Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Advancements in Semantic Segmentation Using Deep Learning Approaches

Chenhuan Ni *

Wenzhou No.22 Senior Middle School, Wenzhou, 325000, China

* Corresponding Author Email: nichenhuan@outlook.com

Abstract. Semantic segmentation plays an important part in computer vision by assigning semantic labels to each pixel in an image. Many models have been created to enhance segmentation performance as deep learning develops, from traditional models CNNs, FCNs, and U-Net, to the advanced models DeepLab series, GANs, and Transformer-based models. This paper offers a comprehensive overview of semantic segmentation methods based on deep learning. It begins with an introduction to the task and related background concepts, followed by a review of traditional and advanced models. A summary of current issues, including labeled data scarcity, complex object boundary definition, poor model generalization, and multi-scale information handling, is also provided to assist in future study and real-world application. The study also provides a comparative evaluation of the models' computational and performance characteristics. The overall goal of this work is to give a clear overview of the progress made in semantic segmentation using deep learning.

Keywords: Semantic segmentation; deep learning; computer vision.

1. Introduction

A fundamental issue in computer vision is semantic segmentation, which gives each pixel in an image a semantic label to allow for fine-grained scene analysis. Numerous real-world applications rely on it, such as autonomous driving, where it facilitates lane and object detection; medical imaging, where it helps segment organs or diseases; and remote sensing, where it allows for accurate land cover classification. Segmentation accuracy and robustness have significantly improved because of the transition from conventional image processing methods to deep learning. Subsequent advancements like Transformer-based models, DeepLab, and GANs have greatly enhanced segmentation performance [1-3]. DeepLab enhances boundary precision using atrous convolution, GANs refine segmentation through adversarial learning, and Transformers capture global context via self-attention mechanisms.

Despite these advancements, semantic segmentation still faces several challenges. The scarcity of labeled data—especially in specialized domains like medical imaging—hinders the training of robust models. Class imbalance is another issue, as some categories are underrepresented in training datasets, which typically leads to biased model predictions. Furthermore, it is still challenging to precisely define the boundaries of complex objects, especially when objects are small, overlap, or have different scales. In addition, models trained on specific datasets frequently suffer from poor generalization, struggling to maintain performance when deployed in different environments or under varying conditions. Lastly, it is essential but challenging to handle multi-scale information—where objects vary significantly in size and shape between images.

Reviewing recent developments in deep learning-based semantic segmentation and exploring how they overcome the limitations of conventional techniques are the objectives of this research. This study aims to give researchers in computer vision and related fields guidance by reviewing both foundational and cutting-edge models. It also provides guidance for the future development of more effective and better segmentation frameworks.

2. Early Deep Learning Models in Semantic Segmentation

2.1. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) have played an important part in the development of semantic segmentation [4]. CNNs are a kind of deep learning models that were created especially for visual data. The models are suitable for image-based applications because they use convolutional operations to extract spatial hierarchies of information.

CNNs primarily comprise three types of layers: (i) Convolutional layers, which extract features by convolving a kernel (or filter) of weights. (ii) Nonlinear layers, which apply an activation function on feature maps (typically element-wise) to allow the network to model non-linear functions. (iii) Pooling layers, which reduce spatial resolution by replacing a small neighborhood of a feature map with some statistical information about the neighborhood [5]. As the network deepens, these components cooperate to capture local spatial patterns and increasingly abstract information. CNNs' capacity to share weights via convolutional kernels is one of its main advantages; this allows for translational invariance in image processing and significantly reduces the number of parameters when compared to fully connected networks [6].

With architectures like AlexNet, VGGNet, and ResNet setting new standards, CNNs have shown outstanding performance in image classification tasks [7-9]. These architectures have not only pushed the limits of accuracy but also inspired adaptations for dense prediction problems where pixel-level precision is needed— such as in semantic segmentation. Semantic segmentation gives each pixel a class label, in contrast to classification tasks that produce a single label per image. This requires models that maintain spatial information across the depth of the network. Although they are quite good at extracting high-level semantic characteristics, standard CNNs have difficulty with dense prediction because they decrease spatial resolution through pooling operations. Nevertheless, the powerful feature extraction capability of CNNs became the foundation of many segmentation models.

CNN's efficiency, scalability, and strong inductive biases (like locality and shift invariance) continue to make them crucial in both research and real-world applications, and they are still used as the basis encoders in many cutting-edge segmentation frameworks, or combined with other modules to balance local detail capture and global context modeling, even as newer architectures like Transformers are gaining traction.

2.2. Fully Convolutional Networks

Fully Convolutional Networks (FCNs), proposed by Long, Shelhamer, and Darrell in 2015, represent an important step in the field of semantic segmentation by adapting traditional convolutional neural networks (CNNs) for dense pixel-level prediction tasks [10]. CNNs were mostly employed for image classification tasks before FCNs, with a single label for the entire image serving as the output. However, a new architecture that maintained the spatial resolution across the network was needed for semantic segmentation, where the objective is to give each pixel in the image a class label.

The primary architectural evolution of FCNs is a replacement of convolutional layers for the fully connected layers present in conventional CNNs. This enables dense pixel-wise predictions in segmentation tasks. Convolutional, pooling, and upsampling layers are the three primary parts of the network. From raw picture pixels, convolutional layers extract hierarchical feature representations. Each filter identifies local patterns, which are then gradually incorporated in deeper layers to create more abstract information. In order to concentrate on higher-level features, pooling layers—like those in conventional CNNs—reduce the spatial dimensions of the feature maps; nevertheless, this lowers spatial resolution, which is crucial for segmentation. FCNs use upsampling layers, also known as deconvolution layers, to recover this lost resolution. This allows for precise pixel-by-pixel segmentation predictions by mapping the low-resolution feature maps back to the original picture dimensions.

A key modification in FCNs, compared to traditional CNNs, is the introduction of skip connections. These connections bridge the feature maps from earlier layers with those from deeper layers. By combining rich semantic context with fine spatial detail, this combination of low- and high-level characteristics enhances segmentation accuracy, particularly around boundaries and small objects. Through its hierarchical structure, FCNs also make use of multi-scale information. Deeper layers capture more abstract information, while lower layers catch finer-grained data. This multi-scale method works especially well when dealing with items of different sizes in an image.

A wide range of practical applications have been significantly impacted by FCNs. They are employed for tasks including obstacle avoidance, vehicle detection, and road segmentation in autonomous driving. To help with precise diagnosis and treatment planning, FCNs help segment organs, tumors, and other anatomical structures from medical scans. Better environmental monitoring and analysis are made possible using FCNs in remote sensing for land cover classification and feature extraction from satellite data.

2.3. U-Net and Encoder-Decoder Networks

U-Net, introduced by Ronneberger et al. in 2015, is a deep learning architecture specifically designed for biomedical image segmentation tasks [11]. The U-Net architecture follows an encoder-decoder structure, where the encoder captures high-level semantic information through a series of convolutional and pooling layers, while the decoder gradually recovers spatial resolution through upsampling and deconvolution layers. The distinctive characteristic of U-Net is its symmetric construction, which enables it to provide precise predictions at the pixel level while preserving spatial accuracy, even in small objects or objects with complex boundaries. To capture abstract features and minimize the spatial dimensions of the image, the encoder part of U-Net uses max-pooling after repeatedly applying convolutional layers. To restore the image's spatial resolution, the decoder component then upsamples the feature maps using deconvolution layers, sometimes referred to as transposed convolution.

One of the key innovations of the U-Net architecture is its use of skip connections, which directly pass feature maps from the encoder to the decoder. By combining these feature maps at appropriate layers, the network can preserve fine-grained spatial features that could otherwise be lost while downsampling. This enhances boundary localization and segmentation accuracy, particularly for tasks like organ segmentation or tumor detection in medical imaging, by enabling U-Net to integrate low-level characteristics, such as edges and textures, with high-level semantic information [11]. Compared to models like FCN, where skip connections are simpler, U-Net’s approach of concatenating feature maps enhances segmentation performance by better preserving spatial details and improving the network's ability to handle complex structures [11].

U-Net has been widely adopted in various segmentation tasks, particularly in medical imaging, where it effectively segments structures like organs, tumors, and blood vessels in MRI, CT scans, and ultrasound images. It is particularly useful for identifying cancers or lesions because of its capacity to handle small, irregularly shaped objects, where the regions of interest are frequently smaller than the surrounding structures [11]. U-Net has been used in remote sensing for urban planning and land cover categorization, where precise building, road, and vegetation segmentation is crucial. It is perfect for high-resolution segmentation tasks because of its capacity to catch minute details in intricate scenes. U-Net is also utilized in industrial and agricultural applications, including precision farming's crop segmentation and manufacturing's defect detection, which automate quality control and allow for the detection of minor flaws or plant illnesses from high-resolution photos.

3. Advanced Deep Learning Models in Semantic Segmentation

3.1. DeepLab series

The DeepLab series has been an important step in the development of semantic segmentation, famous for introducing methods that preserve spatial resolution while enhancing multi-scale context understanding [2]. The encoder-decoder structure in DeepLabv3+, atrous convolution, Atrous Spatial Pyramid Pooling (ASPP), and the earlier use of fully connected Conditional Random Fields (CRFs) for boundary refining are some of its key advancements.

Atrous convolution was one of the foundational techniques introduced by DeepLab [2]. By adding spaces between kernel pieces, this method increases the receptive field of filters rather than decreasing spatial resolution through pooling. The approach is straightforward but effective: capture a broader background while keeping small details. In segmentation tasks demanding both accuracy and structure, this approach enables the network to obtain more information without adding parameters or reducing the precision of the feature map.

Based on atrous convolution, DeepLabv2 introduced ASPP which applies multiple atrous convolutions with different dilation rates at the same time [12]. This setup makes it possible for the model to obtain information from many spatial scales, which proves critical when working with objects of varying sizes and complex backdrops. Global average pooling and more efficient feature merging techniques were included in later iterations, such as DeepLabv3, to further refine this approach and make the model more adaptable to fine detail and global context [13].

The third major improvement came with DeepLabv3+, where the architecture was extended to include a decoder component with the ASPP-enhanced encoder [13]. The segmentation results were much clearer thanks to this encoder-decoder arrangement, especially at edges and object boundaries. Rich semantic information is gathered by the encoder, while spatial details are reconstructed by the decoder through feature combination and upsampling. With this architectural decision, DeepLab's capacity to strike a balance between semantic depth and spatial accuracy is more in line with models such as U-Net.

3.2. Generative Models and GANs for Segmentation

Generative Adversarial Networks (GANs) have become a promising tool in semantic segmentation, particularly for tasks that demand precise structural detail and smooth region delineation [3]. A typical GAN consists of two neural networks: a generator, which attempts to produce realistic outputs, and a discriminator, which learns to distinguish between real and generated samples. The discriminator assesses how "realistic" the segmentation masks look with relation to ground truth annotations, and the generator predicts the segmentation masks. Conditional GANs (cGANs) enhance this setup by adding an input image as a condition, which enables the generator to produce more consistent and spatially accurate masks.

The capacity of GANs to provide more precise and refined outputs is the main reason behind their integration into segmentation frameworks. Conventional segmentation models frequently just use pixel-wise losses, like Dice loss or cross-entropy, which can lead to predictions that are too smooth or rough. GANs enable the generator to generate masks that show realistic structural patterns and match the ground truth by introducing adversarial loss. This results in improved spatial consistency and precise detail capture, particularly at complex object boundaries [14, 15].

Models based on Generative Adversarial Networks (GANs) demonstrate unique advantages in enhancing boundary quality. When the discriminator is trained to focus on local mask structures, it can effectively guide the generator to restore delicate or fragmented edges. In image segmentation tasks, edge processing constitutes a critical component, and this guidance mechanism significantly mitigates common issues such as jagged or discontinuous contours. Practical implementations of models like SegAN and GAN-UNet reveal that introducing adversarial training mechanisms improves the smoothness and continuity of segmentation results without requiring significant

architectural modifications. This characteristic endows GAN-based models with high practical value across real-world applications, whether in medical image analysis or general image segmentation tasks in computer vision domains.

These strengths prove particularly critical in demanding scenarios. Take segmenting small-sized or faint-colored objects, for instance: conventional algorithms often miss such targets, but GAN-based approaches reliably detect them. When processing blurry or noisy low-quality images, GANs utilize their image synthesis capabilities to produce segmentation results with clearer details and more complete shapes. This unique trait drives the growing adoption of GAN-enhanced segmentation in medical imaging scans and satellite image analysis—fields where razor-sharp boundary delineation remains a stringent requirement.

3.3. Transformer-Based Models

The Transformer architecture was initially designed for natural language processing but has gradually gained traction in the field of computer vision. Its core mechanism—self-attention—enables the model to capture global dependencies within input sequences. This ability is particularly valuable for vision tasks that require an understanding of the entire image context. The emergence of the Vision Transformer (ViT) marked a major turning point in this domain. Studies have shown that, when trained on sufficiently large datasets, Transformer-based models can match or even surpass convolutional neural networks in image classification tasks [16]. This breakthrough has sparked widespread interest, leading to the rapid development of Transformer-based models across dense prediction tasks such as object detection and semantic segmentation.

The Swin Transformer utilizes shifted window-based attention to enhance efficiency: it divides images into non-overlapping windows for local self-attention, reducing computational complexity from quadratic to linear. Shifted window mechanisms between layers enable cross-window interactions, capturing long-range dependencies without global computation. Its hierarchical architecture merges patches across stages, generating multi-scale features for tasks like semantic segmentation. This design ensures linear scaling with image resolution, maintaining high performance on dense prediction tasks while minimizing resource costs [17].

DETR redefines object detection and segmentation by framing the task as a direct set prediction problem [18]. Traditional hand-designed elements like region suggestions, anchor boxes, and non-maximum suppression are no longer required. Instead, DETR utilizes a transformer encoder-decoder architecture in which each output token predicts the bounding box and class of an object. By including mask heads for pixel-wise prediction, variations like DETR-based Panoptic Segmentation expand on this method for segmentation. This paradigm change simplifies the design process and demonstrates how effective transformers are at combining many vision tasks into a single framework.

4. Comparative Analysis of Semantic Segmentation Models

4.1. Accuracy vs. Efficiency Trade-Off

In semantic segmentation, model accuracy and computational efficiency are often at odds. DeepLab models, especially DeepLabv3+, have a high segmentation accuracy because of breakthroughs like ASPP and atrous convolution, however they are computationally expensive. Similar to this, transformer-based designs like Swin Transformer and DETR model global context to achieve competitive or better accuracy, although they usually take more memory and are slower during inference. FCNs and U-Net, on the other hand, generally strike a better balance between speed and performance, particularly in fields like medical imaging where real-time processing isn't necessarily required. Although GAN-based models raise edge accuracy and precision even more, their adversarial training makes them more difficult and could cause deployment delays. As a result, choosing a model frequently represents balancing accuracy with real-world limitations like processing time and hardware availability.

4.2. Model Complexity and Computational Requirements

Computational demands are significantly impacted by model architecture. As the foundations of segmentation networks, CNN-based models such as VGG or ResNet tend to be effective and have broad implementation support. Although having a simplified structure, FCNs and U-Net still need a substantial amount of resources, especially when deep encoders are used. DeepLab variations require additional GPU memory and training time because of their wide receptive fields and multi-scale modules. Due to their reliance on long-range attention processes that do not scale well with image quality, transformer-based models like ViT and DETR are particularly resource-intensive. Swin Transformer still outperforms CNN models in terms of computing load, but it does so by employing local attention within windows. GAN-based techniques are frequently lightweight in inference, they are computationally costly to optimize since they require dual-network training (generator and discriminator). As a result, deployment environments must be carefully considered when choosing an architecture.

4.3. Generalization across Domains

Different models exhibit different levels of adaptability when applied beyond their original training domains. Originally created for biological segmentation, U-Net has demonstrated effectiveness in low-data settings and adapts well to tasks such as industrial inspection and remote sensing [11]. Because of their hierarchical feature learning and modular structure, FCNs also generalize rather well across datasets. DeepLab can adapt to various domains and performs well in urban scene segmentation, however its multi-scale design may need to be adjusted. GAN-based models have shown resilience in low-quality or noisy data and are excellent at getting small details, which is useful in satellite analysis or medical imaging [15]. Despite their powerful representational capabilities, transformer-based models typically require extensive annotated datasets and fine-tuning to sustain performance across domains. While pretraining and hybrid methods are being investigated to increase resilience, this sensitivity to domain transition poses a challenge.

5. Challenges and Future Directions

5.1. Challenges in Semantic Segmentation Models

With the development of different models, several challenges also hinder the further progress. The absence of labeled data is one of the most persistent problems, especially in specialized fields like remote sensing and medical imaging. Large, high-quality datasets are necessary for deep learning model training in order for them to generalize effectively, but obtaining them is costly and time-consuming. This lack of labeled data limits the potential for training robust segmentation models, especially in rare or underrepresented classes.

Although many models are great at capturing global context and big structures, they sometimes have problems with precise boundary delineation and small object segmentation. Achieving pixel-perfect segmentation near object edges remains a major difficulty, even with methods like skip connections and adversarial loss in models such as U-Net and GAN-based approaches [11, 15].

The computational burden of high-performing models such as GANs, transformer-based architectures, and DeepLab is significant. These models are hard to implement in real-time applications or on edge devices with limited processing power because they need substantial amounts of memory, storage, and computing resources.

5.2. Future Directions in Semantic Segmentation

Semantic segmentation is a rapidly developing field, and there are several of promising directions that could improve performance and increase application.

In order to lessen reliance on big labeled datasets, self-supervised and weakly-supervised learning approaches are being considered as an appealing future direction. Weakly-supervised approaches use less thorough kinds of supervision to direct segmentation learning, and self-supervised learning can use unlabeled data to pre-train models.

It has been demonstrated that using multi-modal data improves segmentation accuracy, especially in domains like autonomous driving and remote sensing. Combining several data sources enables models to capture more robust and comprehensive features, which improves performance in difficult situations and enables better segmentation of complex environments.

Real-time segmentation is becoming increasingly important in applications such as robotics and autonomous driving, which is driving up demand for lightweight models that can operate well on devices with limited resources. Real-time deployment is made possible without sacrificing performance by models like SegFormer and other lightweight transformer-based models that try to find a balance between accuracy and computational efficiency [2].

6. Conclusion

This paper presents a comprehenzsive review of semantic segmentation models, from early convolutional neural networks to more recent approaches. The discussion includes models such as FCN, U-Net, DeepLab, GAN-based frameworks, and Transformer-based architectures. Through a structured analysis, the paper compares these models in terms of architecture, performance, and application adaptability. It also stresses typical problems in the field, such as the lack of data, computational expense, and cross-domain generalization. Possible future directions are also explored, including the creation of effective models for real-time tasks, multi-modal data fusion, and the use of self-supervised learning. Given the strong performance of deep learning in this area, it is expected that semantic segmentation will continue to evolve, with many new techniques and research emerging.

References

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017.

[2] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell., 2017, 40(4).

[3] Souly N, Spampinato C, Shah M. Semi Supervised Semantic Segmentation Using Generative Adversarial Network. Proc. IEEE Int. Conf. Comput. Vis., 2017.

[4] LeCun Y, Bottou L, Bengio Y, et al. Gradient - based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11).

[5] Minaee Shervin, Boykov Yuri, Porikli Fatih, et al. Image Segmentation Using Deep Learning: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(7).

[6] Biscione V, Bowers J. Learning translation invariance in cnns. arXiv preprint arXiv:2011.11757, 2020.

[7] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 2012.

[8] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[9] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[10] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[11] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer - Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5 - 9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015.

[12] Chen L - C, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. arXiv preprint, arXiv:1706.05587, 2017.

[13] Chen L - C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder - decoder with atrous separable convolution for semantic image segmentation. Proc. Eur. Conf. Comput. Vis. (ECCV), 2018.

[14] Luc P, Couprie C, Chintala S, Verbeek J. Semantic segmentation using adversarial networks. arXiv preprint arXiv:1611.08408, 2016.

[15] Xue Y, Xu T, Zhang H, Long L R, Huang X. SegAN: Adversarial network with multi - scale L1 loss for medical image segmentation. Neuroinformatics, 2018, 16(3 - 4).

[16] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, … Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the International Conference on Learning Representations (ICLR), 2021.

[17] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin Transformer: Hierarchical vision transformer using shifted windows. Proceedings of the International Conference on Computer Vision (ICCV), 2021.

[18] Carion N, Massa F, Synnaeve G, et al. End - to - End Object Detection with Transformers. Computer Vision – ECCV 2020, 16th European Conference, Glasgow, UK, August 23 – 28, 2020, Proceedings, Part I, 2020.

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