The Application of Semantic Segmentation on 2D images

Yutong Li*

School of Statistics and Data Science, Naikai University, Tianjin, China

*Corresponding author: 1910071@mail.nankai.edu.cn

Abstract. A fundamental problem in computer vision is semantic segmentation, which calls for the algorithm to categorize each pixel in the picture and provide the precise details of the category. Semantic segmentation is being employed extensively in a variety of applications, including autonomous vehicles and medical imaging. An overview of similar semantic segmentation approaches is given in this study. First, this paper gives a brief overview of the history and vocabulary of semantic segmentation. The key datasets for semantic segmentation, conventional segmentation models, and fundamental deep learning techniques for semantic segmentation will then be covered. In particular, traditional methods centered on context models and deep learning-centered methods are discussed in detail. Finally, we review several assessment techniques, including their benefits and drawbacks, and outline the key issues facing semantic segmentation today. In addition, this study seeks to provide an overview of the relevant literature and the difficulties in semantic segmentation. Finally, the paper summarizes the semantic segmentation and prospects the future.

Keywords: Semantic segmentation; Deep learning; Conditional Random Field; 2D images.

1. Introduction

With the advancement and development of computer vision, we need computers to have a deeper and more accurate understanding of images. Semantic segmentation tasks classify every pixel in an image, and clustering pixels belonging to the same object and class in the image. The difficulty of the semantic segmentation task is that there are many object categories in the image, some of which have high similarity and are difficult to identify, which may lead to errors in annotation. Secondly, the stability of semantic segmentation under lighting, smoke, occlusion and other environment is also a difficulty. Since the results of semantic segmentation can provide semantic information at the pixel level, semantic segmentation tasks are starting to become increasingly important, for example, autonomous driving and medical diagnosis [1].

When the concept of semantic segmentation was proposed, the meaning of the term was similar to image segmentation at the beginning. But unlike image segmentation, semantic segmentation places more emphasis on the semantic information of the segmented regions. Object segmentation task proposed in the 1990s further distinguished all classes of semantic objects from the background, resulting in the inability to segment multiple objects of different classes in an image. One solution to this problem is to use the constrained parametric min cuts (CPMC) algorithm, which helps find the position of an object in an image. However, the disadvantage of the two-class image segmentation problem is that it cannot distinguish the category of the segmented image, so we extend it to multi-class labeling, which is the current semantic segmentation to distinguish the position and category of objects in the scene [2-3].

When achieving high-quality semantic segmentation, we need to focus on two core issues. First, efficient feature representations need to be proposed to distinguish different classes of objects. Secondly, how the algorithm model uses contextual information to ensure the matching between pixels and labels is also a problem we need to care about. For the first problem, the traditional solution is to use artificial features. The representative of these features is the Histograms of Oriented Gradient (HOG) [4]. In recent years, computer vision tasks including semantic segmentation have made great progress under deep learning models. Deep learning models can automatically extract image features through neural networks, so that there is no need to manually extract features from images. And for the second problem, the most common algorithm is to use contextual models framed by graph models. By setting the connections between adjacent pixels, these graphical models can easily exploit various

relationships between classes. In the semantic segmentation task, the more representative context models are Markov Random Field and its variant Conditional Random Field.

Although there are many models and algorithms to solve the above problems, these algorithms are not mature enough and face many challenges. First, there is currently no universally accepted handcrafted feature, and second, the inference of Markov Random Field or Conditional Random Field usually requires approximate computations, making their inference a challenging problem. Therefore, we need to create new and creative semantic segmentation algorithms. In recent years, due to the great success of deep learning in computer vision tasks, semantic segmentation methods centered on deep learning models have become mainstream, and have achieved significant improvements in evaluation indicators such as accuracy compared to traditional methods.

This paper gives a comprehensive description of semantic segmentation, mainly analyzes commonly used methods and points out the problems and corresponding strategies of these methods. The current semantic segmentation algorithms are mainly divided into traditional methods centered on context models and methods centered on deep learning. The above two methods do not have obvious boundaries and can be combined. We will give the corresponding basic algorithms for each method and then introduce and analyze them. Additionally, we will introduce commonly used datasets for semantic segmentation, evaluation metrics, and existing challenges.

The rest of this review is: Section 2 introduces commonly used 2D datasets in semantic segmentation, Section 3 introduces traditional methods for semantic segmentation, Section 4 introduces deep learning related methods, Section 5 introduces evaluation metrics for semantic segmentation, Section 6 presents the existing challenges and concludes with a brief summary in section 7.

2. Datasets

For computer vision tasks, a good dataset is very important. In recent years, semantic segmentation has mainly focused on 2D images, but some works have also tried using 3D images. In addition, there are some works that propose the concept of co-segmentation, which is to jointly segment similar parts from a given set of images. This article mainly introduces some of the more popular 2D datasets. In general, the available datasets will vary depending on the specific application. In medical imaging we usually use grayscale images, but otherwise basically color images.

PASCAL VOC: PASCAL VOC is a computer vision related competition which has held a challenge competition since 2005 [5]. The content of each year is different. From the initial image classification task, to the later content of detection, semantic segmentation, human body layout, action recognition, etc., the capacity and type of data sets are also constantly increasing. additions and improvements, the competition ended in 2012. Many excellent models in computer vision, especially some object detection algorithms, are launched on the PASCAL VOC competition, such as the R-CNN and YOLO, etc. Since 2007, the competition has added a segmentation challenge, and in the 2012 semantic segmentation competition, the dataset has a total of 21 categories including cat, dog, cow, person. The dataset has 1464 image as training set and 1449 images as validation set. This dataset is one of the most popular semantic segmentation datasets and has received widespread use.

Microsoft COCO: COCO is a computer vision dataset released by Microsoft, which is mainly used for tasks such as object detection and semantic segmentation. In terms of semantic segmentation, the large-scale dataset has 82783 images as training set, 40504 images as validation set and 80000 images as test set [6]. This dataset has more than 80 categories for semantic segmentation. Because of the large size of this dataset, it has grown in importance and popularity among researchers.

SYNTHIA SYNTHIA is a semantic segmentation dataset with virtual city as the background, which can be used to understand urban scenes or autonomous driving. the dataset has a total of 11 categories including building, car, road, sky, void, and provides pixel-level annotations. The training set of this dataset has 13,407 images and exhibits diversity in terms of scene, season, and weather.

Cityscapes Cityscapes is a dataset for semantic segmentation based on street scenes. It classifies images from 30 different categories into 8 categories, some of which are human, vehicle, sky, void, and provides semantic information, instances, and dense pixel annotations. The dataset consists of training set and validation set, consisting of 22973 and 500 images, respectively. The images in the dataset were collected in 50 different cities under good weather conditions, and are characterized by a variety of scenes and backgrounds [8].

MINC MINC is a dataset for semantic segmentation and image classification of materials. The dataset contains 23 different categories with pixel-level annotations. The dataset contains 7061 training set images, 2500 validation set images and 5000 test set images. Besides, some semantic segmentation datasets such as PASCAL Context, Semantic Boundaries Dataset (SBD) KITTI and other datasets are also frequently used. Table 1 below shows a summary of the datasets mentioned above, and gives the name, year, category, the number of images in the training sets, validation sets, and test sets [9].

Table 1 Summary of the common dataset

Datasets	Year	Classes	Training sets	Validation sets	Test sets
SBD	2011	21	8498	2857	/
PASCAL VOC	2012	21	1464	1449	/
PASCAL Context	2014	540	10103	/	9637
Microsoft COCO	2014	80+	82783	40504	81434
MINC	2015	23	7061	2500	5000
Cityscapes	2015	8	22973	500	/
KITTI	2015	11	170	/	46
SYNTHIA	2016	11	13407	/	/

3. Traditional Algorithm

Traditional algorithms for semantic segmentation use artificial features for feature extraction. Traditional algorithms mainly include two models. The first is the Contextual Models, which is modeled by a graphical model to solve the posterior distribution. The second is the superpixel model, which represents entities by connecting pixels to form superpixels.

3.1. Contextual Models

As we all know, graph algorithms are widely used in computer vision tasks. An undirected graph $G = \{V, E\}$ is defined, where E and V are the set of edges and the set of graph vertices respectively, in addition, we need to apply a weight to each edge in E to indicate the importance of the edge. In semantic segmentation, the vertices of the graph represent the pixels of the image, the edges of the graph represent the relationship between two adjacent pixels, and the weights are related to the features of the two pixels connected by the edge [10]. When applied, graph theory is often combined with Markov theory to build context models. Typical examples are Conditional Random Field and Markov Random Field. Spatial coherence can be ensured using contextual models to show relationships between modeled pixels or regions. The semantic segmentation problem is actually a labeling problem. Let the number of classes, the number of pixels and the label of pixel i be l、N and y_i respectively, then the labeling result of the image can be expressed as: $y = \{y_1, \dots, y_i, \dots, y_N\}$, where $y_i \in \{1,2, \dots, L\}$. Let the possible results of all annotations be Y, then the number of elements in Y is L^N.

3.1.1 Markov Random Field

In the contextual model based on the Markov Random Field, the prior $P(y)$ of the label y is usually modeled as a Markov random field (MRF) with a neighborhood system ε defined on a pixel lattice. According to Bayes' theorem, given an image observation x, the posterior distribution of y is:

$$P(y \mid x) = \frac{P(x|y)P(y)}{P(x)} \propto P(x \mid y)P(y) \quad (1)$$

Among them, when calculating $P(x \mid y)$, it is usually assumed that the pixels are independent and identically distributed, so:

$$P(x \mid y) = \prod_{i=1}^{N} P(x_i \mid y_i) \quad (2)$$

In this way, the original problem is transformed into maximizing the distribution function $P(x \mid y)$ (MAP):

$$y^* = \arg\max_{y \in Y} P(y \mid x) = \arg \max_{y \in Y} P(y) \prod_{i=1}^{N} P(x_i \mid y_i) \quad (3)$$

However, some work pointed out that contextual models based on Markov Random Field have the following disadvantages when applied to semantic segmentation tasks. First, pixels need to strictly satisfy the property of independent and identical distribution, which is difficult to achieve in reality. Second, the inference of MAP can be complex despite the simple posterior distribution. So the method of CRF-MAP is more popular.

3.1.2 Conditional random fields CRF

Conditional random fields (CRF) can be considered as a variant of MRF [11]:

$$P(y \mid x) = \frac{1}{Z} \exp(-U(y)) = \frac{1}{Z} \exp\left(-\sum_{c \in C} \Psi(y_c \mid x_c)\right) \quad (4)$$

where y is the label, $U(y)$ is the energy corresponding to the label y, Z is a normalization constant that ensures that the sum of the distributions is 1, and $\Psi(y_c \mid x_c)$ is defined in the A set of connected pixels c∈C, so we transform the MAP problem into a problem of energy minimization:

$$y^* = \arg \max_{y \in Y} P(y \mid x) = \arg \min_{y \in Y} \sum_{c \in C} \Psi(y_c \mid x_c) = \arg \min_{y \in Y} E(y \mid x) \quad (5)$$

The commonly used energy functions are:

$$E(y) = \sum_{i=1}^{N} \varphi_i(y_i \mid x_i) + \sum_{(i,j) \in \varepsilon} \varphi_{ij}(y_i, y_j \mid x_{ij}) \quad (6)$$

Among them, φ_i is called unary potential, and φ_{ij} is called pairwise potential. Some work show that the unary potential represents the degree of matching between the pixel i and the label y_i, and is usually expressed as the negative log-likelihood of the probability $P(y_i \mid x_i)$:

$$\varphi_i(y_i) = -\log P(y_i \mid x_i) \quad (7)$$

With the above method, the problem of maximizing the posterior distribution of conditional random fields is equivalent to minimizing the energy function. However, due to the existence of many local minima, the global minimization of the energy function is very difficult. The Simulated Annealing in intelligent optimization algorithms is often used to solve the problem of global minimization difficulty [12].

3.2. Superpixels

Work proposed that, compared to pixels, a superpixel consisting of a set of similar and connected pixels is more suitable for representing entities, which can greatly reduce the computational complexity and allow a region provide much more cues. Due to the advantages of superpixels, they are increasingly used in the field of semantic segmentation, such as mean-shift. However, the superpixels generated by traditional methods usually lack control, resulting in inefficient feature extraction. To improve this shortcoming, the Turbopixel algorithm generates superpixels by dilate a set of seed locations using geometric flow, and the Simple Linear Iterative Clustering (SLIC) algorithm generates superpixels by using the localized K-means algorithm clusters pixels to produce superpixels.

4. Deep-learning-based semantic segmentation methods

4.1. Basic Network Architecture

In the past few years, deep learning has made rapid progress. Certain networks, such as CNN, VGG-16, ResNet, GAN, Transformer, have been used as the basic block for semantic segmentation tasks, including. The following sections provide a brief review of these modules.

4.1.1 VGG Net

Visual Geometry Group (VGG) is a convolutional neural network proposed by the Visual Geometry Group of Oxford University [13]. The network proves that the depth of the neural network has a crucial impact on the results. Unlike AlexNet, VGG's first convolutional layer has a smaller receptive field, which results in fewer parameters and faster model convergence. The network architecture of VGG-16 is shown in the figure 1.

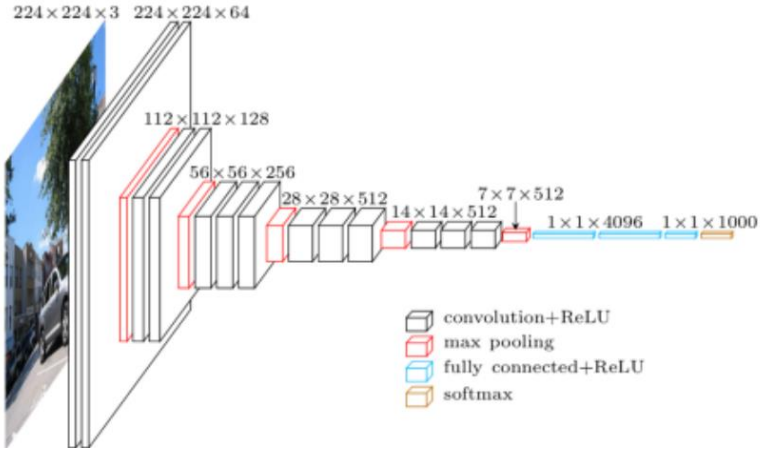


Fig. 1 The network architecture of VGG-16

4.1.2 ResNet:

For deep learning models, the depth of the model is of great significance and can directly affect the effect of the model. The advantage of the deep network structure is that it can help us extract deeper and more complex features of the image Through trial and error, the researchers found that deep networks can cause problems with degradation: when the network gets deeper and the structure becomes more complex, the accuracy does not improve as expected, but decreases [13]. ResNet was the 2015 ILSVRC champion, and the intuitive idea of ResNet is that it uses residual connections to ensure that what the next layer learns is different from what has been learned before. The structure of the residual connection is shown in the figure 2.

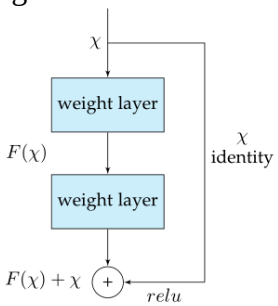


Fig. 2 Schematic diagram of residual linking

4.2. Deep Learning Methods

After reviewing the basic structure, the basic deep learning model of semantic segmentation is introduced below. We mainly introduce supervised methods. The supervised semantic segmentation model assumes that there is enough labeled training data, i.e., enough raw images and corresponding

pixel semantically annotated images. In 2011, Lucchi et al. pointed out that contextual information can sense semantic information, complement semantic segmentation models and improve the accuracy of semantic segmentation. Therefore, some work proposed a context-based method: DilatedNet [14]. Unlike traditional convolutional neural networks, this method aggregates multi-scale contexts by dilated convolution. DilatedNet extracts contextual information from images by using dilated convolutions of different scales. Typically, the network uses three different expansion rates of 2, 3, and 8. From the results, DilatedNet also achieved good results, the MIoU index of DilatedNet reached 67.7% on the PASCAL VOC test set [15]. Likewise, as a context information algorithm, PSPNet introduces a multi-scale context aggregation module (PPM), as shown in the figure 3.

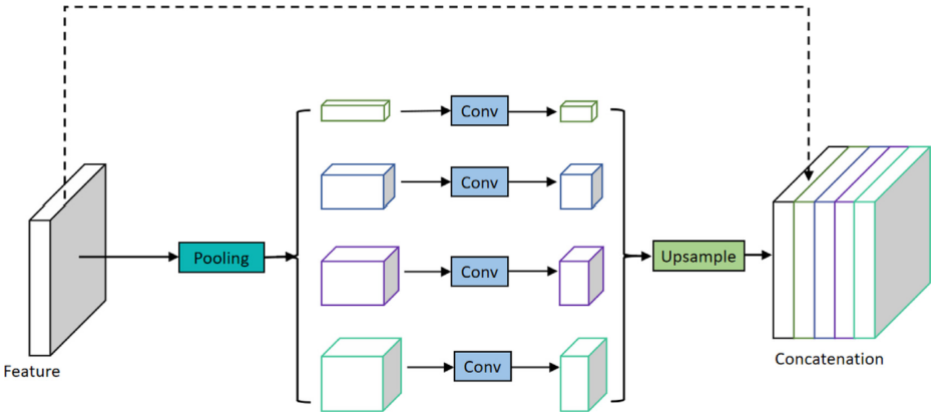


Fig. 3 The network of PPM.

In addition to the algorithms based on contextual information, other algorithms such as: methods based on feature enhancement, methods based on deconvolution, methods based on Recurrent Neural Network [16], methods based on Generative Adversarial Network are also frequently used as fully supervised deep learning methods for semantic segmentation. Figure 4 below shows the Macro-Micro Adversarial Network (MMA Net) commonly used in methods based on Generative Adversarial Networks.

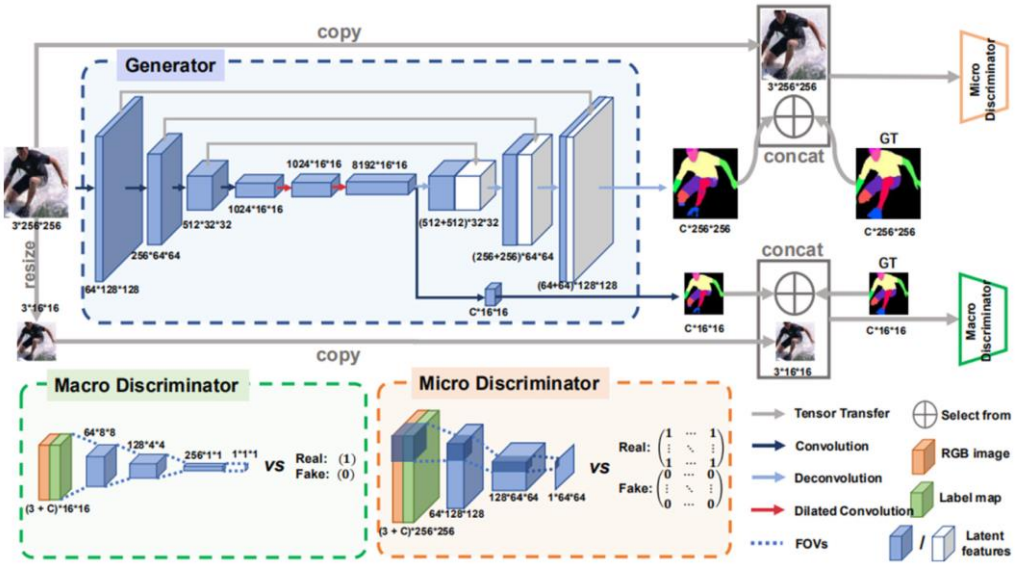


Fig. 4 The network of MMA Net.

5. Evaluation

In order to measure and compare the effects of different semantic segmentation models more accurately, specific metrics must be used to evaluate algorithms. In order to comprehensively

different algorithms, the evaluation metrics must involve many aspects: the algorithm running time (i.e. speed), accuracy, stability and other metrics.

Speed: In some application fields, algorithms need to meet certain real-time requirements. For example, in autonomous driving, some models divide pixels into streets and non-streets, and then perform semantic segmentation. In order to ensure real-time performance, the processing time of each image have to within 20ms, which is a delay. For the above semantic segmentation tasks, real-time performance is very important, in order to improve the execution speed, the accuracy can be sacrificed to a certain extent.

Stability: For semantic segmentation algorithms, the segmentation results need to remain stable when the input image changes slightly. For example, when the image is affected by factors such as smoke, light, occlusion, etc., the segmentation result does not change or only slightly changes.

Accuracy: When evaluating the performance of models used for semantic segmentation, the following metrics are used to represent accuracy. Let $x_{ij} \in \mathbb{N}_0$ be the number of pixels belonging to class i but labeled as class j, and $l \in \mathbb{N}$the number of classes, where $i, j \in 1, \dots, l$. Then $(x_{ij})_{l \times l}$ is the confusion matrix, and $X_i = \sum_{j=1}^{l} x_{ij}$ is the total number of pixels of the i-th class. From this we can define a per-pixel rate to represent the pixel-wise accuracy:

$$\text{Pixel Accuracy} = \frac{\sum_{i=1}^{l} x_{ii}}{\sum_{i=1}^{l} X_i} = \frac{\sum_{i=1}^{l} x_{ii}}{\sum_{i=1}^{l} \sum_{j=1}^{l} x_{ij}} \tag{8}$$

Among them, $\sum_{i=1}^{l} x_{ii}$ represents the sum of the number of correctly classified pixels, and $\sum_{i=1}^{l} X_i$ represents the sum of all pixels, so this ratio represents the proportion of correctly predicted pixels.

However, using the above accuracy has the following disadvantages:

Certain semantic segmentation tasks, such as autonomous driving, can achieve classification accuracy of 30% and above with only prior knowledge. Specifically, our model might learn from prior knowledge that one area is mostly "roads" and another location is "sky" most of the time, etc.

When manually labeling images, the labeling results are coarser. For example, a pixel or area may be labeled as a bicycle during manual classification, and the model algorithm may divide the area into a general bicycle and a more specific bicycle wheel.

Also, a known pixel cannot be assigned to a known class.

To address the above shortcomings, we use other accuracy metrics to evaluate the performance of the model. Work uses different accuracy metrics to address problem 1:

Mean pixel accuracy (MPA):

$$\frac{1}{l} \cdot \sum_{i=1}^{l} \frac{x_{ii}}{X_i} = \frac{1}{l} \cdot \sum_{i=1}^{l} \frac{x_{ii}}{\sum_{j=1}^{l} x_{ij}} \in [0,1] \tag{9}$$

Mean intersection over union(MioU):

$$\frac{1}{l} \cdot \sum_{i=1}^{l} \frac{x_{ii}}{X_i - x_{ii} + \sum_{j=1}^{l} x_{ji}} = \frac{1}{l} \cdot \sum_{i=1}^{l} \frac{x_{ii}}{\sum_{j=1}^{l} x_{ij} - x_{ii} + \sum_{j=1}^{l} x_{ji}} \in [0,1] \tag{10}$$

Frequency weighted intersection over union (FWIoU):

$$\left(\sum_{i=1}^{l} X_i\right)^{-1} \sum_{i=1}^{l} X_i \cdot \frac{x_{ii}}{X_i - x_{ii} + \sum_{j=1}^{l} x_{ji}} \tag{11}$$

$$= \left(\sum_{i=1}^{l} \sum_{j=1}^{l} x_{ij}\right)^{-1} \sum_{i=1}^{l} \sum_{j=1}^{l} x_{ij} \cdot \frac{x_{ii}}{\sum_{j=1}^{l} x_{ij} - x_{ii} + \sum_{j=1}^{l} x_{ji}} \in [0,1]$$

Among them, $\sum_{j=1}^{l} x_{ji}$ represents the number of all pixels marked as category i, and X_i represents the number of pixels that are actually category i. The intersection of the two is x_{ii} which is actually category i and then marked as the number of pixels for class i. Among all the metrics,

MIoU can well represent the results of semantic segmentation and is concise and clear, so it is widely used by researchers.

There are also works using confusion matrix methods to solve problems 1 and 2. But if there are many classes in the training set, then this method is not feasible. This work also uses the methods of the void category and to solve problem 3. Specifically, this work uses a void class that is different from the class known in the training set. We classify pixels that do not belong to a known class as void class, so at this point the total number of pixels becomes the image's $height * width - number\ of\ pixels\ in\ the\ void\ class$. Then we can use the above-mentioned accuracy index to calculate to solve problem 3.

For binary classification semantic segmentation task, F-measure is usually used to evaluate the accuracy of semantic segmentation models:

$$F_\beta = (1+\beta)^2 \frac{\mathrm{tp}}{(1+\beta^2)\cdot \mathrm{tp} + \beta^2 \cdot \mathrm{fn} + \mathrm{fp}} \tag{12}$$

Where β represents the parameter, in most cases β=1, tp is the true positive rate, and fn is the false positive rate.

Finally, it is important to note that for non-semantic segmentation, we propose many other methods to improve segmentation accuracy. One such measure of improved accuracy is the normalized probabilistic rand (NPR) index. Only the method of semantic segmentation is introduced here, and the method of non-semantic segmentation is not involved.

6. Existing Issues and Challenges

Although deep neural networks have achieved many achievements in semantic segmentation recently, there are still many problems in the semantic segmentation task waiting for us to solve.

First, how to balance the accuracy and real-time performance of semantic segmentation algorithms is an important issue. Many times, in academic research, the implementation of semantic segmentation algorithms is not very important, and researchers tend to focus on the accuracy of the model. However, the real-time performance of semantic segmentation algorithms is more important for applications such as autonomous driving and urban safety. Usually, models with high accuracy tend to have poor real-time performance, while models with high real-time performance cannot provide more accurate results.

Second, today's semantic segmentation tasks rely on high quality data for training. To train a model to the desired effect, researchers often need to acquire enough high-quality images and label each pixel, a laborious and time-consuming task. The solution to this problem is to use semi-supervised or weakly supervised learning methods, but the effect of both is not good and further research is needed.

7. Conclusion

This article provides a brief review of the semantic segmentation task from five directions: datasets, traditional methods, basic deep learning methods, evaluation methods, and problems and challenges. We present different datasets and semantic segmentation algorithms in the form of pictures and tables and compare these methods. Nowadays, with the widespread use of deep learning in this field, one breakthrough after another in semantic segmentation, and it can be applied to multiple application directions such as autonomous driving. But there are still many challenges in semantic segmentation waiting to be solved by new models.

References

[1] Carreira, Joao, and Cristian Sminchisescu. CPMC: Automatic object segmentation using constrained parametric min-cuts. IEEE transactions on pattern analysis and machine intelligence 2011, 34.7: 1312-1328.

[2] Dalal, Navneet, and Bill Triggs. Histograms of oriented gradients for human detection. 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). 2005 1: 129-137.

[3] Everingham, Mark, et al. The pascal visual object classes challenge: A retrospective." International journal of computer vision 2015,111.1: 98-136.

[4] Ros, German, et al. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016 261:731-740.

[5] Cordts, Marius, et al. The cityscapes dataset. CVPR Workshop on the Future of Datasets in Vision. 2015, 2:121-129.

[6] Bell, Sean, et al. Material recognition in the wild with the materials in context database. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, 33:249-252.

[7] Hariharan, Bharath, et al. Semantic contours from inverse detectors. 2011 international conference on computer vision. 2011, 23:762-771.

[8] Ros, German, et al. Vision-based offline-online perception paradigm for autonomous driving. 2015 IEEE Winter Conference on Applications of Computer Vision. IEEE, 2015,13:53-61.

[9] Zhang, Kunlei, Jun Deng, and Wenmiao Lu. Segmenting human knee cartilage automatically from multi-contrast MR images using support vector machines and discriminative random fields. 2011 18th IEEE International Conference on Image Processing. 2011, 392:14-23.

[10] Kirkpatrick, Scott, C. Daniel Gelatt Jr, and Mario P. Vecchi. Optimization by simulated annealing. science 1983, 220.4598: 671-680.

[11] Comaniciu, Dorin, and Peter Meer. Mean shift: A robust approach toward feature space analysis. IEEE Transactions on pattern analysis and machine intelligence 2002, 24.5: 603-619.

[12] Levinshtein, Alex, et al. Turbopixels: Fast superpixels using geometric flows. IEEE transactions on pattern analysis and machine intelligence 2009, 31.12: 2290-2297.

[13] Achanta, Radhakrishna, et al. "SLIC superpixels compared to state-of-the-art superpixel methods." IEEE transactions on pattern analysis and machine intelligence 2012: 2274-2282.

[14] Noh, Hyeonwoo, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. Proceedings of the IEEE international conference on computer vision. 2015, 26.5: 512-519.

[15] Pinheiro, Pedro, and Ronan Collobert. Recurrent convolutional neural networks for scene labeling. International conference on machine learning. 2014, 244:529-535.

[16] Luo, Yawei, et al. Macro-micro adversarial network for human parsing. Proceedings of the European conference on computer vision. 2018,121:1-11.