Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ24684

Semantic Segmentation in the Era of Foundation Models: Technical Evolution and Applications

Zhaolin Yu

Beijing University of Posts and Telecommunications, Beijing, China
zhuzaiyuhudie@gmail.com

Abstract. Semantic segmentation has undergone a remarkable transformation from traditional computer vision approaches to sophisticated deep learning architectures, culminating in the revolutionary capabilities introduced by foundation models. This comprehensive survey examines the technical progression of semantic segmentation methodologies, with particular emphasis on vision foundation models, such as the Segment Anything Model (SAM) and Contrastive Language-Image Pre-training (CLIP). This paper systematically analyzes how these large-scale pretrained models enable previously unattainable capabilities, including zero-shot learning and cross-domain generalization while identifying persistent challenges regarding computational efficiency and boundary precision. The investigation encompasses critical applications across medical imaging, remote sensing, and video understanding domains, revealing both transformative benefits and technical limitations. It concludes that foundation models represent a fundamental paradigm shift requiring hybrid approaches that effectively combine general capabilities with domain-specific optimizations.

Keywords: Semantic Segmentation, Foundation Models, SAM, CLIP, Zero-shot Learning

1. Introduction

Semantic segmentation, the task of assigning semantic labels to every pixel in an image, stands as a crucial computer vision task enabling detailed scene understanding for applications from autonomous driving to medical diagnosis. The field has evolved significantly over the past decade, progressing from traditional handcrafted feature-based approaches to deep learning architectures that substantially improved accuracy on standard benchmarks.

Fully convolutional networks marked an important milestone by enabling end-to-end learning for pixel-level classification, establishing a paradigm refined through various architectural innovations, including encoder-decoder structures and attention mechanisms. The recent introduction of vision foundation models represents the latest transformative shift in semantic segmentation research. Models, such as Segment Anything Model (SAM) and Contrastive Language-Image Pre-training (CLIP), have demonstrated capabilities for general-purpose segmentation with minimal task-specific training, offering new possibilities for addressing longstanding challenges, including limited annotated data and cross-domain generalization.

DOI: 10.54254/2755-2721/2025.BJ24684

This survey examines how foundation models fundamentally change the segmentation paradigm, analyzing their quantifiable benefits and limitations while exploring effective adaptation strategies for different application domains. We investigate the key technical barriers preventing widespread deployment and identify promising research directions. The analysis provides systematic technical evolution mapping from traditional methods to foundation models, quantitative performance assessment across paradigms, and examination of domain-specific applications in medical imaging, remote sensing, and video understanding. This study aims to guide researchers in understanding how foundation models are reshaping semantic segmentation techniques and workflows, while addressing challenges in efficiency, boundary precision, and domain adaptation.

2. Technical evolution of semantic segmentation

The evolution of semantic segmentation techniques can be understood through three major paradigms, each building upon the limitations and insights of its predecessors, where traditional computer vision approaches dominated the early years of the field, establishing fundamental concepts and revealing key challenges that would drive subsequent innovations through early semantic segmentation that relied heavily on carefully engineered features designed to capture relevant visual patterns and classical machine learning algorithms for classification. Texture-based methods represented one of the primary approaches during this era, employing descriptors such as Local Binary Patterns, Gabor filters, and Textons to characterize local surface properties, proving effective for segmenting natural textures like grass, water, or sky but struggling with structured objects and complex scenes where semantic understanding required more sophisticated reasoning, while color-based segmentation utilized color space transformations and histogram analysis to group pixels with similar chromatic properties, remaining computationally efficient but highly sensitive to illumination changes and unable to handle objects with similar colors but different semantic meanings [1]. Multi-scale feature extraction emerged as a key innovation during this period, recognizing that relevant visual patterns exist at different spatial scales through techniques like image pyramids and scale-space analysis that enabled extraction of features at multiple resolutions, improving robustness to scale variations that plagued single-scale approaches, while probabilistic and graph-based models provided more principled frameworks for incorporating spatial relationships into segmentation decisions [2]. Markov random fields and conditional random fields treated segmentation as an energy minimization problem, where the optimal labeling minimizes a global energy function combining data terms capturing pixel-level evidence with smoothness terms enforcing spatial consistency. The energy function typically takes the form:

$$E(S) = \Sigma_i \varphi_i(s_i) + \Sigma_{i,j} \psi_{ij}(s_i, s_j) \quad (1)$$

where φ_i represents unary potentials and ψ_{ij} represents pairwise potentials. Graph-cut algorithms provided efficient solutions for certain classes of energy functions, enabling pixel-level optimization of segmentation quality.

However, traditional methods faced fundamental limitations including manual feature design requiring extensive domain expertise, limited representational capacity for complex visual patterns, and poor computational scalability, ultimately leading to their displacement by deep learning approaches. The deep learning revolution transformed semantic segmentation through learnable feature representations and end-to-end optimization, with fully convolutional networks marking the paradigmatic shift by replacing fully connected layers with convolutional layers, enabling pixel-level learning while processing arbitrary-sized images [3]. FCNs introduced key innovations

DOI: 10.54254/2755-2721/2025.BJ24684

including transposed convolutions for learnable upsampling, skip connections for multi-scale feature fusion, and transfer learning from pretrained classification networks, with progressive refinement from FCN-32s to FCN-8s demonstrating the importance of combining features at different semantic levels. The encoder-decoder paradigm formalized systematic downsampling followed by upsampling for spatial recovery, with U-Net [4] emerging as the most influential design through symmetric skip connections that directly transfer features between corresponding encoder and decoder layers, establishing key principles of symmetric structure, skip connections for spatial preservation, and progressive channel expansion.

Subsequent innovations refined the encoder-decoder paradigm, with SegNet [5] introducing pooling indices for memory-efficient upsampling and DeepLabV3+ [6] combining encoder-decoder architecture with Atrous Spatial Pyramid Pooling for multi-scale context capture. Attention mechanisms further advanced the field, where channel attention from SENet [7] enabled selective feature emphasis and PSPNet [8] pioneered pyramid pooling modules for multi-scale context encoding. The adaptation of Transformer architectures opened new possibilities, with Vision Transformers [9] demonstrating competitive performance against CNNs and SETR [10] representing the first major ViT adaptation to segmentation through patch-based processing. Self-attention provides advantages over convolutions by adaptively focusing on relevant regions regardless of spatial distance, enabling superior long-range dependency modeling and naturally handling scale variations, thereby establishing the foundation for foundation models with unprecedented generalization capabilities.

3. Foundation models in semantic segmentation

Foundation models have introduced a paradigm shift in machine learning through large-scale pretraining on diverse datasets followed by downstream task adaptation, demonstrating remarkable generalization, zero-shot learning, and cross-domain transfer capabilities in computer vision. Contrastive language-image pre-training has revolutionized vision-language understanding by learning joint embeddings through contrastive learning on 400 million image-text pairs [11], enabling open-vocabulary classification for semantic segmentation through joint training of image and text encoders with key adaptations including dense feature extraction, text-visual correlation for pixel-level classification, and multi-scale processing. Methods like DenseCLIP [12] and LSeg [13] effectively adapt CLIP's representations for segmentation while preserving zero-shot capabilities, typically achieving 60%-80% of supervised performance without task-specific training. Complementing these advances, the Segment Anything Model [14] was trained on over 1 billion masks to develop universal segmentation capabilities, featuring a promptable interface that accepts points, boxes, masks, or text to generate high-quality segmentation masks, fundamentally changing segmentation system design through its architecture combining a Vision Transformer-based image encoder, a prompt encoder for various input types, and a lightweight transformer decoder, trained via a three-stage strategy involving assisted-manual labeling, semi-automatic mask proposal, and fully automatic segmentation development.

The adaptation of these foundation models employs various sophisticated techniques, where prompt engineering strategies involve strategic point placement and bounding box constraints for SAM, while CLIP-based segmentation utilizes descriptive text phrases and contextual information, with advanced methods employing learnable prompts optimized by neural networks. Parameter-efficient fine-tuning approaches such as Low-Rank Adaptation [15] introduce decomposition matrices reducing trainable parameters by 90% while maintaining performance, while adapter modules [16] insert lightweight networks between transformer layers and prompt tuning [17]

DOI: 10.54254/2755-2721/2025.BJ24684

optimizes continuous embeddings rather than discrete text. Domain-specific adaptation demonstrates versatility across specialized fields, exemplified by MedSAM [18] for medical applications and remote sensing implementations incorporating multi-spectral capabilities beyond RGB, with hybrid systems integrating foundation models with traditional architectures through feature extraction within established frameworks, coarse-to-fine pipelines, and multi-model ensembles combining complementary capabilities like CLIP's semantics with SAM's localization. Furthermore, open-vocabulary segmentation [19] enables arbitrary category segmentation through natural language descriptions by aligning pixel-level features with text embeddings in shared space, where region-based approaches often outperform direct pixel classification, while few-shot learning leverages foundation models through prototype-based methods, meta-learning for rapid adaptation, and in-context learning without parameter updates, enabling flexible deployment across diverse segmentation tasks with minimal data requirements.

4. Applications and deployment challenges

Foundation models have demonstrated transformative impact across critical application domains. Medical image segmentation benefits particularly from these models given the chronic shortage of annotated medical data. Medical adaptations show remarkable cross-modal generalization, with SAM variants trained on natural images transferring effectively to CT, MRI, and ultrasound modalities [18]. Interactive annotation reduces annotation time by 60%-80%, while few-shot adaptation achieves competitive performance with only 10-50 examples per anatomy.

However, medical imaging presents unique challenges, including multi-modal data integration, 3D volumetric processing, and sub-millimeter precision requirements for treatment planning. Performance analysis shows 85%-92% Dice coefficient for major organ segmentation and 70%-85% sensitivity for tumor detection, with a 5-10x reduction in annotation time.

Remote sensing represents another successful domain for foundation models, effectively addressing the challenge of vast scale variations from individual buildings to entire urban regions. Multi-city remote sensing studies have experienced explosive growth, with publications increasing from fewer than 10 per year in 2000 to over 200 per year by 2022. Recent advances in satellite sensor technologies have enabled comprehensive urban analysis, with 33 satellite sensors now available for multi-city studies distributed across spatial resolution (fine to coarse), data price (low to high), and revisit time (short to long) parameters. Applications demonstrate strong performance across diverse tasks, with studies incorporating anywhere from 2 to over 10,000 cities depending on the research scope. Technical innovations include multi-scale processing architectures that handle heterogeneous urban features, geographic domain adaptation techniques for different climatic and physiographic settings, and multi-source data fusion approaches that integrate optical imagery with synthetic aperture radar (SAR) and LiDAR data sources [20].

Video segmentation requires temporal consistency alongside spatial understanding. Frame-by-frame processing achieves 75%-80% temporal consistency efficiently, while recurrent architectures reach 85%-90% with 2-3x overhead, and 3D spatio-temporal processing attains 90%-95% consistency at 5-10x computational cost. Applications span autonomous driving, video editing, surveillance, and sports analysis.

Deployment faces critical technical challenges. Computational efficiency remains the primary barrier, with SAM requiring 2.4B parameters and 1-3 seconds inference versus 10-50ms for specialized models. Optimization strategies include model compression (quantization, pruning, distillation), efficient architectures like MobileSAM [21] (60x faster) andFastSAM [22], and hardware acceleration through specialized processors.

DOI: 10.54254/2755-2721/2025.BJ24684

Robustness limitations manifest as 10%-20% performance drops in domain shifts (natural to medical images), 5%-15% for indoor-outdoor transitions, and vulnerabilities to adversarial attacks. Enhancement strategies include [23] test-time adaptation (3-8% recovery) [24], uncertainty quantification, and multi-source training.

Boundary precision remains problematic, with foundation models achieving 65-75% boundary F1-score versus 80-90% for specialized architectures. Improvement strategies include multi-scale processing [25] (5-10% improvement), CRF-based refinement [26] (10-15% improvement), and hybrid architectures combining semantic understanding with boundary precision.

5. Conclusion

Foundation models have fundamentally transformed semantic segmentation by shifting from task-specific training to general-purpose adaptation, enabling unprecedented capabilities in zero-shot learning and open-vocabulary segmentation. The analysis reveals quantifiable benefits, including a 80%-90% reduction in annotation requirements for new domains, a 15%-30% improvement in cross-domain performance, and zero-shot segmentation capabilities for novel categories without task-specific training.

However, persistent challenges limit broader adoption. Computational overhead represents a significant barrier, with 10-100x increased requirements constraining deployment in resource-constrained environments and real-time applications. Boundary precision consistently degrades by 10%-20% compared to specialized architectures, particularly problematic for applications requiring precise object boundaries. Robustness issues under domain shift continue to affect performance when deployment conditions differ from training scenarios.

Critical research priorities include developing efficient foundation model architectures through sparse attention mechanisms and mixture of experts approaches, advancing multi-modal integration to handle temporal consistency in video sequences and 3D point clouds, and enabling continual learning capabilities for incremental adaptation without catastrophic forgetting. Domain-specific applications show particular promise, with medical foundation models incorporating anatomy-aware architectures and remote sensing models specialized for satellite imagery and environmental monitoring.

The most promising path forward involves thoughtful integration rather than wholesale replacement of traditional approaches. Hybrid architectures that combine foundation models with specialized components show particular potential, leveraging semantic understanding capabilities while employing specialized elements for boundary refinement and computational efficiency. Success requires efficient adaptation mechanisms that preserve general capabilities while enabling domain-specific optimization, balanced hybrid architectures that maintain both semantic understanding and boundary precision, and robust deployment strategies that handle real-world variations effectively.

As computational resources continue to advance and foundation models become more efficient through ongoing research, we anticipate continued breakthroughs that will expand semantic segmentation capabilities while addressing current limitations. The ultimate goal of achieving human-level understanding with computational efficiency and reliability has moved significantly closer to reality, though realizing this potential requires addressing the technical challenges identified in our comprehensive analysis.

DOI: 10.54254/2755-2721/2025.BJ24684

References

[1] Shi, J., Malik, J., & IEEE. (2000). Normalized cuts and image segmentation. In IEEE Transactions on Pattern Analysis and Machine Intelligence. No. 8; Vol. 22, pp. 888–888.

[2] Boykov, Y. Y., Jolly, M.P.. (2001). Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. Proceedings of "Internation Conference on Computer Vision, " pp.105–106.

[3] Long, J., Shelhamer, E., Darrell, T., & UC Berkeley. (n.d.). Fully convolutional networks for semantic segmentation.

[4] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-NET: Convolutional Networks for Biomedical Image Segmentation, May 18. arXiv.org. https: //arxiv.org/abs/1505.04597

[5] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2015). SEGNet: a deep convolutional Encoder-Decoder architecture for image segmentation, November 2. arXiv.org.

[6] Chen, L., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, February 7. arXiv.org.

[7] Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2017). Squeeze-and-Excitation networks, September 5. arXiv.org. https: //arxiv.org/abs/1709.01507

[8] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2016). Pyramid Scene Parsing network, December 4. arXiv.org. https: //arxiv.org/abs/1612.01105

[9] Dosovitskiy, A., Beyer, L., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, October 22. arXiv.org. https: //arxiv.org/abs/2010.11929

[10] Zheng, S., Lu, J., et al. (2020). Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers, December 31. arXiv.org.

[11] Radford, A., Kim, J. W., et al. (2021). Learning transferable visual models from natural language supervision, February 26. arXiv.org. https: //arxiv.org/abs/2103.00020

[12] Rao, Y., Zhao, W., et al. (2021). DenseCLIP: Language-Guided Dense Prediction with Context-Aware Prompting, December 2. arXiv.org. https: //arxiv.org/abs/2112.01518

[13] Li, B., Weinberger, K. Q., Belongie, S., Koltun, V., & Ranftl, R. (2022, January 10). Language-driven semantic segmentation. arXiv.org. https: //arxiv.org/abs/2201.03546

[14] Kirillov, A., Mintun, E., et al. (2023). Segment anything, April 5. arXiv.org.

[15] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LORA: Low-Rank adaptation of Large Language Models, June 17. arXiv.org.

[16] Houlsby, N., Giurgiu, A., et al. (2019). Parameter-Efficient Transfer Learning for NLP, February 2. arXiv.org. https: //arxiv.org/abs/1902.00751

[17] Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for Parameter-Efficient Prompt Tuning, April 18. arXiv.org. https: //arxiv.org/abs/2104.08691

[18] Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). Segment anything in medical images. Nature Communications, 15(1). https: //doi.org/10.1038/s41467-024-44824-z

[19] Ghiasi, G., Gu, X., Cui, Y., & Lin, T. (2021, December 22). Scaling Open-Vocabulary Image Segmentation with Image-Level Labels. arXiv.org. https: //arxiv.org/abs/2112.12143

[20] Chen, G., Zhou, Y., et al. (2024). Remote sensing of diverse urban environments: From the single city to multiple cities. Remote Sensing of Environment, 285, 113396.

[21] Zhang, C., Han, D., Qiao, Y., Kim, J. U., Bae, S., Lee, S., & Hong, C. S. (2023). Faster segment anything: towards lightweight SAM for mobile applications, June 25. arXiv.org.

[22] Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., & Wang, J. (2023). Fast segment anything, June 21. arXiv.org. https: //arxiv.org/abs/2306.12156

[23] Wang, D., Shelhamer, E., Liu, S., Olshausen, B., & Darrell, T. (2020). TENT: Fully test-time adaptation by entropy minimization, June 18. arXiv.org. https: //arxiv.org/abs/2006.10726

[24] Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? March 15. arXiv.org. https: //arxiv.org/abs/1703.04977

[25] Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2016). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, June 2. arXiv.org. https: //arxiv.org/abs/1606.00915

[26] Krähenbühl, P., & Koltun, V. (2012). Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials, October 20. arXiv.org. https: //arxiv.org/abs/1210.5644