IJPREMS

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 04, Issue 01, January 2024, pp : 443-449

e-ISSN : 2583-1062

Impact Factor : 5.725

A REVIEW PAPER ON RECENT ADVANCES IN SEMANTIC SEGMENTATION

DR. Mohamed Raf1i[1], Swathi M[2], Meghana B Madlur[3]

[1,2,3]Department of Computer Science and Engineering, UBDTCE, India.

ABSTRACT

Semantic image segmentation is a crucial aspect of image processing and computer vision, finding applications in various domains such as medicine and intelligent transportation. This paper reviews both traditional and recent Deep Neural Network (DNN) methods for semantic segmentation. Traditional methods and datasets are briefly summarized, followed by a comprehensive investigation of DNN methods across eight aspects. These aspects include fully convolutional networks, upsampling techniques, joint methods with Conditional Random Fields (CRF), dilated convolution approaches, advancements in backbone networks, pyramid methods, multi-level and multi-stage methods, and various supervised, weakly-supervised, and unsupervised techniques.

Keywords: DNN, CNN, image semantic segmentation

1. INTRODUCTION

Semantic image segmentation, also known as pixel-level classification, is a fundamental task in computer vision that involves clustering image parts belonging to the same object class (Thoma 2016). It stands apart from image-level classification, where each image is treated as a single category, and object detection, which involves localizing and recognizing objects. Semantic image segmentation can be considered as pixel-level it classifies each pixel into its corresponding category. Additionally, there is a task called instance segmentation, which combines object detection and segmentation (Lin et al. 2014; Li et al. 2017a).

The applications of semantic image segmentation are diverse, ranging from detecting road signs (Maldonado-Bascon et al. 2007) and segmenting colon crypts (Cohen et al. 2015) to land use and land cover classification (Huang et al. 2002). In the medical field, it finds applications in detecting brains and tumors (Moon et al. 2002) and tracking medical instruments during operations (Wei et al. 1997). Advanced Driver Assistance Systems (ADAS) and self-driving cars heavily rely on scene parsing, which, in turn, depends on semantic image segmentation (Fritsch et al. 2013; Menze and Geiger 2015; Cordts et al. 2016).

With the resurgence of Deep Neural Networks (DNN), segmentation accuracy has seen significant improvement. Traditional segmentation methods, predating DNN, are briefly reviewed in this paper, with a focus on recent progress achieved through the adoption of DNN. The paper also presents a survey of datasets used in image segmentation and discusses evaluation metrics.

4 1. January 2024 January

The organization of the paper includes a review of semantic image segmentation datasets and evaluation metrics in Section 2, a brief summary of traditional methods in Section 3, and an introduction to recent progress in Section 4. Finally, Section 5 provides a brief summary of the research, presumably discussing findings and potential future directions.

The research of semantic segmentation, aiming to assign semantic labels to each pixel, is foundational in computer vision, applicable to augmented reality devices, autonomous driving, video surveillance, and more. Recent algorithms for real-time semantic segmentation employ three main approaches to accelerate models: (1) restricting input size, (2) channel pruning, and (3) network stage dropping (ENet). However, these approaches often compromise accuracy for speed, making them suboptimal in practice. The next segment of the paper delves into related work, exploring the evolution of convolutional networks in visual recognition problems, including image classification, object detection, and semantic segmentation. The transition from coarse to fine inference is discussed, highlighting the need for pixel-wise prediction. Fully Convolutional Networks (FCNs) emerge as a pivotal approach, enabling dense prediction and learning from supervised pre-training. The shortcomings of traditional FCN-based methods, such as limited receptive fields and loss of detailed structures, lead to the introduction of deep deconvolution networks.

The paper introduces a novel strategy for semantic segmentation based on CNNs, leveraging a deep deconvolution network comprising deconvolution, unpooling, and ReLU layers. The network is applied to individual object proposals, enabling instance-wise segmentations. The proposed approach surpasses existing FCN-based methods in accuracy and performance. The paper concludes with a comprehensive organization, summarizing contributions and outlining the subsequent sections. Semantic segmentation remains a challenging computer vision problem with applications in various domains. The paper presents a transformer-based approach, departing from traditional convolutional architectures, to address the limitations of local interactions in capturing global image context. Transformers, known for their success in natural language processing, are adapted to the sequence-to-sequence problem of semantic segmentation.The proposed method, termed Segmenter, splits the image into patches and utilizes transformer architecture for both encoding and decoding stages. The approach achieves state-of-the-art results on challenging datasets, outperforming convolution-based methods. The paper contributes to the field by introducing a transformer-based paradigm for semantic segmentation, emphasizing global context at every layer of the model.

In the related work section, the evolution of semantic segmentation methods is discussed, with a focus on Fully Convolutional Networks (FCNs). Recent advances incorporate attention mechanisms to capture long-range dependencies, but limitations persist due to the inherent bias towards local interactions. The proposed transformer-based approach aims to overcome these limitations, presenting a pure transformer architecture that captures global context during both encoding and decoding stages.

2. LITERATURE SURVEY

It encapsulates the evolution of image segmentation methodologies, spanning from classical techniques to recent advancements, with a focus on the prominent datasets, evaluation metrics, and the progression of deep learning models. Notably, datasets like PASCAL VOC, MS COCO, ADE20K, Cityscapes, and KITTI have played pivotal roles in shaping segmentation research. Evaluation metrics, including pixel accuracy, mean accuracy, region intersection upon union, and frequency-weighted IU, offer a comprehensive framework for assessing model performance. Traditional methods, predating the era of deep neural networks, leveraged techniques such as thresholding, clustering, edge-based detection, support vector machines, and Markov Random Networks for segmentation.

The literature review unfolds the paradigm shift brought about by Fully Convolutional Networks (FCNs), which revolutionized image segmentation by replacing fully connected layers with convolutional layers. The utilization of up-sample methods, dilated convolutions, and backbone networks like VGG, ResNet, and ResNeXt has become integral to the modern segmentation landscape. Pyramid strategies, encompassing image pyramids, atrous spatial pyramid pooling, pooling pyramids, and feature pyramids, have emerged as effective tools for multi-scale context integration.

The U-shape structure in segmentation networks, employing deconvolution layers, skip connections, Laplacian Pyramid Reconstruction Network, multi-path refinement, and channel attention blocks, signifies a powerful trend in capturing and refining spatial information. Context information, essential for semantic segmentation, is systematically addressed through dilation, pyramid pooling, large kernels, and specialized modules like ASPP and PSP. The integration of attention mechanisms, spanning scale-dependent and channel attention, further refines feature extraction and recognition.

The review extends to real-time segmentation challenges, highlighting strategies for achieving speed without compromising accuracy. The proposed Bilateral Segmentation Network (BiSeNet) stands out as an exemplar, presenting a lightweight model that balances spatial information, receptive field, and context. The introduction of the Spatial Path and Context Path, complemented by the Attention Refinement Module (ARM) and Feature Fusion Module (FFM), underscores the iterative development and ablation studies conducted for performance enhancement.

In terms of background and motivation, the survey draws inspiration from the success of deep neural networks in image classification, emphasizing the transition from convolutional networks to FCNs. Dense prediction problems, such as segmentation, restoration, and depth estimation, are addressed through adaptations of classic classification nets like LeNet, AlexNet, and VGGNet. Techniques like shift-and-stitch and upsampling are explored for obtaining refined predictions, with a comparative analysis of patchwise versus whole-image training strategies.

The survey delves into the experimental landscape, with FCNs fine-tuned for segmentation and skip connections introduced for improved spatial precision. Results from the PASCAL VOC 2011 segmentation challenge showcase the effectiveness of FCN-VGG16, particularly with additional training data. Training details, including stochastic gradient descent with momentum, and considerations for class distribution balance and augmentation techniques, further contribute to the comprehensive exploration of the FCN paradigm. The literature survey culminates in the discussion of weakly supervised settings, system architecture, training challenges, inference processes, and the experimental evaluation of proposed networks on benchmark datasets.

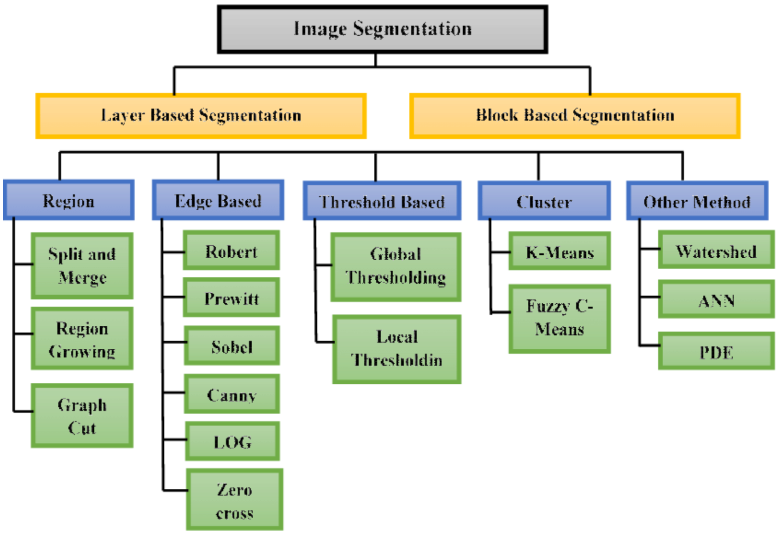

Figure.1

Layer-based semantic segmentation typically involves the use of deep neural networks, such as convolutional neural networks (CNNs). In this approach, the network consists of multiple layers that learn hierarchical representations of the input image. Each layer extracts features at different levels of abstraction, capturing both low-level details and high-level semantics. The final layer produces a pixel-wise classification map, where each pixel is assigned a label corresponding to a specific object or class.

On the other hand, block-based semantic segmentation divides the image into non-overlapping blocks or regions and classifies each block independently. This approach is often used in combination with traditional machine learning algorithms or shallow neural networks. Each block is processed separately, and the resulting classifications are combined to create the final segmentation map. While this method may be computationally less intensive compared to layer-based approaches, it may struggle with capturing intricate details and context across different regions.

Both layer-based and block-based semantic segmentation methods have their advantages and limitations. Layer-based approaches, especially with deep neural networks, tend to excel at capturing intricate spatial dependencies and contextual information, making them suitable for complex scenes. However, they can be computationally expensive and may require substantial computational resources. Block-based approaches, on the other hand, are computationally more efficient but may struggle with capturing fine details and context.

3. METHODOLOGY

By performing semantic segmentation, the original image is transformed into a detailed map where each pixel is assigned a specific class label, such as identifying objects, boundaries, or regions. This process not only enhances the interpretability of visual data but also provides crucial information for subsequent tasks like object recognition, scene understanding, and autonomous navigation. Semantic segmentation plays a pivotal role in fields like medical imaging, where it aids in identifying and delineating anatomical structures, as well as in autonomous vehicles, where it contributes to real-time environmental perception. The ability to convert actual images into semantically segmented ones empowers machines to comprehend and interact with their visual surroundings, facilitating advancements in diverse domains and paving the way for more sophisticated and context-aware artificial intelligence systems.

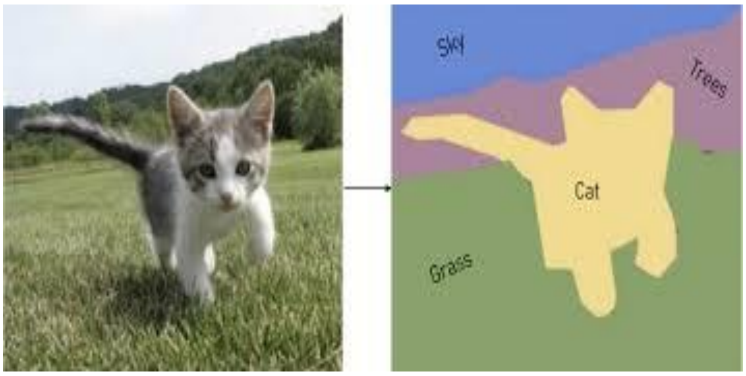

Figure.2

Actual image SemanticSegmented image Converting an actual image into a semantic segmented image is a task often associated with computer vision and image processing. Semantic segmentation involves classifying each pixel in an image into specific categories or classes, such as identifying objects, boundaries, or regions with similar characteristics. The understanding is crucial for applications ranging from image editing and virtual reality to industrial automation. In essence, the conversion of actual images into semantic segmented images is a pivotal step toward endowing machines with the ability to interpret and respond to visual information in a manner akin to human perception, thereby advancing the capabilities of numerous artificial intelligence and computer vision applications.

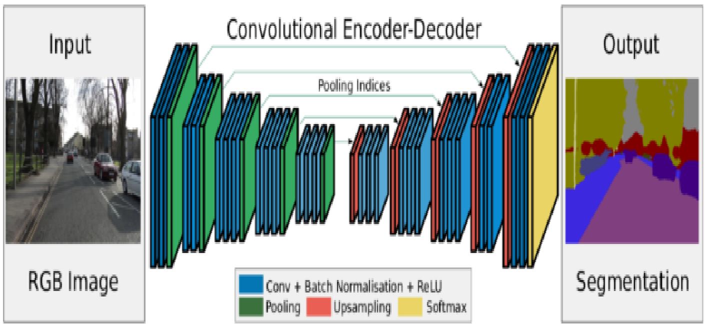

Figure 3: Fully Convolution Network

Compared with traditional statistical methods, methods based on CNNs have natural advantages. They can automatically extract semantic features, so they can replace manual feature extraction in original methods. CNNs have end-to-end processing structures, which take images as an input and generate pixel-level labels through multiple layers of convolution and deconvolution. The fully convolutional networks (FCNs) proposed by Long et al. were the first application of CNN in semantic segmentation. FCNs perform pixel-level classification of images. CNNs have achieved great success in semantic segmentation tasks. They use fully connected layers in convolutional layers to obtain fixed-length feature vectors. Meanwhile, FCNs accept images of any size and use deconvolution layers to upsample the feature maps of the last convolutional layer and restore them to the same size as the input image. Subsequently, predictions are generated for each pixel while retaining the spatial information of the original input image. Pixel classification is then performed on the upsampled feature maps. FCN network structure as shown in Figure 2. It encompasses the evolution and future directions of semantic segmentation technology, with a specific focus on encoder–decoder systems, skip connections, spatial pyramid pooling, dilated convolutions, knowledge distillation, domain adaptation, and few-shot/zero-shot semantic segmentation. The development trend of encoder–decoder architectures, such as Visual Transformers, is discussed, emphasizing the need for adaptive feature fusion, task-oriented encoder design, dynamic decoder optimization, integration of Vision Transformers (ViTs), guiding segmentation with prior knowledge, adaptive domain adaptation, and multimodal information integration. Skip connections are highlighted as crucial for combating gradient vanishing and exploding problems, with a call for scientific determination of optimal positions and quantities. Spatial pyramid pooling is introduced as a solution for real-time segmentation, with future directions focusing on optimization, multiscale feature fusion, and attention mechanisms. Dilated convolutions are recognized for expanding receptive fields, with research directions including addressing checkerboard effects, combining with other operations, and practical applications. Knowledge distillation is presented as a means to reduce model size and complexity, with potential enhancements through adaptive methods, task-driven approaches, weakly supervised learning, and integration with model architecture search. Domain adaptation is discussed as a critical challenge, with proposed future research directions including domain-specific information transfer, adversarial training, generative models for data augmentation, and the incorporation of unsupervised or weakly supervised learning. Lastly, few-shot/zero-shot semantic segmentation is acknowledged for its practicality in data-scarce scenarios, with ongoing research focusing on weakly supervised variants and improved generalization abilities through multimodal fusion and knowledge-graph-based approaches. Semantic image segmentation in deep learning relies heavily on Convolutional Neural Networks (CNNs), serving as the foundational architecture for cutting-edge models. These networks leverage convolutional layers to extract hierarchical features from input images, a crucial step in discerning semantic information. Fully Convolutional Networks (FCNs) extend the capabilities of CNNs by preserving spatial information through the use of convolutional and pooling layers, replacing fully connected layers with 1x1 convolutions to maintain spatial dimensions. The prevalent encoder-decoder architectures in semantic segmentation involve the extraction of hierarchical features by the encoder and the subsequent upsampling of these features by the decoder to generate detailed segmentation maps. Skip connections play a vital role in retaining fine-grained details during the upsampling process, connecting encoder and decoder at various levels to combine low-level and high-level features effectively. U-Net, a widely adopted architecture, exhibits a U-shaped design with a contracting path (encoder) and an expansive path (decoder), proving particularly effective for semantic segmentation tasks. The DeepLab series introduces atrous (dilated) convolutions to capture multi-scale contextual information, with DeepLabv3 incorporating atrous spatial pyramid pooling (ASPP) for enhanced segmentation. Mask R-CNN, an extension of Faster R-CNN, integrates a pixel-level segmentation branch, featuring a Region Proposal Network (RPN) for object detection alongside a mask branch. Cross-entropy loss serves as a common choice for pixel-wise classification in semantic segmentation, complemented by additional loss functions like Dice loss or IoU to further refine segmentation accuracy. Due to limited annotated data availability, data augmentation is crucial for training robust models, incorporating techniques such as rotation, flipping, scaling, and adjustments in brightness and contrast. Transfer learning plays a pivotal role, utilizing pre-trained models on extensive datasets like ImageNet. Fine-tuning these models for specific segmentation tasks often leads to improved performance. Post-processing techniques, such as conditional random fields (CRFs), contribute to refining segmentation masks for enhanced accuracy. The applications of semantic segmentation are diverse, spanning autonomous vehicles, medical image analysis, satellite imagery, and more. As the field continues to evolve, researchers explore innovative architectures and techniques, requiring a delicate balance between computational resources, dataset size, and application-specific requirements in the implementation and fine-tuning of these models. Semantic segmentation aims to classify and assign a specific label to each pixel in an image, grouping them into meaningful segments or regions based on their content, such as distinguishing between different object classes or background elements. It provides a detailed understanding of the scene by segmenting it into semantically meaningful parts. On the other hand, object detection focuses on locating and classifying multiple objects within an image, providing bounding boxes around each object along with their corresponding class labels. While semantic segmentation offers pixel-level accuracy for understanding the overall scene, object detection is more focused on identifying and localizing individual objects, making it suitable for tasks where precise object boundaries and locations are crucial, such as in autonomous vehicles or video surveillance. Semantic segmentation involves assigning a specific class label to each pixel in an image, enabling a detailed understanding of the scene by segmenting it into semantically meaningful parts. This pixel-level accuracy is valuable in applications like image editing, medical image analysis, and autonomous driving, where precise delineation of objects is crucial. Object detection, on the other hand, identifies and localizes multiple objects within an image by providing bounding boxes and class labels. It's widely used in applications like video surveillance, object tracking, and robotics.

Advantages and Disadvantages of segmentation technique.

Segmentation technique	Desscription	Advantages	Disadvantages
Thresholding method	Based on histogram peaks of image to find particular threshold values	No need of previous information, simplest method	Highly dependent on peaks, spatial details are not considered
Edge based method	Based on discontinuity direction	Good for images having better contrast between objects.	Not suitable for wrong detected or too many edges
Region based methods	Basesd on partitioning image into homogeneous regions	More immense to noise, useful when it is easy to define similarity criteria	Expensive method in terms of time and memory
Clustering method	Based on division into homogeneous regions	Fuzzy uses partial membership therefore more useful for real problems	Determining membership function is not easy
Watershed method	Based on topological interpretation	Results are more stable,detected boundaries are continuous	Complex calculation of gradients.
PDE based method	Based on the working of differential equations	Fastest method, best for time critical applications	More computational complexity
ANN based method	Based in the simulation of learning process for decision making	No need to write complex programs	More wastage of time in training

4. RESULTS

The trajectory of semantic segmentation technology is advancing along multiple fronts, with a concerted emphasis on refining encoder-decoder systems, incorporating skip connections, leveraging spatial pyramid pooling, integrating dilated convolutions,employing knowledge distillation, addressing domain adaptation challenges, and exploring the realm of few-shot/zero-shot semantic segmentation. Encoder-decoder architectures,exemplified by Visual Transformers, are evolving towards adaptive feature fusion, task-oriented encoder design, and dynamic decoder optimization.

The integration of Vision Transformers (ViTs) and the strategic use of prior knowledge in guiding segmentation underscore the quest for more efficient and context-aware models. Skip connections play a pivotal role in mitigating gradient vanishing and exploding issues, urging for a scientific determination of their optimal placements and quantities.

Spatial pyramid pooling, introduced for real-time segmentation, is poised for further optimization, multiscale feature fusion, and the incorporation of attention mechanisms. Dilated convolutions, acknowledged for expanding receptive fields, are under ongoing research to address challenges such as checkerboard effects and to explore practical applications. Knowledge distillation emerges as a technique not only to reduce model size but also for potential enhancements through adaptive methods, task-driven approaches, and integration with model architecture search. The critical challenge of domain adaptation is being tackled with directions including domain-specific information transfer, adversarial training, generative models for data augmentation, and exploration of unsupervised or weakly supervised learning. Furthermore, the practicality of few-shot/zero-shot semantic segmentation in data-scarce scenarios is being explored, with ongoing research emphasizing weakly supervised variants and improved generalization capabilities through multimodal fusion and knowledge-graph-based approaches. These collective efforts signal a dynamic and innovative landscape in the continuous evolution of semantic segmentation methodologies.

5. CONCLUSION

In conclusion, this paper provides a thorough exploration of semantic image segmentation, examining both traditional and contemporary Deep Neural Network (DNN) methods. The importance of semantic segmentation in fields like medicine and intelligent transportation is underscored, motivating the need for accurate and efficient segmentation techniques. The review begins by summarizing traditional segmentation methods and datasets, paving the way for a detailed investigation of DNN methods across various dimensions. These dimensions include fully convolutional networks, upsampling techniques, joint methods with Conditional Random Fields (CRF), dilated convolution approaches, advancements in backbone networks, pyramid methods, multi-level and multi-stage methods, and a spectrum of supervised, weakly-supervised, and unsupervised techniques. This comprehensive analysis provides a holistic view of the advancements in semantic segmentation. The paper organizes its content systematically, starting with a review of datasets and evaluation metrics, followed by a concise summary of traditional methods. Recent progress is then explored, emphasizing the evolution of convolutional networks and the emergence of Fully Convolutional Networks (FCNs). Recognizing the limitations of FCN-based methods, the introduction of deep deconvolution networks signifies a pivotal advancement.

6. REFERENCES

[1] Recent progress in semantic image segmentation Xiaolong Liu1 · Zhidong Deng1 · Yuhan Yang2 © The Author(s) 2018

[2] Segmenter: Transformer for Semantic Segmentation. Robin Strudel*, Inria ,Ricardo Garcia* Inria†, Ivan Laptev Inria Cordelia Schmid Inria†.

[3] BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. National Key Laboratory of Science and Technology on Multispectral Information Processing, School of Automation, Huazhong University of Science & Technology, China

[4] Fully Convolutional Networks for Semantic Segmentation Jonathan Long∗ Evan Shelhamer∗ Trevor Darrell UC Berkeley

[5] Learning Deconvolution Network for Semantic Segmentation Hyeonwoo Noh Seunghoon Hong Bohyung Han Department of Computer Science and Engineering, POSTECH, Korea

[6] D Semantic Segmentation: Recent Developments and Future Directions by Yu Guo Guigen Nie Wenliang Gao and Mi Liao

[7] Analysis of Various Image Segmentation Techniques for Flower Images Isha Patel, Sanskruti Patel Faculty of Computer Science and Applications, Charotar University of Science and Technology, Changa, India .