Research On Semantic Segmentation Based on Deep Learning

Xiang Li[1], Jiaye Wu[2, 3], Yujie Yang[1]

[1]School of Automation and Information Engineering, Sichuan University of Science & Engineering, Yibin, China

[2]School of Mechanical Engineering, Southwest Petroleum University, Chengdu, China

[3]Sichuan Central Inspection Technology Inc., Zigong Sichuan, China

Abstract

With the development of deep learning technology, the integration of semantic segmentation and deep learning has made great technical breakthroughs. Image semantic segmentation has become one of the research hotspots in computer vision field. This technology has been widely used in medical image segmentation, remote sensing image detection, intelligent robot and other fields. This paper first describes the basic network model of semantic segmentation in detail, then introduces the application of semantic segmentation in different fields, and finally looks forward to the future research focus of semantic segmentation.

Keywords

Semantic segmentation; Computer vision; Deep learning; Convolutional neural network.

1. INTRODUCTION

Image semantic segmentation technology refers to the process of assigning individual class labels to each pixel in an image based on its grayscale, color, texture, and other features. This division of the image into visually meaningful and distinct regions allows for a more precise understanding of the image [1]. With the advancement of intelligent living, semantic segmentation technology has become increasingly important in various fields such as autonomous driving, medical image processing, video surveillance, virtual interaction, and augmented reality. Traditional methods of semantic segmentation include threshold-based[2], region-based[3], edge detection[4], clustering[5], graph-based approaches that utilize mathematical theories[6], as well as machine learning methods such as texture primitive forests or random forests for constructing pixel classifiers[7]. However, these traditional methods have limitations in terms of efficiency and accuracy. They are less efficient in handling low-level semantic features such as color, shape, and texture in images, resulting in longer segmentation times and lower accuracy. Additionally, these methods struggle to recognize occluded objects[8]. With the advancement of hardware performance and the rise of deep learning, Deep Convolutional Neural Networks (DCNNs) have emerged as powerful tools for semantic segmentation[9]. In 2015, Long et al. introduced Fully Convolutional Networks (FCNs)[10], which applied DCNNs to semantic segmentation. This marked the beginning of an era dominated by DCNNs in the field of semantic segmentation. DCNNs allow for end-to-end training, enabling the extraction and learning of semantic-level image features. This allows the network to actively infer the semantic information of each pixel and classify them, leading to higher segmentation accuracy and computational efficiency. Existing literature provides comprehensive overviews of semantic segmentation, covering various methods and data patterns. Another study focuses on the issue of varying quality in training datasets for semantic segmentation models, analyzing both fully supervised and weakly supervised training approaches[11][12]. To provide a more in-depth analysis of the current development of deep neural networks in semantic segmentation, this paper supplements and enriches the existing knowledge by exploring different technical characteristics in the field. The advantages and disadvantages of each category are analyzed, and experimental results of different models are compared using commonly used datasets. Finally, the current and future development trends, as well as the remaining challenges, are discussed.

2. SEMANTIC SEGMENTATION ALGORITHM

2.1. Traditional image segmentation algorithm

Although deep learning currently dominates the research in image segmentation, the advantages of traditional image segmentation algorithms cannot be denied. They offer faster and more efficient problem-solving capabilities. Traditional image segmentation algorithms can still provide valuable insights for solving problems in deep learning. This article will introduce three traditional image segmentation methods: threshold-based segmentation, edge detection-based segmentation, and region-based segmentation.

2.1.1. Threshold based segmentation method

Thresholding is particularly suitable for images with different grayscale levels for the background and the target. The basic idea is to calculate one or multiple grayscale thresholds based on the grayscale features of the image. Each pixel's grayscale value is compared with the computed thresholds, and based on the comparison results, the pixels are classified into appropriate categories. Therefore, the key aspect of this method is to determine the optimal grayscale thresholds based on certain criterion functions. If the image contains only two classes, the target and the background, a single threshold can be selected for segmentation, which is known as single-threshold segmentation. However, if there are two or more types of targets in the image, the single-threshold segmentation method is not applicable. In such cases, multiple thresholds are used to segment the targets, which is known as multi-threshold segmentation. Common methods in thresholding include fixed threshold segmentation, histogram-based bimodal method, iterative threshold image segmentation, adaptive threshold image segmentation, maximum interclass variance method, mean-based method, and optimal thresholding.

2.1.2. Edge based segmentation method

Edge detection segmentation is a commonly used method for image segmentation. In different regions of an image, there are grayscale and color changes, which result in abrupt transitions at the edges between these regions. Grayscale-based edge detection is an observation-based method where the edges between different regions exhibit step or roof-like changes in grayscale values. When transforming the image from the spatial domain to the frequency domain, edges correspond to high-frequency components. The differential operator is a commonly used edge detection algorithm that utilizes the extremum of first-order derivatives and the zero-crossings of second-order derivatives to determine edges. To achieve better segmentation results, edge detection algorithms can be used in conjunction with complementary segmentation methods.

2.1.3. Region based segmentation method

Region-based segmentation methods determine a base region based on certain criteria and use it as a starting point for segmentation. There are two fundamental forms of region segmentation: region growing and global approaches. In region growing, the segmentation starts from a seed pixel and gradually expands by merging neighboring pixels with similar properties. In the global approach, the entire image is treated as a whole and segmented into different sub-regions. Common region-based segmentation algorithms include seed region growing, region splitting and merging, and watershed segmentation.

2.2. Semantic segmentation algorithm based on convolutional neural network

With the introduction of fully convolutional neural networks (FCNs), their superior feature extraction performance compared to traditional segmentation methods has made them the mainstream approach in semantic segmentation. Contrary to traditional image segmentation methods, FCNs can extract high-level semantic information from images, thus improving segmentation accuracy. Since the proposal of FCNs, classic networks such as U-net, PSPnet, and Deeplab have emerged, greatly influencing the development of subsequent semantic segmentation networks.

2.2.1. FCN

Fully Convolutional Networks (FCNs) marked the beginning of semantic segmentation and since then, the field has rapidly progressed. The end-to-end training of network models is also achieved through FCNs. FCNs have made significant contributions in three aspects: fully convolutional, upsampling, and skip connections. In terms of being fully convolutional, while conventional CNN classification networks have a fixed input image size determined by the network's design structure, FCNs allow for inputs of varying sizes. FCNs replace the last three fully connected layers of the CNN classification network with convolutional layers, preserving both the spatial information of the image and integrating the output features of the CNN. Regarding upsampling, after a series of convolution and pooling operations, the resulting feature map size is much smaller than the original image size. To associate the pixels in the feature map with those in the original image for pixel-level prediction and minimize segmentation accuracy loss, the authors employ deconvolution operations. During feature map decoding, deconvolution is used to resize the feature map to match the original image size. As for skip connections, FCNs lose many fine-grained details after convolution, pooling, and deconvolution operations. By employing skip connections, the shallow-level information and high-level semantic information are combined to enhance the model's robustness. Although FCNs achieve pixel-level image prediction, they overlook global contextual information.

2.2.2. U-net

U-net[13] was initially designed as a segmentation network for medical image segmentation. It utilizes an encoder-decoder structure and incorporates skip connections to fuse shallow features with high-level semantics. In the encoder section, the image undergoes four downsampling operations through a combination of convolutional layers and max-pooling layers. With each downsampling step, the channel dimensions of the feature maps double. In the decoder section, after each upsampling operation, the feature maps are fused with corresponding downsampling feature maps, followed by a halving of the channel dimensions. In the final layer of the decoder, a 1x1 convolution is used to adjust the output channel to the desired number of classes for classification. However, U-net has notable drawbacks. It suffers from slow training speed, as the same features are trained multiple times, resulting in GPU resource wastage. It also runs the risk of overfitting, leading to poor generalization of the trained network. Furthermore, U-net struggles to simultaneously obtain accurate object localization and contextual information. The use of larger patches requires more max-pooling operations, which can degrade localization accuracy by losing spatial relationships between target pixels and their surroundings. On the other hand, smaller patches can only capture limited local information and may lack sufficient background context.

2.2.3. PSPnet

The main innovation of PSPnet[14] is the introduction of the pyramid pooling module, which aggregates contextual information from different positions of the target and improves the performance of capturing global information. Additionally, auxiliary loss functions are incorporated to accelerate the convergence speed during network training. Given an input image (a), the last convolutional layer's feature map (b) is obtained using CNN. Then, the pyramid pooling module is applied to acquire representations of different sub-regions. Subsequently, upsampling and concatenation layers form the final feature representation (c), which carries both local and global contextual information. Finally, the representation is fed into a convolutional layer to obtain the final pixel prediction (d). The pyramid pooling module integrates features from four pyramid scales. he first parallel branch employs global pooling to generate a global feature map. The other parallel branches perform pooling operations on the feature map to obtain feature maps of different regions, which are then fused together. Each parallel branch utilizes different pooling and 1x1 convolution operations to obtain feature maps of different sizes. The low-dimensional feature maps are directly upsampled using bilinear interpolation to match the size of the original feature map. Ultimately, the fused feature maps from the four parallel branches serve as the global features of the pyramid pooling module.

2.2.4. Deeplab series

DeepLab is a semantic segmentation method proposed by Google. In DeepLabv1[15], two main issues of deep convolutional neural networks were addressed. The problem of losing positional information due to repeated downsampling was solved by using dilated convolutions. The problem of coarse segmentation results caused by the spatial invariance of DCNN was addressed by using Conditional Random Fields. In DeepLabv2[16], an Atrous Spatial Pyramid Pooling (ASPP) module was introduced to tackle the multiscale problem. It involved processing the feature map with four different dilated convolutions with varying rates, summing up the processed feature maps, and then upsampling them. Additionally, ResNet was used as the backbone network in DeepLabv2. In DeepLabv3[17], improvements were made to the ASPP module. It included a 1×1 ordinary convolution, three 3×3 dilated convolutions with dilation rates of 6, 12, and 18, and the addition of batch normalization (BN) layers. To obtain image-level features, global average pooling was applied to the feature map output by the backbone network. The new ASPP module had five branches. In DeepLabv3+[18], to achieve better segmentation results, the entire DeepLabv3 was used as the encoder for feature extraction. In the decoder, the features obtained from the encoder were processed with 1×1 convolutions and upsampled by a factor of 4. Low-level features from the backbone network were adjusted with 1 × 1 convolutions to match the channel dimension and then fused with the upsampled feature maps. Finally, the fused feature map underwent two 3×3 convolutions and was upsampled by a factor of 4 to restore the original image size.

2.3. Semantic segmentation algorithm based on Transformer

Inspired by the Transformer model in natural language processing, many researchers have attempted to apply Transformer to the field of semantic segmentation. By leveraging the attention mechanism of Transformer to establish long-range dependencies, significant achievements have been made. Currently, Transformer remains a hot research topic in this domain.

2.3.1. Segmenter

Segmenter[19] is an encoder-decoder structure based on the Transformer model. Unlike CNN, before feeding the image into the encoder, the image needs to be divided into multiple patches and flattened into one-dimensional sequences, which are then encoded with positional encodings. Compared to convolution-based methods, the encoder of Segmenter is used to model the global contextual information of the image. Segmenter has two types of decoders: a linear decoder, which performs simple linear mapping, deformations, upsampling, and softmax operations on the patches to generate predicted images; and a mask decoder based on

Transformer, which differs from the linear decoder by incorporating a set of learnable class embedding vectors into the decoder's input. Experimental results show that the mask decoder based on Transformer outperforms the linear decoder in terms of segmentation performance.

2.3.2. SegFormer

In response to the issues of large parameters and computational complexity of ViT, as well as the unfriendliness of the columnar structure for semantic segmentation, the authors of SegFormer[20] designed a hierarchical Transformer encoder. When embedding patches, they introduced overlapping designs to ensure local continuity of features, and replaced positional encodings with deep convolutions to convey positional information. The encoder of SegFormer consists of only six linear layers, resulting in smaller parameter size and computational complexity, yet it achieves excellent segmentation performance. Compared to CNN networks like Deeplabv3+, SegFormer exhibits stronger robustness.

3. APPLICATION OF SEMANTIC SEGMENTATION

3.1. Semantic segmentation of remote sensing images

Remote sensing is the process of acquiring information and monitoring the characteristics of an area without any physical contact. There are two main types of remote sensing technologies: active sensors, such as radar and lidar, and passive sensors, such as satellite imagery [21]. These high-resolution images of the Earth's surface provide a wide range of use cases, including updating world maps, analyzing forest degradation, and monitoring surface changes. Remote sensing images, combined with computer vision and artificial intelligence (AI), are widely used for analyzing and processing large-scale Earth surface areas with complex feature distributions. Images collected by satellites or unmanned aerial vehicles (UAVs) provide extensive information for applications such as urban planning, disaster management, traffic management, climate change, wildlife conservation, and crop monitoring. Datasets that include these high-resolution images and their respective segmentation masks [22] form the foundation for using computer vision and AI to analyze remote sensing images. Neural networks enable the processing of large amounts of image data for tasks such as object detection, semantic segmentation, and change detection. The development in the field of remote sensing has further improved satellite sensors, and the introduction of UAV technology is crucial for capturing finer details of the Earth's surface. This has led to precise and accurate data processed using AI techniques [23]. Remote sensing images of the Earth's surface provide information about land cover, which can be divided into different segment classes. Each category assigns a label to each pixel while preserving the spatial resolution of the image. Many datasets containing high-resolution remote sensing images and their segmentation masks are available for various applications such as change detection, land cover segmentation, and classification. Common land cover categories covered by pixel-level classification include forests, crops, buildings, water resources, grasslands, and roads. Researchers have used ViT architecture models to effectively add layers and attention mechanisms, improving performance for semantic segmentation ofhigh-resolution remote sensing images, such as Efficient Transformer and Wide-Context Transformer.

3.2. Semantic segmentation of medical images

Medical image analysis has advanced and integrated scanning and visualization technologies. Segmentation techniques are crucial as they can identify and segment medical images to assist in further diagnosis and intervention. By identifying and highlighting regions of interest (ROIs) in medical images, various important diagnoses can be made, such as detecting brain tumor boundaries from MRI images, identifying pneumonia impacts in X-rays, and detecting cancer in biopsy sample images. Recently, there has been a growing demand for image segmentation in this type of analysis, leading to extensive research in developing more accurate and efficient models and algorithms. Medical images used for image segmentation tasks can be grouped based on imaging modalities, including MRI, CT scans, X-rays, ultrasound, microscopy, dermoscopy, and more. Each category includes datasets collected under medical supervision, some of which are publicly available. Due to the existence of these various modalities, the technology systems used for medical imaging can vary greatly. Medical imaging system developers build these systems according to the requirements of healthcare professionals. The generated images are subject to limitations imposed by existing technologies and require the involvement of medical personnel for examination [24]. Therefore, the segmentation of these images in different biomedical fields requires domain experts who are knowledgeable about these systems and spend a significant amount of time reviewing them. To overcome these challenges, the capability of automatic feature extraction has been introduced through deep learning-based techniques, which have proven valuable in the context of medical image analysis. As segmentation analysis techniques have evolved, many researchers have introduced models with improved performance using medical images. One well-known architecture is U-Net, which was initially introduced for medical image analysis. Building upon this, several improved versions have been developed using medical image datasets for cardiac, lesion, and liver segmentation. This demonstrates how improvements in segmentation can greatly benefit the medical environment. In recent years, emerging architectures like ViT have also been applied in the medical field, including TransUNet [25] and Swin-Unet [26]. They are hybrid Transformer architectures that leverage the advantages of U-Net and exhibit better accuracy in applications such as cardiac and multi-organ segmentation. Medical imaging faces certain limitations, such as the relatively limited availability of images compared to natural image datasets (e.g., landscapes, people, animals, and cars) that consist of millions of images. In the field of medicine, there are several imaging modalities, and expertise in each medical domain is required for annotating medical images. MRI and microscopy images, in particular, are challenging to annotate. Typically, these datasets contain fewer images and are easier to annotate with less complex structures and fine boundaries compared to datasets consisting of ultrasound, X-ray, and lesion data obtained using existing scanning systems. However, there are still significant limitations in accessing these images due to privacy and other medical policies[27]. To overcome these limitations in certain datasets, image segmentation challenges with publicly available, well-annotated medical image datasets are held several times a year. The majority of improvements made in semantic segmentation models are based on these challenge datasets, and they serve as benchmark datasets for segmentation tasks.

3.3. Video semantic segmentation

Computer interaction, augmented reality, autonomous driving cars, and image search engines are some applications in the field of complete scene understanding. For these types of applications, semantic segmentation contributes more to the complete scene understanding of videos. Typically, the idea is to apply semantic segmentation to each frame of high-resolution videos, treating the video as a collection of unrelated still images [28]. The common challenge in this type of semantic segmentation is the computational complexity of scaling the spatial dimension of videos with the temporal frame rate. In video segmentation, it is meaningless to remove temporal features and only focus on spatial frame-by-frame features. Considering the temporal context of the video is an important factor in video semantic segmentation, even if it is computationally expensive. Research has been conducted to reduce this high computational cost on videos, and solutions such as feature reuse and feature distortion have been proposed. Cityscapes and CamVid are some of the largest video segmentation datasets available for frame-by-frame methods. Recent papers have proposed segmentation methods such as selectively re-executing feature extraction layers, feature distortion based on optical flow, and fixed-budget keyframe selection strategies based on LSTM. However, a major issue with these methods is that they pay less attention to the temporal context of the video. Researchers have demonstrated that utilizing the optical flow of videos as temporal information to accelerate uncertainty estimation is meaningful in order to meet the spatial and temporal context requirements. VisTR, TeViT, and SegFormer are some Transformer models used for video segmentation tasks.

4. CONCLUSION

This paper summarizes traditional segmentation methods and deep learning-based semantic segmentation methods, and introduces the application areas of semantic segmentation technology. Based on existing research achievements, the future research focuses of semantic segmentation are discussed. (1) Real-time semantic segmentation. In line with practical application needs, lightweight segmentation networks need to be developed for real-time segmentation. It is crucial to ensure both accuracy and efficiency in segmentation so that semantic segmentation can be widely applied in practical scenarios. (2) Sample imbalance problem. Data is the foundation of semantic segmentation. It is important to explore methods that can achieve high accuracy with limited sample data and allow the network to converge quickly even in the presence of challenging samples. (3) Unsupervised domain adaptation. Due to the difficulty in obtaining ground truth labels for data and the poor generalization ability across scenes, the development of unsupervised domain adaptation methods has been promoted. Unsupervised domain adaptation utilizes deep learning models for feature extraction and alignment to improve model transferability. Further research is needed to better understand how to perform feature alignment effectively.

REFERENCES

[1] Otsu N. A threshold selection method from gray-level histograms[J]. IEEE transactions on systems, man, and cybernetics, 1979, 9(1): 62-66.WESZKA J S. A survey of threshold selection techniques[J]. Computer Graphics and Image Processing, 1978, 7 (2): 259-265. DOI: 10.1016/0146-664X(78) 90116-8

[2] Adams R, Bischof L. Seeded region growing[J]. IEEE Transactions on pattern analysis and machine intelligence, 1994, 16(6): 641-647.ANGULO J, JEULIN D. Stochastic watershed segmentation [C]//8th ISMM 2007. Rio de Janeiro: ISMM, 2007: 265-276

[3] Koller D, Friedman N. Probabilistic graphical models: principles and techniques[M]. MIT press, 2009.

[4] Rother C, Kolmogorov V, Blake A. Interactive foreground extraction using iterated graph cuts, 2004[C]//SIGGRAPH. 2004.

[5] Shotton J, Johnson M, Cipolla R. Semantic texton forests for image categorization and segmentation[C]//2008 IEEE conference on computer vision and pattern recognition. IEEE, 2008: 1-8.

[6] Mahapatra D. Analyzing training information from random forests for improved image segmentation[J]. IEEE Transactions on Image Processing, 2014, 23(4): 1504-1512.

[7] Aloysius N, Geetha M. A review on deep convolutional neural networks[C]//2017 international conference on communication and signal processing (ICCSP). IEEE, 2017: 0588-0592.

[8] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[9] Yu H, Yang Z, Tan L, et al. Methods and datasets on semantic segmentation: A review[J]. Neurocomputing, 2018, 304: 82-103.

[10] Xuan T, Liang W, Qi D. Review of image semantic segmentation based on deep learning[J]. J. Softw, 2019, 30(2): 440-468.

[11] ZHAO H, SHI J, QI X. Pyramid scene parsing network; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition[J]. 2017.

[12] Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. arXiv preprint arXiv:1412.7062, 2014.

[13] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 834-848.

[14] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv preprint arXiv:1706.05587, 2017.

[15] Yu H, Yang Z, Tan L, et al. Methods and datasets on semantic segmentation: A review[J]. Neurocomputing, 2018, 304: 82-103.

[16] Mo Y, Wu Y, Yang X, et al. Review the state-of-the-art technologies of semantic segmentation based on deep learning[J]. Neurocomputing, 2022, 493: 626-646.

[17] Hao S, Zhou Y, Guo Y. A brief survey on semantic segmentation with deep learning[J]. Neurocomputing, 2020, 406: 302-321.

[18] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.

[19] Strudel R, Garcia R, Laptev I, et al. Segmenter: Transformer for semantic segmentation [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 7262-7272.

[20] Xie E, Wang W, Yu Z, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers[J]. Advances in Neural Information Processing Systems, 2021, 34: 12077-12090.

[21] L. Zhu, J. Suomalainen, J. Liu, J. Hyypp¨a, H. Kaartinen, H. Haggren, et al., A review: Remote sensing sensors, Multi-purposeful application of geospatial data (2018) 19–42.

[22] M. Schmitt, J. Prexl, P. Ebel, L. Liebel, X. X. Zhu, Weakly super- vised semantic segmentation of satellite images for land cover mapping– challenges and opportunities, arXiv preprint arXiv:2002.08254 (2020).

[23] L. P. Olander, H. K. Gibbs, M. Steininger, J. J. Swenson, B. C. Murray, Reference scenarios for deforestation and forest degradation in support of redd: a review of data and methods, Environmental Research Letters 3 (2) (2008) 025011.

[24] F. Pacifici, F. Del Frate, C. Solimini, W. J. Emery, An innovative neural-net method to detect temporal changes in high-resolution op- tical satellite imagery, IEEE Transactions on Geoscience and Remote Sensing 45 (9) (2007) 2940–2952.

[25] A. Boguszewski, D. Batorski, N. Ziemba-Jankowska, T. Dziedzic, A. Zambrzycka, Landcover. ai: Dataset for automatic mapping of build- ings, woodlands, water and roads from aerial imagery, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog- nition, 2021, pp. 1102–1110.

[26] L. P. Osco, J. M. Junior, A. P. M. Ramos, L. A. de Castro Jorge, S. N. Fatholahi, J. de Andrade Silva, E. T. Matsubara, H. Pistori, W. N. Gon¸calves, J. Li, A review on deep learning in uav remote sensing, International Journal of Applied Earth Observation and Geoinformation 102 (2021) 102456.

[27] L. Ding, D. Lin, S. Lin, J. Zhang, X. Cui, Y. Wang, H. Tang, L. Bruz- zone, Looking outside the window: Wide-context transformer for the semantic segmentation of high-resolution remote sensing images, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–13.

[28] S. D. Olabarriaga, A. W. Smeulders, Interaction in the segmentation of medical images: A survey, Medical image analysis 5 (2) (2001) 127–142.