Table 4 Results of super-resolution reconstruction methods on Cityscapes

Name	Backbone	Contribution(s)	mIoU (%)
HRNet [26]	HRNetV2	Fusion from high to low resolution	81.6
DFN [22]	ResNet101	Border Network, Smooth Network	79.3
GCN [25]	ResNet152	Larger convolution kernel	76.9
RefineNet [24]	ResNet101	RCU, MRF, CRU	73.6
BiSeNet [27]	Xception39	Dual-path architecture	71.4
Fast-SCNN [28]	Xception39	Weight share, Deep separable Conv	68.0
FCN [12]	VGG-16	Full convolution, Skip connection	65.3

Table 5 The mIoU of semantic enhancement methods (%)

Name	Core methods	Pascal VOC2012	Cityscape
OCRNet [54]	Object context representation	–	83.6
SegFormer [60]	Overlap patch, MLP		83.1
DNLNet [46]	Disentangled Non-Local	–	83.0
CCNet [47]	Positional attention	–	81.9
OCNet [50]	Channel, Positional attention	–	81.7
SETR [58]	Transformer, MLA`		81.6
DANet [49]	Dual branch architecture	–	81.5
EMANet [48]	EM, Low rank reconstruction	87.7	81.2
Deeplab V3+ [38]	ASPP, Upsampling	87.8	82.1
EncNet [52]	Channel attention, Atrous conv	85.9	79.5
Deeplab V3 [36]	HDC, ASPP	85.7	81.3
Auto-Deeplab [41]	Neural architecture search	85.6	82.1
PSPNet [31]	Adaptive pooling	85.4	78.4
DMNet [32]	Dynamic convolution modules	84.4	–
Deeplab V2 [35]	ASPP	79.7	70.4
Large FOV [34]	Atrous conv, CRF	72.7	–
ParseNet [29]	Global average pooling	69.8	–

of complex semantic segmentation datasets are not ideal. There is still a long way to go in the field of semantic segmentation.

4.3.3 Boundary correction

The boundary correction method should receive more attention due to its unique two-branch structure. Its backbone is not a traditional network such as VGG and ResNet, but a mature semantic segmentation network such as Deeplab V3. Therefore, it is not correct to simply look at the final result, but to look at the improvement compared with the backbone. Table 7 shows the results of boundary correction methods. Based on the original network, whether it is Cityscape or ADE20K, the segmentation results can be greatly improved. The most important thing is that the dual-branch network creates a unique idea, which integrates different methods to deal with common problems.

Table 6 The mIoU of semantic enhancement methods on complex datasets (%)

Name	ADE20K	Pascal Context
Swin Transformer [57]	53.5	–
SegFormer [60]	51.8	–
SETR [58]	50.3	–
OCRNet [51]	45.3	54.8
DMNet [32]	45.5	54.4
DANet [52]	45.2	52.6
EncNet [49]	44.7	51.7
Fast-FCN [40]	44.3	53.1
PSPNet [31]	43.3	47.8

Table 7 Results of boundary correction methods

Name	backbone	Datasets	mIoU (%)
Gated-SCNN [73]	–	Cityscapes	82.8
PointRend [68]	Deeplab V3	Cityscapes	78.4 (+1.2)
STEAL [74]	CASENet	Cityscapes	71.3 (+2.4)
GUN [67]	–	Cityscapes	70.4
Segfix [75]	Deeplab V3	ADE20K	45.4 (+1.3)

4.4 Conclusion

Recently, many semantic segmentation methods based on fully supervised learning are leading the way in many computers vision. To familiarize readers with the development of relevant research in the semantic segmentation, this paper provides a broad and organized survey of existing methods. We discuss the key issues affecting the segmentation effects and explore the inner connection between past research and ongoing research. We provide a brief summary of aforementioned methods.

(1) The main goal of the super-resolution reconstruction method in segmentation is to restore the location information of the pixels and make predictions. The optimization of the structure focuses on the up-sampling part and the fusion part between up-sampling and down-sampling.

(2) The main goal of semantic enhancement methods is to extract richer features. It includes the extraction of context and the addition of global features. The improvement of the structure mainly focuses on the down-sampling part.

(3) The main goal of the boundary correction method is to refine the edge and improve the segmentation accuracy. It is a useful post-processing module for any deep learning based semantic segmentation architecture. Its manifestation in the network structure can be either at the end of the network or as a new parallel branch.

Based on the above discussion, we found that three kinds of methods do not affect each other. The down-sampling part, up-sampling part, fusion part, and a new parallel branch do not overlap each other in structure. Therefore, it is feasible to combine existing methods to enhance semantic segmentation performance. It is of great significance to establish such a complete semantic segmentation system.

5 Future research directions

At the current stage, fully-supervised semantic segmentation is basically mature. Based on this survey of current methods, we present a series of challenging directions.

5.1 Real-time semantic segmentation

To enhance its application on mobile devices, semantic segmentation networks not only require higher segmentation accuracy, but also require fewer parameters and memory consumption. Although this paper does not classify separately for real-time segmentation methods, SegNet, Bayesian SegNet, BiSeNet, Fast-SCNN, and GUN are all real-time segmentation networks. Under the circumstance of ensuring accuracy, “how to further increase the operating speed and design a more portable network” is worthy of further study. This will speed up the process of applying semantic segmentation to practice, such as autonomous driving field.

5.2 Complex scene parsing

From the model itself, current semantic segmentation methods have high segmentation accuracy for one category in a single image. However, there are still great difficulties when more semantic categories and instance categories are involved. On one hand, it is still insufficient to capture the context and semantic information. On the other hand, the target size becomes smaller and information is easily lost when the number of categories increases. The network lacks the ability to discriminate small targets. Therefore, the maintenance of high resolution and the acquisition of deep semantic information are still the primary problems of semantic segmentation. Although HRNet has made some attempts, this problem still needs further solutions according to its results on ADE20K.

Acknowledgments We convey our sincere appreciation to Qingdao University of Technological and Machine Vision Laboratory. We also wish to acknowledge the younger martial brothers and sisters who provide help but do not appear in the ranks of the authors.

Code availability Not applicable.

Author contributions Conceptualization: Zhou YG; Methodology: Ren YB, Zhou YG; Formal analysis and investigation: Zhou LJ; Writing - original draft preparation: Ren YB; Writing - review and editing: Zhou YG, Xu EY, Liu SL; Funding acquisition: Zhou YG; Resources: Zhou YG; Supervision: Zhou YG. All authors have read and agreed to the published version of the manuscript.

Funding This work was supported by a National Natural Science Foundation (Grant No. 62171247) and two Natural Science Foundation of Shandong Province (Grant No. ZR2020MF001 and Grant No. ZR2020QF101).

Data availability The data presented in this study are available on request from the corresponding author.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Mardia KV, Hainsworth TJ (1988) A spatial thresholding method for image segmentation. IEEE Transactions on Pattern and Machine Intelligence 10(6):919–927
2. Shotton J, Johnson M, Cipolla R (2008) Semantic texton forests for image categorization and segmentation. in Proceedings of 26th IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2008.4587503.
3. Li SZ (1994) Markov random field models in computer vision. In proceedings of computer vision—ECCV 1994 - 3rd European conference on computer vision, pp. 361-370. https://doi.org/10.1007/bfb0028368.

4. Lafferty JD, Mccallum A, Pereira FCN (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In proceedings of the eighteenth international conference on machine learning, pp. 282-289.
5. Adams R, Bishof L (1994) Seeded region growing. IEEE Trans Pattern Anal Mach Intell 16(6):641–647. https://doi.org/10.1109/34.295913
6. Lakshmi S, Sankaranarayanan DV (2010) A study of edge detection techniques for segmentation computing approaches. International Journal of Computer Applications 1:35–41
7. Liu ST, Yin FL (2012) The basic principle and its new advances of image segmentation methods based on graph cuts. Acta Automat Sin 38(6):911–922. https://doi.org/10.3724/SP.J.1004.2012.00911
8. Simonyan K, Zisserman (2015) Very deep convolutional networks for large-scale image recognition. in Proceedings of 3rd International Conference on Learning Representations.
9. Krizhevsky A, Sutskever I, Hinton G (2012) ImageNet classification with deep convolutional neural networks. in Proceedings of 26th Annual Conference on Neural Information Processing Systems 2:1097–1105
10. Szegedy C, Liu W, Jia Y, et al. (2015) Going deeper with convolutions. In proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 1-9. https://doi.org/10.1109/CVPR.2015.7298594.
11. He K, Zhang X, Ren S, et al. (2016) Deep residual learning for image recognition. In proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90.
12. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In proceedings of IEEE conference on computer vision and pattern recognition, pp. 3431-3440. https://doi.org/10.1109/CVPR.2015.7298965.
13. Garcia-Garcia A, Orts-Escolano S, Oprea S, et al. (2017) A review on deep learning techniques applied to semantic segmentation. arXiv preprint arXiv:1704.06857.
14. Ghosh S, Das N, Das I, et al. (2019) Understanding deep learning techniques for image segmentation. ACM Computing Surveys, vol. 52, no. 4, pp. 40. https://doi.org/10.1145/3329784.
15. Guo Y, Liu Y, Georgiou T, Lew MS (2018) A review of semantic segmentation using deep neural networks. International journal of multimedia information retrieval 7(2):87–93
16. Dumoulin V, Visin F. (2016). A guide to convolution arithmetic for deep learning. arXiv preprint arXiv:1603.07285. https://arxiv.org/abs/1603.07285.
17. Noh H, Hong S, Han B (2015) Learning deconvolution network for semantic segmentation. In proceedings of the IEEE international conference on computer vision, pp. 1520-1528. https://doi.org/10.1109/ICCV.2015.178.
18. Badrinarayanan V, Kendall A, Cipolla R (2017) SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 39(12):2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615
19. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In proceedings of 32nd international conference on machine learning, pp. 448-456.
20. Kendall A, Badrinarayanan V, Cipolla R (2017) Bayesian segnet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. in Proceedings of British Machine Vision Conference. https://doi.org/10.5244/c.31.57.
21. Zhang Z, Zhang X, Peng C, et al. (2018) ExFuse: enhancing feature fusion for semantic segmentation. in Proceedings of Computer Vision – ECCV 2018 - 15th European Conference, vol. 10, pp. 273–288. https://doi.org/10.1007/978-3-030-01249-6_17.
22. Yu C, Wang J, Peng C, et al. (2018) Learning a discriminative feature network for semantic segmentation. In proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp. 1857-1866. https://doi.org/10.1109/CVPR.2018.00199.
23. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In proceedings of medical image computing and computer-assisted intervention – MICCAI 2015 - 18th international conference, pp. 234-241.https://doi.org/10.1007/978-3-319-24574-4_28.
24. Lin G, Milan A, Shen C, et al. (2017) RefineNet: multi-path refinement networks for high-resolution semantic segmentation. In proceedings of 30th IEEE conference on computer vision and pattern recognition, pp. 5168-5177. https://doi.org/10.1109/CVPR.2017.549.
25. Peng C, Zhang X, Yu G, et al. (2017) Large kernel matters — improve semantic segmentation by global convolutional network. In proceedings of 30th IEEE conference on computer vision and pattern recognition, pp. 1743-1751. https://doi.org/10.1109/CVPR.2017.189.
26. Sun K, Xiao B, Liu D, et al. (2019) Deep high-resolution representation learning for human pose estimation. In proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp. 5686-5696. https://doi.org/10.1109/CVPR.2019.00584.

27. Yu C, Wang J, Peng C et al (2018) BiSeNet: bilateral segmentation network for real-time semantic segmentation. in Proceedings of Computer Vision – ECCV 2018 - 15th European Conference 13:334–349. https://doi.org/10.1007/978-3-030-01261-8_20
28. Poudel RPK, Liwicki S, Cipolla R (2019) Fast-SCNN: fast semantic segmentation network. in Proceedings of 30th British Machine Vision Conference 2019.
29. Liu W, Rabinovich A, Berg AC (2015) ParseNet: Looking Wider to See Better. arXiv preprint arXiv:1506.04579.
30. Farabet C, Couprie C, Najman L, LeCun Y (2013) Learning hierarchical features for scene labeling. Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence 35(8):1915–1929. https://doi.org/10.1109/TPAMI.2012.231
31. Zhao H, Shi J, Qi X, et al. (2017) Pyramid scene parsing network. In proceedings of 30th IEEE conference on computer vision and pattern recognition, pp. 6230-6239. https://doi.org/10.1109/CVPR.2017.660.
32. He J, Deng Z, Qiao Y (2019) Dynamic multi-scale filters for semantic segmentation. In proceedings of the IEEE international conference on computer vision, pp. 3561-3571. 10.11-09/ICCV.2019.00366.
33. Li Y, Song L, Chen Y, et al. (2020) Learning dynamic routing for semantic segmentation. In proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp. 8550-8559. https://doi.org/10.1109/CVPR42600.2020.00858.
34. Yu F, Koltun V (2016) Multi-scale context aggregation by dilated convolutions. in Proceedings of 4th International Conference on Learning Representations.
35. Chen L, Papandreou G, Kokkinos I, et al. (2015) Semantic image segmentation with deep convolutional nets and fully connected CRFs. in Proceedings of 3rd International Conference on Learning Representations.
36. Chen L, Papandreou G, Kokkinos I et al (2018) DeepLab: semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848. https://doi.org/10.1109/TPAMI.2017.2699184
37. Chen L, Papandreou G, Schroff F, et al. (2017) Rethinking Atrous Convolution for Semantic Image Segmentation arXiv preprint arXiv: 1706.05587.
38. Wang P, Chen L, Yuan Y, et al. (2018) Understanding convolution for semantic segmentation. In proceedings of 2018 IEEE winter conference on applications of computer vision, pp. 1451-1460. https://doi.org/10.1109/WACV.2018.00163.
39. Chen LC, Zhu Y, Papandreou G et al (2018) Encoder-decoder with Atrous separable convolution for semantic image segmentation. in Proceedings of Computer Vision – ECCV 2018 - 15th European Conference 7:833–851. https://doi.org/10.1007/978-3-030-01234-2_49
40. Wu H, Zhang J, Huang K, et al. (2019) FastFCN: rethinking dilated convolution in the backbone for semantic segmentation. arXiv preprint arXiv:1903.11816.
41. Liu C, Chen L C, Schroff F, et al. (2019) Auto-deeplab: hierarchical neural architecture search for semantic image segmentation. In proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp. 82-92. https://doi.org/10.1109/CVPR.2019.00017.
42. Devlin J, Chang MW, Lee K, et al. (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. in Proceedings of NAACL HLT 2019–2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 4171–4186.
43. Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. In proceedings of advances in neural information processing systems 30, pp 5999-6009.
44. Hu J, Shen L, Albanie S, Sun G, Wu E (2020) Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell 42(8):2011–2023. https://doi.org/10.1109/TP-AMI.2019.2913372
45. Wang X, Girshick R, Gupta A, et al. (2018) Non-local neural networks. In proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp. 7794-7803. https://doi.org/10.1109/CVPR.2018.00813.
46. Yin M, Yao Z, Cao Y et al (2020) Disentangled non-local neural networks. In. Proceedings of Computer Vision – ECCV 2020 - 16th European Conference 12360(15):191–207. https://doi.org/10.1007/978-3-030-58555-6_12
47. Huang Z, Wang X, Huang L, et al. (2019) CCNet: Criss-cross attention for semantic segmentation. In proceedings of the IEEE international conference on computer vision, pp 603-612. https://doi.org/10.1109/ICCV.2019.00069.
48. Li X, Zhong Z, Wu J (2019) EMANet: expectation-maximization attention networks for semantic segmentation. In proceedings of the IEEE international conference on computer vision, pp 9166-9175. https://doi.org/10.1109/ICCV.2019.00926.
49. Fu J, Liu J, Tian H, et al. (2019) Dual attention network for scene segmentation. In proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 3141-3149. https://doi.org/10.1109/CVPR.2019.00326.

50. Yuan Y, Wang J (2018) OCNet: object context network for scene parsing. arXiv preprint arXiv: 1809.00916.
51. Cao Y, Xu J, Lin S, et al. (2019) GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. in Proceedings of 2019 International Conference on Computer Vision Workshop, pp. 1971-1980. https://doi.org/10.1109/ICCVW.2019.00246.
52. Zhang H, Dana K, Shi J, et al. (2018) Context encoding for semantic segmentation. In proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp. 7151-7160. https://doi.org/10.1109/CVPR.2018.00747.
53. Andrew T, Karan S, Bryan C (2020) Hierarchical multi-scale attention for semantic segmentation. arXiv preprint arXiv:2005.10821.
54. Yuan Y, Chen X, Wang J (2020) Object-contextual representations for semantic segmentation. in Proceedings of Computer Vision – ECCV 2020 - 16th European Conference 6:173–190. https://doi.org/10.1007/978-3-030-58539-6_11
55. Galassi A, Lippi M, Torroni P (2020) Attention in natural language processing. IEEE Transactions on Neural Networks and Learning Systems 32:4291–4308. https://doi.org/10.1109/T-NNLS.2020.3019893
56. Dosovitskiy A, Beyer L, Kolesnikov A, et al. (2021) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
57. Liu Z, Lin YT, Cao Y, et al. (2021) Swin transformer: hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030.
58. Zheng S, Lu J, Zhao H, et al.(2021) Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp. 6881–6890.
59. Chen J, Lu Y, Yu Q, et al. (2021) Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.
60. Xie E, Wang W, Yu Z, et al. (2021) Segformer: simple and efficient design for semantic segmentation with transformers. arXiv preprint arXiv:2105.15203v3.
61. Liu Z, Li X, Luo P, et al. (2015) Semantic image segmentation via deep parsing network. In proceedings of the IEEE international conference on computer vision, pp. 1377-1385. https://doi.org/10.1109/ICCV.2015.162.
62. Liu S, De MS, Gu J, et al. (2017) Learning affinity via spatial propagation networks. In proceedings of advances in neural information processing systems, pp. 1521-1531.
63. Krähenbühl P, Koltun V (2011) Efficient inference in fully connected crfs with Gaussian edge potentials. in Proceedings of Advances in Neural Information Processing Systems 24.
64. Vemulapalli R, Tuzel O, Liu M, et al. (2016) Gaussian conditional random field network for semantic segmentation. In proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 3224-3233. https://doi.org/10.1109/CVPR.2016.351.
65. Bertasius G, Shi J, Torresani L (2016) Semantic segmentation with boundary neural fields. In proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp. 3602-3610. https://doi.org/10.1109/CVPR.2016.392.
66. Jaderberg M, Simonyan K, Zisserman A, et al. (2015) Spatial transformer networks. In proceedings of advances in neural information processing systems, pp. 2017-2025.
67. Mazzini D (2018) Guided upsampling network for real-time semantic segmentation. in Proceedings of British Machine Vision Conference 2018.
68. Kirillov A, Wu Y, He K, et al. (2020) PointRend: image segmentation as rendering. In proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp. 9796-9805. https://doi.org/10.1109/CVPR42600.2020.00982.
69. Kittler J (1983) On the accuracy of the Sobel edge detector. Image Vis Comput 1(1):37–42. https://doi.org/10.1016/0262-8856(83)90006-9
70. Canny JF (1986) A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell 8(6): 679–698. https://doi.org/10.1109/T-PAMI.1986.4767851
71. Xie S, Tu Z (2017) Holistically-nested edge detection. Int J Comput Vis 125(3):3–18. https://doi.org/10.1007/s11263-017-1004-z
72. Liu Y, Cheng M, Hu X, Bian JW, Zhang L, Bai X, Tang J (2017) Richer convolutional features for edge detection. IEEE Trans Pattern Anal Mach Intell 41(8):1939–1946. https://doi.org/10.1109/TPAMI.2018.2878849
73. Wang Z, Acuna D, Ling H, et al. (2019) Object instance annotation with deep extreme level set evolution. In proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp. 7492-7500. https://doi.org/10.1109/CVPR.2019.00768.

74. Acuna D, Kar A, Fidler S (2019) Devil is in the edges: learning semantic boundaries from noisy annotations. In proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp. 11067-11075. https://doi.org/10.1109/CVPR.2019.01133.
75. Yuan Y, Xie J, Chen X, et al. (2020) SegFix: model-agnostic boundary refinement for segmentation. In proceedings of computer vision – ECCV 2020 - 16th European conference, pp. 489-506. https://doi.org/10.1007/978-3-030-58610-2_29.
76. Shao J, Huang X, Cao K (2019) A review on deep learning techniques applied to semantic segmentation. Dianzi Keji Daxue Xuebao/Journal of the University of Electronic Science and Technology of China 48(5): 644–654. https://doi.org/10.3969/j.issn.1001-0548.2019.05.001
77. Brostow G, Fauqueur J, Cipolla R (2009) Semantic object classes in video: A high-definition ground truth database. Pattern Recognition Letters, vol. 30, no. 2, pp. 88–97. 10.101–6/j.patrec.2008.04.005.
78. Cordts M, Omran M, Ramos S, et al. (2016) The cityscapes dataset for semantic urban scene understanding. In proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp. 3213-3223. https://doi.org/10.1109/CVPR.2016.350.
79. Geiger A, Lenz P, Stiller C, Urtasun R (2013) Vision meets robotics: the KITTI dataset. Int J Robot Res 32(11):1231–1237. https://doi.org/10.1177/0278364913-491297
80. Suyash S (2016) Application of convolutional neural network for image classification on Pascal VOC challenge 2012 dataset. arXiv preprint arXiv:1607.03785.
81. Mottaghi R, Chen X, Liu X, et al. (2014) The role of context for object detection and semantic segmentation in the wild. In proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 891-898. https://doi.org/10.1109/CVPR.2014.119.
82. Zhou B, Zhao H, Puig X, et al. (2017) Scene parsing through ADE20K dataset. In proceedings of 30th IEEE conference on computer vision and pattern recognition, pp. 5122-5130. https://doi.org/10.1109/CVPR.2017.544.
83. Hariharan B, Arbelaez P, Bourdev L, et al. (2011) Semantic contours from inverse detectors. In proceedings of the IEEE international conference on computer vision, pp. 991-998. https://doi.org/10.1109/ICCV.2011.6126343.
84. Staal J, Abramoff M, Niemeijer M et al (2004) Ridge-based vessel segmentation in color images of the retina. IEEE Trans Med Imaging 23(4):501–509. https://doi.org/10.1109/TMI.2004.825627
85. Menze B, Jakab A, Bauer S et al (2015) The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging 34(10):1993–2024. https://doi.org/10.1109/TMI.2014.2377694
86. Paisitkriangkrai S, Sherrah J, Janney P, van den Hengel A (2016) Semantic labeling of aerial and satellite imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 9(7):2868–2881. https://doi.org/10.1109/JSTARS.2016.2582921
87. Maggiori E, Tarabalka Y, Charpiat G, et al. (2017) Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark. International Geoscience and Remote Sensing Symposium (IGARSS), pp. 3226–3229. https://doi.org/10.1109/IGARSS.2017.8127684.