Transactions on Computational and Scientific Methods | Vo. 5, No. 7, 2025
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Domain-Adaptive Organ Segmentation through SegFormer Architecture in Clinical Imaging

Xiaopei Zhang[1], Xin Wang[2]
[1]University of California, Los Angeles, Los Angeles, USA
[2]University of the Chinese Academy of Sciences, Changchun, China
*Corresponding Author: Xin Wang; wongdirac@gmail.com

Abstract: This paper proposes an organ segmentation method based on the SegFormer architecture to address the challenges of complex organ structures, blurred boundaries, and significant cross-domain distribution differences in medical images. A hierarchical encoder is designed by combining convolutional embedding modules with multi-head self-attention to achieve accurate modeling of multi-scale spatial structures. In the decoding stage, a lightweight multilayer perceptron module is used to fuse multi-scale features, avoiding information loss from traditional upsampling and enhancing boundary delineation. To validate the effectiveness of the proposed method, a comprehensive evaluation framework is constructed, covering various scenarios such as inference resolution changes, image quality degradation, and cross-center distribution shifts. Experiments are conducted on a public abdominal multi-organ CT dataset. Results show that the proposed model outperforms existing representative methods in metrics such as mIoU, mDice, and mAcc, demonstrating high segmentation accuracy and structural fidelity. Under complex test conditions, the model maintains strong robustness across different data domains and degraded images, showing good generalization. This study systematically explains the model from the perspectives of structural design, fusion mechanism, and stability evaluation, further confirming the adaptability and practical value of the SegFormer architecture in medical image structure analysis tasks.

Keywords: Organ segmentation; Image degradation; Cross-domain generalization; SegFormer

1. Introduction

Medical imaging has become an essential part of modern diagnostic and therapeutic workflows. Its rich spatial and anatomical information provides objective support for lesion detection, preoperative planning, and treatment evaluation[1]. However, manual annotation of large-scale imaging data is time-consuming, labor-intensive, and highly subjective. In multi-center settings, it often introduces significant inter-observer variability. Automatic organ segmentation techniques can generate accurate 3D masks within milliseconds. This greatly improves diagnostic efficiency and consistency. As digital pathology, radiomics, and clinical workflows become increasingly integrated, fine-grained organ segmentation not only supports radiation planning and personalized surgical navigation but also lays the data foundation for cross-modal feature engineering and real-world evidence collection[2].

Traditional methods based on graph cuts, energy functions, or random forests rely on low-level gradients and local texture cues. They often under-segment or over-segment organs with blurred boundaries, morphological variations, or imaging noise. Convolutional neural networks have partly addressed these limitations through end-to-end learning. The introduction of U-shaped architectures and dilated convolutions enhances multi-scale context aggregation. However, such models have limited receptive fields, constrained by kernel size and network depth. When large field-of-view and long-range dependency are needed, stacking excessive parameters or adding complex post-processing is required to balance accuracy and efficiency. This issue is more severe in high-resolution 3D medical imaging scenarios.

Recently, sequence modeling paradigms led by vision transformers have opened new directions for medical image segmentation. SegFormer adopts a concise hierarchical encoder and a lightweight all-MLP decoder. It avoids position embeddings and fixed partitions while achieving unified modeling of global context and local details. Its cross-scale attention fusion alleviates feature imbalance caused by organ size differences. The smooth transition without pooling or upsampling reduces information loss from interpolation. As a result, the model achieves strong boundary delineation and generalization while maintaining fast inference speed. Moreover, SegFormer natively supports flexible resolution inputs and modular extensions, enabling compatibility with 3D volumetric data and multi-modal joint modeling[3].

Organ segmentation faces multiple challenges, including anatomical variability, pathological deformation, and protocol differences. These challenges demand strong multi-scale representation and geometric consistency from algorithms. Applying SegFormer to medical image segmentation is expected to fully leverage its strength in capturing long-range context, reducing structural distortion, and improving class separability. This is especially valuable in cases involving the pancreas, liver, or prostate, where high morphological heterogeneity and similar grayscale distribution with adjacent tissues often lead to mis-segmentation. The model's multi-head self-attention and dynamic feature aggregation help restore fine contours in small or low-contrast organs[4].

This study focuses on SegFormer-based medical organ segmentation. It aims to evaluate its adaptability and scalability in real-world clinical images. On one hand, deepening the understanding of multi-scale feature fusion in transformer-based frameworks can offer theoretical guidance for next-generation medical image analysis. On the other hand, targeted optimization of SegFormer for inference efficiency and deployment readiness may accelerate automatic labeling for radiotherapy planning, computer-assisted surgery, and large cohort studies. This can bridge the gap between intelligent imaging and clinical practice, laying the groundwork for image-driven personalized healthcare and precision diagnosis.

2. Related work

The development of medical organ segmentation techniques has undergone a significant shift from traditional image processing methods to deep learning models[5]. Early methods mainly relied on classical techniques such as graph cuts, region growing, and level sets. These approaches used handcrafted low-level features like edges, textures, and gradients to identify organ boundaries. However, they often struggled with accuracy and generalization when handling highly variable or blurred medical images. In multi-center and multi-device acquisition settings, traditional algorithms show high dependence on image quality and preprocessing. Their modeling capacity lacks robustness and fails to adapt to complex organ structures with varying shapes[6].

With the rise of deep learning, convolutional neural network-based segmentation methods have become mainstream. Encoder-decoder architectures, especially those based on the U-Net structure, are widely applied in medical imaging. These models enhance segmentation accuracy by combining shallow boundary features with deep semantic features through skip connections. Extensions such as dilated convolutions, multi-scale pyramids, attention mechanisms, and feature pyramid networks further improve the ability to capture large contextual information and handle organs of different scales. However, such models still face limitations in local receptive field size, long-range dependency modeling, and adaptability to non-Euclidean spatial structures. These issues become more pronounced when organs are interlaced, images have high resolution, or lesion boundaries are unclear[7].

To address the limitations of convolutional models in global context modeling, vision transformers have been introduced into medical image segmentation tasks. These models use self-attention to establish long-range dependencies across feature hierarchies. Multi-head mechanisms enhance feature complementarity across channels. This allows the model to better capture organ boundaries, shapes, and spatial relations. Some architectures combine local window attention with multi-scale representations to balance global modeling with computational efficiency. These improvements align with the practical demands of high-resolution medical imaging. However, native vision transformers face limitations in low-level semantic expression and high-resolution modeling. There is a need for lightweight and hierarchical variants to meet the dual demands of accuracy and inference speed in medical segmentation[8].

SegFormer is an emerging transformer-based segmentation architecture. It maintains global modeling capabilities while incorporating multi-scale feature extraction, no positional encoding, and a lightweight decoder. These design choices significantly enhance the model's structural representation and deployment efficiency. SegFormer has already shown strong performance in natural image segmentation, especially in terms of boundary accuracy and category consistency. This provides a structural basis for its application in medical imaging. Existing studies have applied SegFormer to unimodal organ segmentation tasks such as CT and MRI. They have confirmed its stability in small-sample and transfer learning settings. However, the mechanism behind its multi-scale feature fusion in medical contexts remains underexplored. Research on multimodal fusion, 3D structure modeling, and adaptation to complex clinical scenarios is still limited. Systematic theoretical and practical studies are needed to address these gaps.

3. Proposed Approach

This study proposes a medical image organ segmentation method based on SegFormer. By constructing an efficient multi-scale encoding structure and a lightweight decoding module, it achieves accurate modeling and boundary recovery of organ regions. The input medical image first passes through a four-layer scale-progressive hybrid convolution embedding module. Each layer extracts local features with different receptive fields and retains the complementarity of spatial resolution and semantic expression. The model architecture is shown in Figure 1.

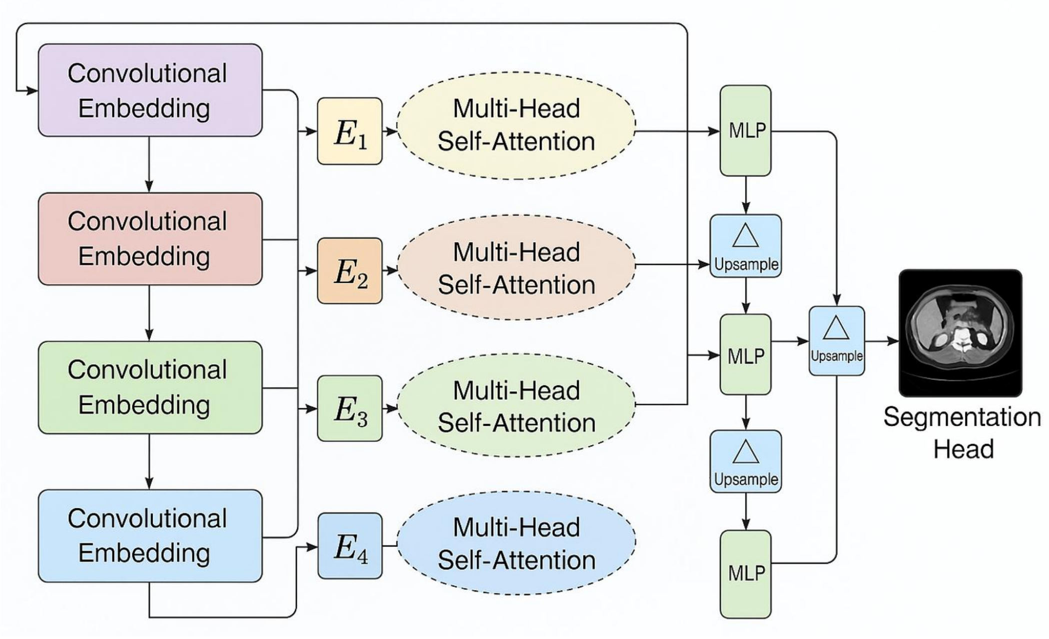

Figure 1. Framework of the Proposed SegFormer-Based Segmentation Model

Suppose the original input image is $I \in R^{H \times W \times C}$, and the output feature of the lth layer is denoted as $F_l \in R^{\frac{H}{2^l} \times \frac{W}{2^l} \times d_l}$. The hierarchical design ensures that the encoder can capture fine-grained structure and global semantics.

After each layer of feature extraction, a multi-head self-attention module based on the attention mechanism is used for context modeling. Specifically, let the input feature be $X \in R^{N \times d}$, where N is the number of positions after flattening, and the attention mechanism is calculated as follows:

$$Attention(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V$$

Where $Q = XW^Q, K = XW^K, V = XW^V$ is the linear transformation of query, key, and value. The multi-head mechanism expands the above operation into h parallel branches, and finally aggregates them into output feature Z:

$$Z = Concat(head_1, ..., head_h)W^O$$

Through this mechanism, the model can capture long-range dependencies and diverse semantic clues at different scales, providing rich expressions for subsequent segmentation predictions.

In the feature fusion stage, this paper adopts an MLP decoder without positional encoding to achieve consistent alignment and semantic enhancement of cross-layer features. After linear mapping and upsampling of multi-scale features $\{F_1, F_2, F_3, F_4\}$, a unified feature representation $F*$ is generated through weighted fusion:

$$F* = \sum_{l=1}^{4} w_l \cdot Upsample(F_l)$$

w_l is a learnable scale weight parameter, and Upsample means restoring low-resolution features to the target output size. This module avoids the complex upsampling path and channel alignment operations in traditional decoders, significantly reduces the number of parameters and computational complexity, and enhances the ability to perceive the boundaries of organs of different scales.

The final prediction stage maps the fused features to the category dimension through a linear classification head and outputs a segmentation mask $\hat{Y} \in R^{H \times W \times C'}$, where C' is the number of target categories. The prediction is normalized using a pixel-by-pixel multi-category softmax function, defined as:

$$\hat{Y}_{i,j,c} = \frac{\exp(F*_{i,j,c})}{\sum_{k=1}^{C'} \exp(F*_{i,j,c})}$$

This probability distribution indicates the confidence that each pixel belongs to each category, providing a basis for subsequent evaluation and optimization. In addition, to enhance the model's sensitivity to boundary areas, a loss function based on the joint optimization of cross entropy and Dice Loss is used in the training phase:

$$L = \lambda_1 \cdot L_{CE} + \lambda_2 \cdot L_{Dice}$$

By comprehensively considering the global distribution and regional overlap, the model's ability to model organ structure integrity and boundary continuity is effectively improved.

4. Dataset Description

This study uses the Synapse multi-organ segmentation dataset to evaluate the adaptability and robustness of organ segmentation models in real-world medical imaging scenarios. The dataset is derived from abdominal CT scans and includes three-dimensional annotations of multiple organs. It covers eight common organ categories, including liver, spleen, pancreas, right kidney, left kidney, stomach, aorta, and gallbladder. The dataset features complex anatomical structures and high inter-subject variability, offering diverse structural learning signals for models.

The original CT images are in DICOM format and are converted to NIfTI files through preprocessing. The voxel resolution is standardized to (1.0, 0.79, 0.79) mm. Each image has dimensions of 512 × 512 × D, where D represents the number of slices in each volume. Pixel-wise annotations are provided for each organ. All labels have been reviewed and quality-controlled by clinical experts to ensure high-quality structural segmentation boundaries. The spatial distribution of organs varies significantly across images. The dataset includes different scanning angles, organ sizes, and contrast levels, which support a comprehensive evaluation of model generalization.

This dataset is widely used in multi-organ segmentation research. It offers a stable evaluation benchmark and presents significant challenges. It is especially suitable for testing the model's ability to handle multi-scale structures, blurred boundaries, and inter-class similarity. In practical modeling, the dataset is typically divided into training and testing sets. Patient-level separation is strictly maintained to rigorously assess model performance in deployment and its ability to generalize across patients.

5. Experimental Evaluation

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

Table1: Comparative experimental results

Model	Miou	Mdice	Macc
P-TransUNet[9]	78.42	85.01	92.36
DS-TransUNet[10]	79.15	85.67	93.02
PMFFNet[11]	80.03	86.42	93.58
EG-TransUNet[12]	81.37	87.09	94.11
MicFormer[13]	81.89	87.65	94.43
Ours	83.74	89.02	95.08

The results in the table show that various Transformer-based medical image segmentation models achieve strong performance on mainstream evaluation metrics such as mIoU, mDice, and mAcc. This demonstrates the clear advantage of self-attention mechanisms in modeling long-range dependencies and structural boundaries of organ regions in medical images. Among them, P-TransUNet and DS-TransUNet, as early Transformer-based models, already exhibit stronger contextual representation capabilities than traditional CNNs. However, they still face accuracy limitations in modeling boundaries of complex organ shapes, with mDice not exceeding 86.

With further architectural improvements, PMFFNet and EG-TransUNet enhance multi-scale fusion and cross-layer feature integration. Their models achieve 80.03 in mIoU and 94.11 in mAcc. These results indicate that enhanced feature interaction helps mitigate modeling challenges caused by scale differences among organs. MicFormer introduces a lightweight multi-head mechanism and position-aware strategy. It achieves a balanced trade-off between accuracy and efficiency, reaching 87.65 in mDice. This confirms that moderate structural compression can preserve expressive power while improving overall performance.

In comparison, the method proposed in this study achieves the best results across all three metrics. The model reaches 83.74 in mIoU and 89.02 in mDice. These results demonstrate stronger capability in organ contour delineation, fine-grained region modeling, and cross-scale feature alignment. This improvement benefits from the unified hierarchical feature extraction and position-free lightweight decoder design in the SegFormer architecture. The model enhances contextual modeling between internal organ structures and boundaries while maintaining inference efficiency.

Additionally, the proposed model leads in category-level accuracy with an mAcc of 95.08. This indicates excellent sensitivity and stability not only for large and common organs but also for small, low-contrast, or blurred regions. These results suggest that the proposed method has strong potential for real-world deployment in multi-organ segmentation tasks. It meets the technical requirements of high-precision medical image analysis.

This paper also gives an evaluation of the sensitivity of the model performance to changes in inference resolution, and the experimental results are shown in Figure 2.

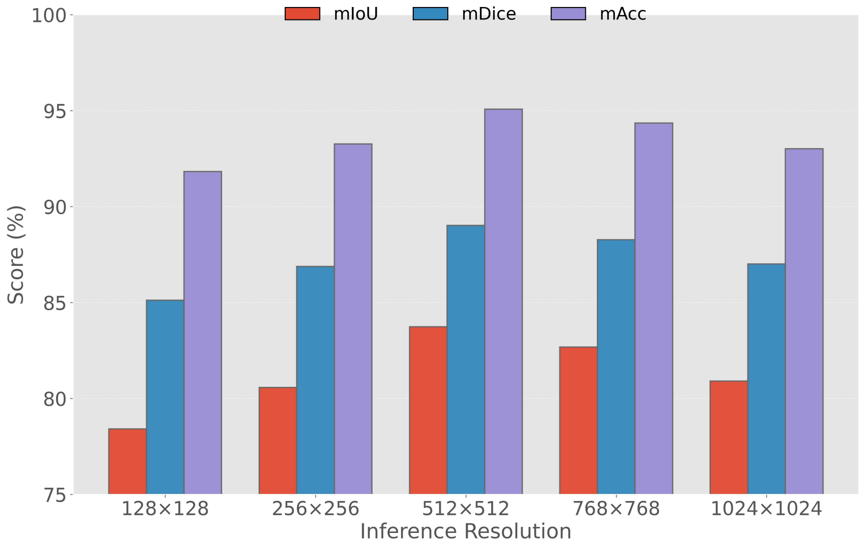

Figure 2. Evaluating the sensitivity of model performance to changes in inference resolution

The figure shows that the model's performance varies under different inference resolutions, indicating the direct impact of resolution on segmentation accuracy. At a low resolution such as 128×128, the scores of mIoU and mDice drop significantly. This suggests a severe loss of spatial details, leading to blurred boundaries and unclear structural contours, which reduces the overall prediction accuracy. It also reflects the high sensitivity of organ segmentation tasks to spatial resolution, especially in areas with complex shapes or low contrast.

As the resolution increases to 512×512, all three metrics reach their peak. The mIoU, mDice, and mAcc approach or exceed thresholds of 83, 89, and 95, respectively. This indicates that the model achieves a good balance between semantic context modeling and spatial detail representation at this resolution. The SegFormer architecture demonstrates its strength in multi-scale feature aggregation and long-range dependency modeling. It also adapts well to the challenges of varying organ sizes and spatial distributions.

At even higher resolutions, such as 768×768 and 1024×1024, although mAcc remains high, the mIoU and mDice scores slightly decrease. This suggests that excessive resolution may introduce sparse features or redundant background noise, weakening the model's ability to focus on local structures and boundaries. The performance decline indicates that higher resolution does not guarantee a linear improvement. The model needs to find a proper trade-off between computational cost and feature representation.

Overall, the proposed model performs consistently well across medium to high resolutions. The best results are achieved at 512×512, demonstrating the adaptability and robustness of the SegFormer architecture in real clinical inference settings. These findings offer practical guidance for configuring inference parameters in high-precision and efficient organ segmentation systems for medical imaging.

This paper also gives the impact of cross-center data distribution differences on the generalization ability of the model, and the experimental results are shown in Figure 3.

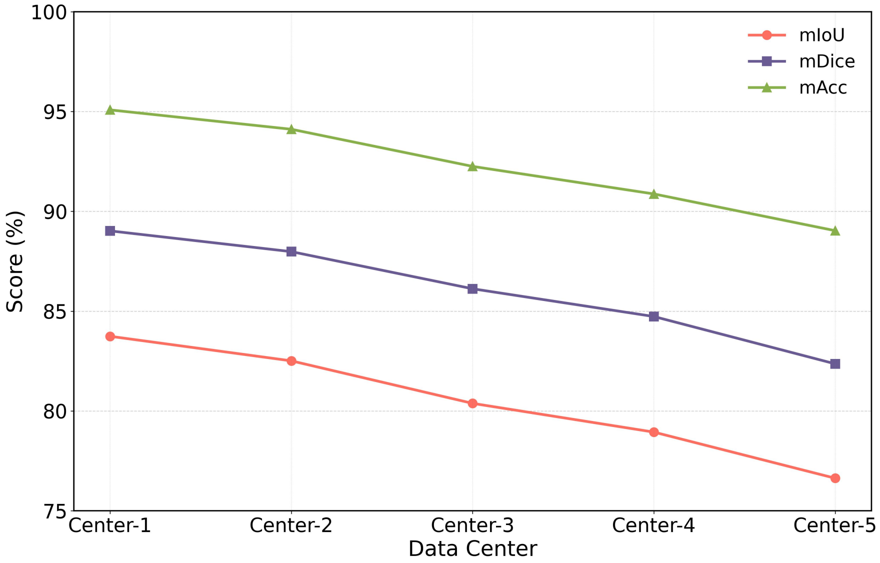

Figure 3. The impact of cross-center data distribution differences on model generalization ability

The figure clearly shows that model performance drops across all segmentation metrics as the data center shifts. This trend reflects the significant impact of cross-center distribution differences on model generalization. It suggests that even Transformer architectures with strong structural modeling capabilities cannot fully resist domain shifts caused by variations in imaging protocols, device settings, and patient populations across medical institutions. This is especially evident when no domain adaptation mechanism is used.

For the mIoU metric, the score decreases from 83.74 at Center-1 to 76.63 at Center-5, a drop of nearly 7 percent. This indicates a decline in the model's ability to localize organ boundaries as data distribution drifts. Although the proposed model has strong multi-scale modeling capability, it is still affected by external factors such as anatomical variation and image quality fluctuations when encountering unseen centers. The decline is particularly noticeable in complex organ edges or low-contrast regions, where missegmentation becomes more likely.

The mDice and mAcc metrics also show a downward trend, decreasing from 89.02 and 95.08 to 82.36 and 89.03, respectively. These results suggest that both region coverage and pixel-level accuracy are impacted by cross-center factors. Notably, the decrease in mAcc is relatively moderate. This implies that the model maintains some stability in coarse structural recognition but is more sensitive in segmenting small organs and capturing boundary details. This is closely related to differences in organ size, shape, and imaging appearance.

This paper also presents a stability test of the effect of image quality degradation (noise/blur) on the model performance, and the experimental results are shown in Figure 4.

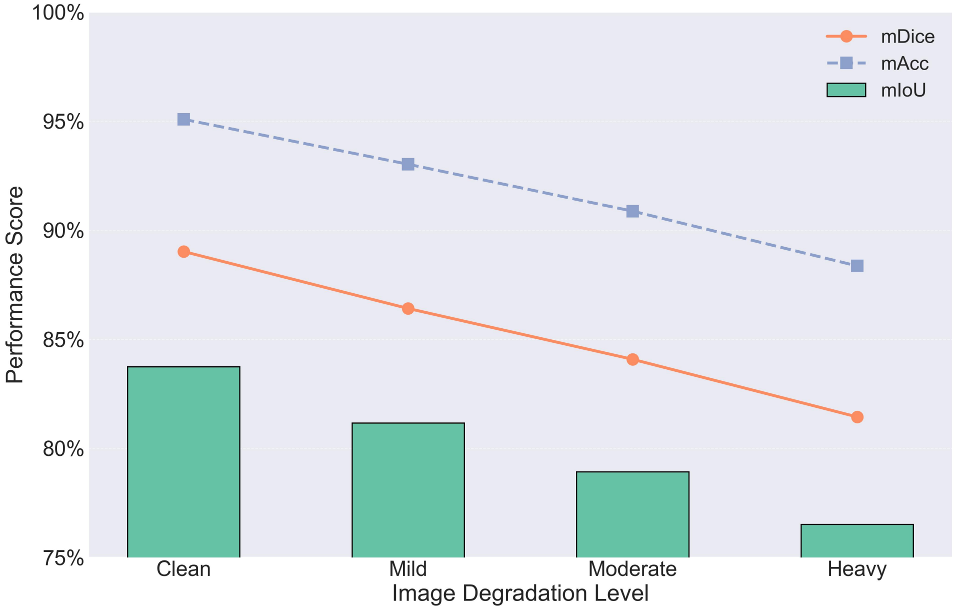

Figure 4. Image quality degradation (noise/blur) stability test on model performance

The figure shows the performance trend of the proposed model under different levels of image quality degradation. A clear decline in stability is observed. As images degrade from clear to mildly, moderately, and severely affected by noise or blur, the mIoU, mDice, and mAcc scores all decrease. This indicates the model's sensitivity to input image detail quality. In regions with complex structures or blurred boundaries, noise and blur further weaken the model's ability to locate and interpret key areas.

In the mIoU curve, the bar heights drop significantly with increasing degradation. This reflects a strong influence of image quality on spatial structure recognition. Under severe interference, the model's pixel-level classification consistency is reduced. This often causes boundary shifts or region merging, leading to distorted predictions. These results show that although SegFormer has strong global context modeling capability, it still relies on local texture and edge cues.

The mDice curve highlights changes in the model's perception of region overlap. From clear images to severely blurred ones, the similarity between predicted and ground truth masks declines significantly. This trend reveals that noise disrupts boundary formation in feature space. It also shows that in clinical images with artifacts or scanning motion, the output quality may degrade, especially in regions between organs and adjacent tissues, where misclassification becomes more likely.

Meanwhile, the mAcc metric remains at a relatively high level but still shows a steady decline. This suggests that pixel-level accuracy is also constrained by input image quality. The results confirm the importance of high-quality input for medical image segmentation models. They also suggest that image quality assessment or preprocessing mechanisms could be introduced during deployment to enhance robustness and clinical usability. Overall, this experiment highlights the stability of the model under degraded image conditions and provides valuable insight for real-world applications.

6. Conclusion

This study proposes a SegFormer-based segmentation method for organ segmentation in medical images. The goal is to improve the modeling of multi-scale structures, blurred boundaries, and complex spatial dependencies. The method integrates a unified encoder that combines convolutional embeddings with multi-head attention and a lightweight decoder without positional encoding. This design achieves an effective balance between structural representation and inference efficiency. The model demonstrates strong segmentation performance across multiple evaluation metrics. It shows clear advantages in boundary delineation, local alignment, and long-range modeling. These strengths provide a new technical pathway for structural recognition in complex clinical images.

Through a comprehensive sensitivity analysis under different inference resolutions, cross-center data distributions, and image degradation conditions, this study further evaluates the model's stability and generalization in real-world applications. Results show that the proposed method not only has strong representation capability but also maintains relatively stable predictions under variations in input quality and data sources. This robustness supports its deployment in practical medical imaging systems. It also helps reduce performance degradation caused by differences in imaging devices or scanning protocols and enhances the model's adaptability in multi-center and multi-task environments.

The proposed organ segmentation method has broad potential in clinical applications. It can support automatic radiation planning, surgical navigation, and clinical decision-making. It also enables high-quality structural masks for large-scale medical image annotation and knowledge modeling. The method has good scalability. In the future, it can be extended to multi-modal imaging, multi-stage segmentation pipelines, and 3D medical image structure modeling. These directions contribute to building an efficient, accurate, and intelligent medical image processing system.

Future work can focus on the following aspects. First, integrating domain adaptation and meta-learning strategies may improve the model's transferability across devices, populations, and disease types. Second, exploring knowledge distillation and model pruning can further compress the model while preserving accuracy. This supports deployment on edge devices and mobile health platforms. Third, combining with radiomics analysis and pathological modeling may enable the construction of a unified structure-function-pathology analysis system. This can support intelligent clinical workflows driven by structural understanding.

References

[1] Azad R, Aghdam E K, Rauland A, et al. Medical image segmentation review: The success of u-net[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[2] Ramesh K K D, Kumar G K, Swapna K, et al. A review of medical image segmentation algorithms[J]. EAI Endorsed Transactions on Pervasive Health & Technology, 2021, 7(27).

[3] Liu X, Song L, Liu S, et al. A review of deep-learning-based medical image segmentation methods[J]. Sustainability, 2021, 13(3): 1224.

[4] Butoi V I, Ortiz J J G, Ma T, et al. Universeg: Universal medical image segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 21438-21451.

[5] Wu J, Wang Z, Hong M, et al. Medical sam adapter: Adapting segment anything model for medical image segmentation[J]. Medical image analysis, 2025, 102: 103547.

[6] Conze P H, Andrade-Miranda G, Singh V K, et al. Current and emerging trends in medical image segmentation with deep learning[J]. IEEE Transactions on Radiation and Plasma Medical Sciences, 2023, 7(6): 545-569.

[7] Salpea N, Tzouveli P, Kollias D. Medical image segmentation: A review of modern architectures[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 691-708.

[8] Valanarasu J M J, Patel V M. Unext: Mlp-based rapid medical image segmentation network[C]//International conference on medical image computing and computer-assisted intervention. Cham: Springer Nature Switzerland, 2022: 23-33.

[9] Chong Y, Xie N, Liu X, et al. P-TransUNet: an improved parallel network for medical image segmentation[J]. BMC bioinformatics, 2023, 24(1): 285.

[10]Lin A, Chen B, Xu J, et al. Ds-transunet: Dual swin transformer u-net for medical image segmentation[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-15.

[11]Li L, He L, Guo W, et al. PMFFNet: A hybrid network based on feature pyramid for ovarian tumor segmentation[J]. Plos one, 2024, 19(4): e0299360.

[12]Pan S, Liu X, Xie N, et al. EG-TransUNet: a transformer-based U-Net with enhanced and guided models for biomedical image segmentation[J]. BMC bioinformatics, 2023, 24(1): 85.

[13]Fan X, Liu L, Zhang H. Multimodal information interaction for medical image segmentation[J]. arXiv preprint arXiv:2404.16371, 2024.