Dean&Francis

ISSN 2959-6157

Image Segmentation Based on Transformer-Class Methods

Yiyang Liu

College of software, Software Engineering, Taiyuan University of Technology, Jinzhong, Shanxi, China
*Corresponding author: 2023006089@link.tyut.edu.cn

Abstract:

Convolutional neural networks (CNNs) have many problems, including the inability to encode long-range details of images, failure to capture the global information, and other problems, like vanishing or exploding gradients during image segmentation tasks. Since its introduction in 1985, the Transformer model has proven to be very beneficial in natural language processing and computer vision. This paper reviews the application and development of Transformer models in image segmentation tasks within the biomedical, industrial manufacturing and agricultural environments in order to promote joint development of Transformer architecture and image segmentation research. The article explains the base and relevance of this study and dwells on the discussion of the three key challenges as follows: increasing contour accuracy, generalization ability and strength, and lowering the computation cost. It is also important to mention that in the future, the Transformer model will be more versatile by optimizing its architecture and algorithms, and decreasing the rate at which the parameters are updated during the process of adapting to new tasks. Moreover, combining the Transformer with additional network architectures is likely to create a dynamic tradeoff between local and long-range dependencies.

Keywords: Transformer, U-Net, Image Segmentation.

1. Introduction

The Transformer architecture was introduced in 2017 by Vaswani et al. and has fundamentally transformed computer vision since its adaptation to image processing tasks. Before Transformers entered the field, image segmentation was dominated by Convolutional Neural Networks (CNNs). Key milestones include the Fully Convolutional Network (FCN) [1], which replaced fully connected layers with convolutional layers to enable dense pixel-wise prediction; the DeepLab family [2–4], which employed atrous convolutions and atrous spatial pyramid pooling to expand receptive fields while preserving spatial resolution; and U-Net [5], proposed in 2015, which introduced symmetric skip connections to effectively combine high-resolution local features with deep semantic context. These architectures dramatically outperformed earlier hand-crafted techniques such as

thresholding, region growing, and edge detection, which lacked generalisation ability and required extensive manual tuning.

Despite their achievements, CNN-based methods suffer from inherent limitations. The local nature of convolutional operations restricts the modelling of long-range dependencies and global contextual information. As network depth increases, vanishing or exploding gradients become more common. Moreover, the strong locality bias of convolutions often leads to incomplete object representations, over-smoothed boundaries, or confusion between visually similar classes—issues that are particularly severe in biomedical, industrial, and agricultural segmentation tasks.

The self-attention mechanism of Transformers directly addresses these shortcomings by allowing every image patch to attend to all others regardless of spatial distance, thereby capturing global context from the earliest layers. Starting with TransUNet in 2021 [6], which integrated a Transformer encoder into a U-Net-like framework, the field has witnessed a rapid proliferation of Transformer-based segmentation models. Prominent examples include pure Transformer architectures such as Swin-UNet [7], UNETR [8], and Swin UNETR [9]; efficient hierarchical designs paired with lightweight decoders, represented by SegFormer [10]; and diverse hybrid strategies that combine CNNs and Transformers in complementary ways.

These advances have driven state-of-the-art performance across biomedical imaging, industrial defect detection, and precision agriculture, where demands for boundary precision, generalisation across varying acquisition conditions, and robustness to limited annotated data are especially stringent. Nevertheless, challenges persist in achieving sharp contour delineation, maintaining high accuracy with small datasets, and reducing computational cost for real-world deployment.

This review consolidates the historical development of CNN-based segmentation in this introductory section and focuses the subsequent sections on the architectural evolution, critical comparative analysis, and domain-specific applications of Transformer-based methods in biomedical, industrial, and agricultural contexts. By doing so, it aims to clarify current best practices, highlight performance trade-offs among competing paradigms, and identify the most promising directions for future research.

2. Hybridisation and Refinement: Architectural Evolution of Transformers in Medical Image Segmentation

Medical image segmentation demands exceptional boundary precision, robust generalisation across scanners, modalities, and pathologies, and reliable performance with limited annotated data. Since 2021, Transformer-based architectures have consistently outperformed purely convolutional models in this domain by directly capturing long-range anatomical dependencies that convolutional networks can only approximate through deep hierarchical stacking. This section provides a systematic critical analysis of the major architectural paradigms, emphasising their fundamental design differences, performance trade-offs, and the specific clinical challenges each paradigm is best suited to address.

The earliest and still most influential paradigm uses the Transformer as a powerful global encoder while retaining a convolutional decoder for accurate spatial localisation. TransUNet, proposed in 2021 [6], pioneered this hybrid strategy by processing image patches through a Transformer encoder and reshaping the resulting contextualised tokens into feature maps that are progressively upsampled via cascaded deconvolutional modules with skip connections from shallow convolutional layers. This design achieved Dice score improvements of 3–5% over U-Net on multi-organ benchmarks such as Synapse and ACDC, particularly for structures exhibiting large shape variation or distant spatial relationships. Its main limitations are high memory consumption and inference latency caused by quadratic-complexity global attention, with boundary quality remaining partly dependent on the convolutional recovery path.

Researchers subsequently explored whether convolutions could be eliminated entirely. Swin-UNet [7] and its three-dimensional extension Swin UNETR [8] demonstrated that fully Transformer-based encoder-decoder architectures are not only viable but often superior. Built exclusively on hierarchical shifted-window self-attention blocks and Patch Expanding upsampling layers, these models deliver 1–2% higher Dice scores than TransUNet on identical datasets while producing markedly sharper boundaries, especially for thin or highly tortuous structures. By contrast, UNETR [9], which connects a standard Vision Transformer encoder to a convolutional decoder at multiple intermediate layers, suffers significant positional detail loss due to unrestricted global attention and consistently underperforms hierarchical alternatives on boundary-critical tasks.

A parallel efficiency-driven lineage is represented by SegFormer [10]. Its hierarchical Transformer encoder naturally generates multi-scale feature maps that are fused by an extremely lightweight all-MLP decoder. Numerous medical adaptations have since confirmed that SegFormer variants attain accuracy comparable to or exceeding TransUNet while requiring 3–5× fewer FLOPs and 50–70% less memory, thereby enabling practical volumetric

ISSN 2959-6157

segmentation on standard clinical hardware.

The latest and most promising direction addresses the chronic scarcity of labelled medical data through parameter-efficient adaptation of large pretrained Transformers. Swin-TUNA [11] exemplifies this approach by inserting lightweight tunable adapters into a frozen Swin Transformer backbone, updating only approximately 4% of parameters during fine-tuning. This strategy dramatically enhances few-shot learning, cross-modality transfer, and cross-scanner generalisation without sacrificing inference efficiency.

In conclusion, although pure Transformer designs initially appeared poised to fully replace convolutions, the medical imaging community has largely converged on two highly effective paradigms: Transformer-as-encoder hybrids that maximise raw accuracy on large datasets, and efficient hierarchical Transformers with lightweight decoding that offer the best accuracy-efficiency balance for real-world clinical deployment. Pure convolution-free models remain competitive only when extensive pretraining or substantial computational resources are available. The most promising future path lies in combining hierarchical Transformers with parameter-efficient adaptation techniques to achieve both state-of-the-art performance and robust generalisation on the small, diverse datasets typical of clinical practice.

3. Small Samples and High Efficiency: Challenges and Evolution in Industrial Image Segmentation Models

Industrial defect segmentation faces extreme challenges: defective samples are rare and heavily imbalanced, high-quality annotations are costly, and inference must be both highly sensitive and fast enough for real-time deployment on production lines. These constraints make most medical-grade Transformer models impractical due to their large parameter counts and dependence on abundant labelled data.

SegFormer marked a major advance for industrial applications. Its hierarchical Transformer encoder and lightweight all-MLP decoder achieve excellent performance on standard surface defect benchmarks such as MVTec AD and various rail/crack datasets while requiring only a fraction of the computation and memory of earlier hybrid models. The naturally multi-scale feature maps from its MiT backbone prove especially valuable for detecting both large anomalies and sub-millimetre cracks in a single forward pass.

In 2025, Gong Yulei et al. further improved SegFormer for concrete and metal surface inspection by integrating coordinate attention and diagonal clustering mechanisms into the decoder. These targeted enhancements significantly strengthen the modelling of narrow elongated defects and long-range spatial relationships, delivering intersection-over-union gains of 4–7 % on thin-crack tasks with almost no additional computational cost.

A separate line of work builds on foundation models. FoodSAM adapts the Segment Anything Model for industrial instances and panoramic segmentation by combining prompt-driven mask generation with learned category embeddings. It demonstrates impressive zero-shot and few-shot generalisation across defect categories, yet its iterative refinement and large frozen backbone result in inference latency and memory demands that remain prohibitive for most production environments.

The most practical and effective solution to date is Swin-TUNA, proposed by Chen Haotian and Xiao Zhiyong in 2025. This method freezes a pretrained Swin Transformer backbone and inserts lightweight TUNA adapter modules into each layer, updating only about 4 % of total parameters when adapting to downstream defect datasets. The approach simultaneously solves three critical industrial problems: severe overfitting on tiny labelled sets is largely eliminated, training time drops from days to hours on a single GPU, and inference speed remains virtually identical to the original backbone. On multiple surface defect benchmarks, Swin-TUNA matches or surpasses fully fine-tuned SegFormer and TransUNet variants while easily satisfying real-time requirements of over 100 frames per second on edge devices.

In summary, industrial defect segmentation has rapidly moved away from compute-heavy hybrids toward lightweight hierarchical Transformers enhanced by targeted attention mechanisms and, most decisively, parameter-efficient adapter techniques. The Swin-TUNA paradigm now offers the best balance of sensitivity to rare defects, strong generalisation from extremely limited annotations, and the low latency essential for deployment on production lines. Future industrial systems are expected to rely increasingly on frozen foundation models combined with lightweight adapters, enabling continuous online learning from streaming defect data without expensive retraining.

4. From U-Net to CNN-Transformer Hybrid Architectures: Precision in Agricultural Image Segmentation

Agricultural image segmentation must contend with extreme variability caused by illumination changes, occlusion by leaves, complex backgrounds, and near-identical visual appearance between crops and weeds. Datasets are

often large in raw image volume yet poor in per-class annotation density, while inference speed remains critical for real-time robotic weeding, harvesting, and drone-based monitoring. These conditions have driven a distinct evolutionary path that favours balanced accuracy, multi-scale reasoning, and efficient generalisation over raw parameter scale.

Early agricultural work relied heavily on U-Net and its lightweight variants because of their proven ability to preserve fine boundaries with limited training data. As higher-resolution satellite and UAV imagery became common, the global context limitation of pure CNNs grew increasingly apparent, particularly for distinguishing interwoven crop rows and detecting subtle phenotyping traits across vast fields.

The introduction of Transformer-based models brought significant advances. SegFormer rapidly gained traction in remote sensing because its hierarchical encoder directly outputs multi-scale features that align naturally with the nested spatial structure of agricultural scenes – individual plants at fine scale, rows at medium scale, and field blocks at coarse scale. Numerous studies have since confirmed that SegFormer consistently outperforms DeepLabv3+ and U-Net on public benchmarks such as DeepGlobe, LoveDA, and rice/weed datasets while requiring far fewer FLOPs.

Pure hierarchical Transformers such as Swin Transformer, were also evaluated extensively. A 2022 comparative study by Xu Huiyao et al. [12] on Sentinel-2 rice mapping showed that Swin Transformer achieved the highest mean intersection-over-union and sharpest field boundaries among U-Net, DeepLabv3+, and itself, but at the cost of dramatically higher parameter count and inference time, rendering it impractical for edge deployment on drones or tractors.

The most successful agricultural solutions have therefore converged on sophisticated CNN-Transformer hybrids that combine local inductive biases with global reasoning. CCTNet, proposed by Wang Hong et al. [13], adopts a two-stream parallel architecture in which a lightweight CNN branch extracts high-resolution local details while a Transformer branch models long-range row alignment and field-level context; the streams are fused progressively through cross-attention. CCTNet significantly reduced confusion between rice and surrounding vegetation compared with single-modality baselines.

An even more refined design, CTFuseNet from Xiang Jianjian et al. in 2023 [14], employs dual parallel CNN and Transformer encoders followed by a dedicated CTFuse module and Feature Pyramid Network head for multi-scale fusion. On large-scale UAV crop-type mapping datasets, CTFuseNet reached a mean intersection-over-union of 85.33 %, outperforming both CCTNet and pure Transformer alternatives while maintaining inference speeds suitable for real-time agricultural robotics.

Recent efforts have begun exploring parameter-efficient adaptation for agricultural scenarios where labelled data remain regionally specific. Lightweight adapters inserted into frozen hierarchical backbones now enable rapid domain transfer – for example, from European wheat fields to Asian rice paddies – using only hundreds of annotated images, a capability that will become increasingly valuable as precision agriculture expands globally.

In summary, agricultural image segmentation has largely abandoned both pure CNNs and pure Transformers in favour of carefully engineered parallel or fused CNN-Transformer hybrids. Designs such as CCTNet and CTFuseNet currently represent the state-of-the-art, delivering superior boundary precision and class discrimination under varying field conditions while satisfying the computational constraints of UAV and tractor-mounted systems. Future progress is expected to centre on parameter-efficient adaptation of large pretrained vision foundation models, enabling continual learning across seasons, crops, and geographies without prohibitive annotation or retraining costs.

5. Current Limitations and Future Prospects

Transformer-based segmentation models still face three core limitations: (1) high data hunger and weak performance on small or imbalanced datasets, (2) significantly higher computational and memory costs than optimised CNNs, hindering edge deployment, and (3) inferior boundary precision without strong convolutional biases or dedicated refinement.

Future directions will centre on parameter-efficient adaptation of large pretrained foundation models, lightweight sparse attention, deeper CNN-Transformer fusion with boundary-aware constraints, and continual/test-time adaptation mechanisms. These advances will deliver accurate, efficient, and deployable systems across biomedical, industrial, and agricultural domains.

6. Conclusions

Transformer models have proven to be a formidable substitute for convolutional neural networks (CNNs) in the area of image segmentation. Their use in biomedicine, industry and agriculture has been the reason behind the successive domain-driven architectural evolutions. The most efficient models now adopt a hybrid architecture with Transformers and CNNs with their specific configu-

ISSN 2959-6157

rations depending on domain needs: in biomedicine, the current paradigm is a Transformers-based global encoder paired with CNN decoders; the industrial world prefers hierarchical Transformers paired with lightweight adapters (e.g., Swin-TUNA) to handle the limitations of small size samples and real-time applications; in agriculture, parallel or fused CNN-Transformer-based hybrid models (e.g., CCTNet) are a good trade- However, Transformer models face three fundamental issues, including a high reliance on labeled data, high costs in computational memory, and the lack of accuracy in segmenting boundaries in challenging situations. These limitations restrict their implementation in limited resources like the environment. In the future, there will be a lot of research on how to apply parameter efficient adaptation methods to big foundational models so that they can learn to provide robust generalisation-on-small samples in the smallest possible training cost; how to create lighter sparse attention mechanisms that can trim down computational costs; and how to further optimise model performance and applicability with stronger CNN-Transformer integration.

References

[1] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 3431–3440.

[2] Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs. 2014.

[3] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 834–848.

[4] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation. 2017.

[5] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015: 234–241. Cham: Springer International Publishing.

[6] Chen J, Lu Y, Yu Q, et al. TransUNet: Transformers make strong encoders for medical image segmentation. 2021.

[7] Cao H, Wang Y, Chen J, et al. Swin-UNet: UNet-like pure transformer for medical image segmentation. European Conference on Computer Vision, 2022: 205–218. Cham: Springer Nature Switzerland.

[8] Hatamizadeh A, Tang Y, Nath V, et al. UNETR: Transformers for 3D medical image segmentation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022: 574–584.

[9] Hatamizadeh A, Nath V, Tang Y, et al. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. International MICCAI Brainlesion Workshop, 2021: 272–284. Cham: Springer International Publishing.

[10] 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: 12077–12090.

[11] Chen H, Xiao Z. Swin-TUNA: A novel PEFT approach for accurate food image segmentation. 2025.

[12] Xu H, Song J, Zhu Y. Evaluation and comparison of semantic segmentation networks for rice identification based on Sentinel-2 imagery. Remote Sensing, 2023, 15(6): 1499.

[13] Wang H, Chen X, Zhang T, et al. CCTNet: Coupled CNN and transformer network for crop segmentation of remote sensing images. Remote Sensing, 2022, 14(9): 1956.

[14] Xiang J, Liu J, Chen D, et al. CTFuseNet: A multi-scale CNN–transformer feature fused network for crop type segmentation on UAV remote sensing imagery. Remote Sensing, 2023, 15(4): 1151.