2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

TR7Net: A Hybrid Transformer-CNN Framework for Endoscopic Image Segmentation with Validation on Spinal Surgery

Shao-Chi Pao[1], Liuyi Yang[1], Lin Lin[1], Fengcheng Mei[1], Xiaoxing Yang[1*], and Bingding Huang[1*[0000-0002-4748-2882]]

College of Big Data and Internet, Shenzhen Technology University, Shenzhen, 518118, China.
(yangxiaoxing, huangbingding)@sztu.edu.cn

Abstract. In endoscopic surgery, accurate segmentation of medical images is indispensable for surgical planning, navigation, and real-time guidance. However, existing segmentation methods often fall short in capturing complex anatomical details and long-range contextual information, which are critical for precise segmentation. To address these challenges, we introduce TR7Net (TransUNet-RSU7-Network), a novel hybrid deep learning framework that integrates the strengths of TransUNet and RSU7 architectures. TransUNet excels in global context modeling through its transformer encoder-decoder structure, while RSU7 is renowned for its robust feature extraction capabilities, particularly in handling intricate image features. TR7Net synergizes these two architectures to achieve superior segmentation performance. Extensive experiments on spinal endoscopic datasets demonstrate that TR7Net outperforms both TransUNet and nnUNet in terms of segmentation accuracy and robustness in surgical scenarios. This work presents a significant advancement in medical image segmentation for spinal endoscopic surgery, offering a more precise and reliable solution for surgical assistance.

Keywords: Spinal endoscopic surgery, Medical image segmentation, Deep learning, TransUNet, RSU7.

1 Introduction

Medical image segmentation [8, 16] is essential in medical imaging, playing a key role in diagnosis, treatment planning, and surgical navigation. In spinal endoscopic surgery [5, 6, 14, 15] in **Fig. 1**, the ability to accurately segment and identify anatomical structures and surgical instruments from endoscopic images is crucial for real-time guidance [14] and robotic assistance. However, traditional segmentation methods, such as the widely used U-Net architecture [7, 9, 11, 18], often fail to capture long-range

Shao-Chi Pao and Liuyi Yang contributed equally to this work. *Corresponding authors: Xiaoxing Yang and Bingding Huang.

dependencies and fine-grained details, essential for precise segmentation. This limitation has motivated the development of more advanced segmentation techniques.

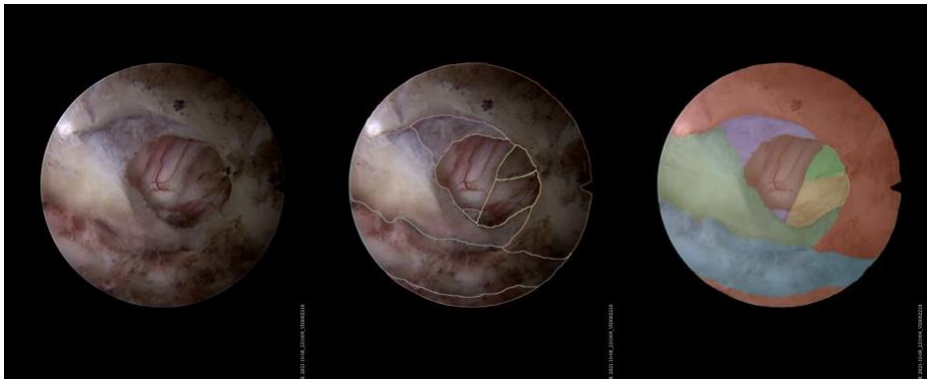

Fig. 1. Spinal endoscopic surgery: This image illustrates the three stages of endoscopic surgical image segmentation: the leftmost image presents the raw view captured during an endoscopic surgery, revealing the tissue structures within the surgical area; the middle image overlays white contour lines on the raw image to delineate the boundaries of the target regions that require segmentation; and the rightmost image displays the final segmentation outcome, where different colors are applied to the original image to distinctly differentiate various tissue areas.

Recent advances in deep learning have introduced Transformer-based architectures [2, 17] that enhance global context modeling through self-attention mechanisms [10]. TransUNet [1], a pioneering model that integrates Transformers [2] into the U-Net [7] framework, has demonstrated superior performance in various medical imaging tasks by leveraging global context extraction and cross-attention refinement. Meanwhile, RSU7 [3] (Recurrent Skip Connections with Multiscale Features) has emerged as a robust solution for extracting features in complex medical images. RSU7 combines multiscale feature extraction and residual learning, making it suitable for handling intricate anatomical structures in spinal endoscopic images. Building on RSU7, U²-Net [3, 4, 12] has shown promising results in medical image segmentation tasks, further highlighting the potential of RSU-based architectures [13].

Despite these advancements, existing models still face challenges in balancing global context modeling and local feature extraction, which are both critical for accurate segmentation. To address these limitations, we propose TR7Net (TransUNet-RSU7-Network), a hybrid deep learning framework that integrates the strengths of TransUNet and RSU7. TR7Net combines the global context modeling capabilities of TransUNet with the robust feature extraction of RSU7 to achieve superior segmentation performance. Our contributions include (1) the development of TR7Net, a novel hybrid architecture that synergizes TransUNet and RSU7; (2) extensive experiments on spinal endoscopic datasets to validate the effectiveness of TR7Net; and (3) a comprehensive analysis of the model's robustness and generalizability in diverse surgical scenarios. Our results show that TR7Net outperforms both TransUNet and U2Net, highlighting its potential as a powerful tool for medical image segmentation in spinal endoscopic

surgery. This work represents a significant step in advancing surgical assistance through more precise and reliable segmentation techniques.

2 Related Work and Background

2.1 TransUNet

TransUNet represents a hybrid architecture that combines the strengths of CNN [7] and Transformers [2], thus overcoming the limitations of traditional segmentation networks. The architecture comprises an encoder-decoder structure, with the encoder based on the Transformer framework and the decoder inspired by the U-Net architecture.

Transformer Encoder The encoder in TransUNet [1] utilizes a Transformer architecture, which excels at capturing global contextual information. The input image is divided into patches that are subsequently embedded into a sequence of tokens. These tokens are processed through multiple Transformer layers, each comprising self-attention mechanisms [10] and feedforward networks. This design enables the model to capture long-range dependencies and global features effectively.

U-Net Decoder The decoder in TransUNet is inspired by U-Net, which is known for its ability to preserve spatial details. The decoder up-samples the feature maps from the encoder and integrates them with high-resolution feature maps from the encoder. This integration of global and local features enhances the model's capacity for precise segmentation.

Integration of CNN and Transformer TransUNet leverages the strengths of both CNNs and Transformers by employing CNNs to extract low-level features and Transformers to capture high-level contextual information. This hybrid approach ensures the model can effectively handle global and local features, thus improving segmentation accuracy.

2.2 RSU7

RSU7 (Recurrent Skip Unit 7) [3] is a specialized module that improves the performance of feature extraction and segmentation. The RSU7 architecture is characterized by its ability to capture multiscale features and preserve spatial details.

Multi-Scale Feature Extraction RSU7 incorporates multiple convolutional layers with varying receptive fields to capture features at different scales. This design enables the module to handle objects of diverse sizes and shapes effectively, making it suitable for segmentation tasks requiring precise boundary detection.

Recurrent Mechanism The recurrent mechanism in RSU7 allows the module to refine the feature maps iteratively. This process helps to preserve spatial details and reduce information loss during feature extraction.

Integration with Encoder-Decoder Framework RSU7 can be seamlessly integrated into the encoder-decoder architecture of segmentation models. By incorporating RSU7 units in both the encoder and the decoder, the model can improve its ability to capture global and local characteristics, leading to improved segmentation accuracy.

The proposed TR7Net architecture effectively integrates the strengths of TransUNet and RSU7, achieving superior performance in complex segmentation tasks. By combining the global attention mechanisms of TransUNet with the multi-scale feature extraction capabilities of RSU7, TR7Net can capture long-range dependencies and fine-grained details. This integration ensures more accurate and robust segmentation results. In general, TR7Net takes advantage of both architectures to address the limitations of traditional segmentation models, providing a powerful framework for various segmentation tasks. Future work will focus on evaluating the performance of TR7Net on diverse datasets and exploring further enhancements to the architecture.

3 Integrating TransUNet and RSU7

The development of TR7Net in **Fig. 2** represents a significant evolution from the traditional TransUNet architecture [1], focusing on enhancing segmentation performance through innovative modifications and integrations. While maintaining the core strengths of TransUNet, TR7Net introduces several architectural advances that address key challenges in image segmentation [8, 16], particularly in medical imaging contexts where class imbalance and complex boundaries are prevalent.

3.1 Architectural Innovations in TR7Net

TR7Net builds on the hybrid structure of TransUNet, which combines the global attention capabilities of Vision Transformers (ViT) [2] with the preservation of spatial detail of U-Net decoders [1, 7, 9, 11, 12, 13, 18]. However, unlike the standard TransUNet, which relies heavily on skip connections to integrate encoder and decoder features, TR7Net adopts a more streamlined approach. Traditional skip connections are replaced with a feature fusion module that efficiently integrates multi-scale features without the redundancy often associated with skip connections. This modification not only simplifies the architecture, but also enhances the model's ability to handle complex segmentation tasks by providing a more coherent feature representation.

3.2 Enhanced Feature Extraction through RSU7 Integration

A key innovation in TR7Net is integrating the Recurrent Skip Unit 7 (RSU7) module, borrowed from the U²-Net architecture [3]. RSU7 is specifically designed to enhance multiscale feature extraction, a critical requirement for accurately segmenting objects

with varying sizes and complex boundaries. Unlike traditional convolutional layers that may struggle to capture diverse scales, RSU7 employs a series of nested convolutional blocks with varying receptive fields. This design in **Fig. 3** allows TR7Net to effectively capture fine-grained details and broader contextual information, thus improving segmentation accuracy. The addition of RSU7 to the TR7Net encoder-decoder framework ensures that the model can robustly handle segmentation tasks on different scales, making it particularly suitable for medical images where precise boundary detection is essential.

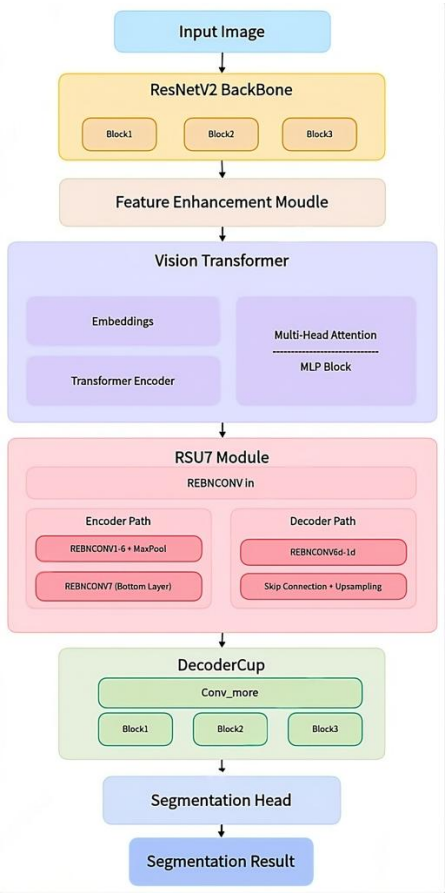

Fig. 2. TR7Net: The input layer receives the images, and the ResNetV2 backbone extracts fundamental features using three Block modules. The features are then enhanced by a dedicated module and further processed by a Vision Transformer incorporating embeddings, encoders, multi-head attention, and MLP Blocks to capture global dependencies. The RSU7 module refines the features with an encoder-decoder structure, utilizing REBNCONV layers and skip connections. The DecoderCup decodes the features into segmentation results via three blocks and a Conv_more layer. Finally, the segmentation head generates the output, producing the final segmentation mask for the endoscopic images.

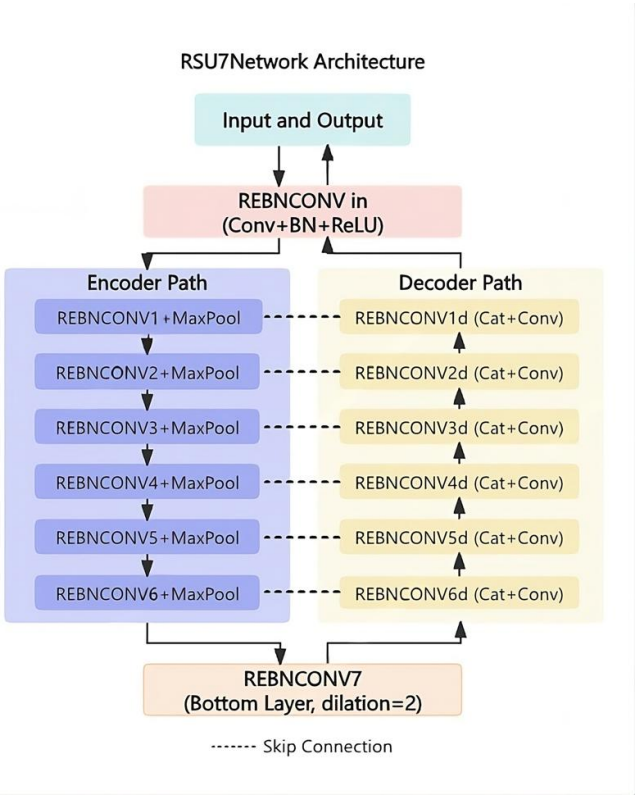

Fig. 3. RSU7 network architecture, a U-shaped architecture, is optimized for image segmentation, especially in medical imaging. It begins with an REBNCONV input layer for feature extraction. The encoder path, comprising six REBNCONV layers with MaxPooling, downsamples the image, while the REBNCONV7 layer at the base uses dilated convolution to broaden the receptive field. The decoder path, with layers from REBNCONV6d to REBNCONV1d, uses Cat+Conv operations and skip connections to integrate and upscale features, facilitating precise segmentation. This design equips RSU7 to tackle intricate segmentation challenges effectively.

3.3 Attention Mechanisms and Class Imbalance Mitigation

Another notable advancement in TR7Net is the incorporation of spatial attention mechanisms [10]. Unlike the standard TransUNet, which relies solely on the inherent self-attention of the Transformer encoder, TR7Net introduces specialized spatial attention gates. These gates are strategically placed within the architecture to focus the model's attention on critical regions of the input image. By selectively focusing on important features, TR7Net achieves a more refined segmentation output, especially in areas with complex textures and overlapping structures.

Addressing class imbalance is a common challenge in medical image segmentation, where certain classes may have significantly fewer instances than others. TR7Net tackles this issue by introducing learnable class weights, which dynamically adjust the model's focus during training. This adaptive weighting strategy ensures that the model pays

adequate attention to underrepresented classes, such as small vessels or rare tissues, thus improving overall segmentation performance. The initial weights for these minority classes are set higher to counteract their scarcity in the training data, leading to more balanced and accurate segmentation results.

3.4 Streamlined Decoder and Input Normalization

The decoder in TR7Net is streamlined to enhance efficiency and performance. Traditional U-Net decoders often rely on skip connections to restore lost spatial details during encoding. However, TR7Net replaces these connections with a more efficient feature fusion mechanism, reducing computational overhead while maintaining or even improving segmentation accuracy. This streamlined decoder architecture ensures the model remains lightweight and scalable, making it suitable for real-time applications and deployment on devices with limited computational resources.

Furthermore, TR7Net incorporates input normalization techniques that standardize the input data before processing. This normalization step, defined as the Equation (1).

$$Normalization = \frac{x-\mu}{\sigma+\epsilon} \tag{1}$$

where μ and σ are the mean and standard deviation of the input, respectively, and ϵ is a small constant to prevent division by zero, ensures that the model operates in a consistent feature space. This preprocessing step is crucial for handling diverse input data, including images with varying intensity distributions and channel configurations. By normalizing the input, TR7Net achieves greater robustness and generalizability across different data sets and imaging modalities.

TR7Net integrates the strengths of TransUNet with the multi-scale feature extraction capabilities of RSU7, offering improved performance in image segmentation tasks. The architectural modifications, including a streamlined decoder, spatial attention mechanisms, and adaptive class weighting, address key challenges in medical image segmentation. These enhancements contribute to more accurate and robust segmentation results. Future work will optimize the architecture and evaluate TR7Net's performance on a broader range of segmentation tasks.

4 Performance Evaluation

To evaluate the performance of TR7Net, we compare it with two state-of-the-art models: nnUNet [18] and TransUNet [1]. The evaluation is based on several key metrics: Precision, Recall, and Dice coefficient. These metrics provide a comprehensive assessment of the ability of the models to segment medical images accurately. The spinal endoscopy dataset used in this study was obtained from the ISICDM 2024 challenge.

4.1 Evaluation Metric

Precision The proportion of correctly predicted pixels among all predicted pixels for a given class. Precision measures the model's accuracy in predicting positive instances, where TP denotes true positives and FP denotes false positives. The precision is calculated as:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

Recall The proportion of correctly identified pixels among all true pixels for a given class. Recall measures the model's ability to detect all relevant instances, where FN denotes false negatives. It is calculated as:

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

Dice Coefficient
A measure of overlap between the predicted segmentation and the ground truth. The Dice coefficient provides a balanced measure of segmentation accuracy, considering both precision and recall. The dice is defined as:

$$\text{Dice} = \frac{2\times TP}{2\times TP+FP+FN} \tag{4}$$

4.2 Results and Discussion

Precision and Recall TR7Net exhibited a balanced trade-off between precision and recall, outperforming nnUNet and TransUNet. This suggests that TR7Net predicts more accurate positive instances and detects a higher proportion of true positives.

Dice Coefficient TR7Net achieved the highest Dice coefficient, indicating a better overlap between the predicted segmentation and ground-truth labels. This metric highlights TR7Net's ability to produce more accurate and refined segmentations, particularly for complex anatomical boundaries.

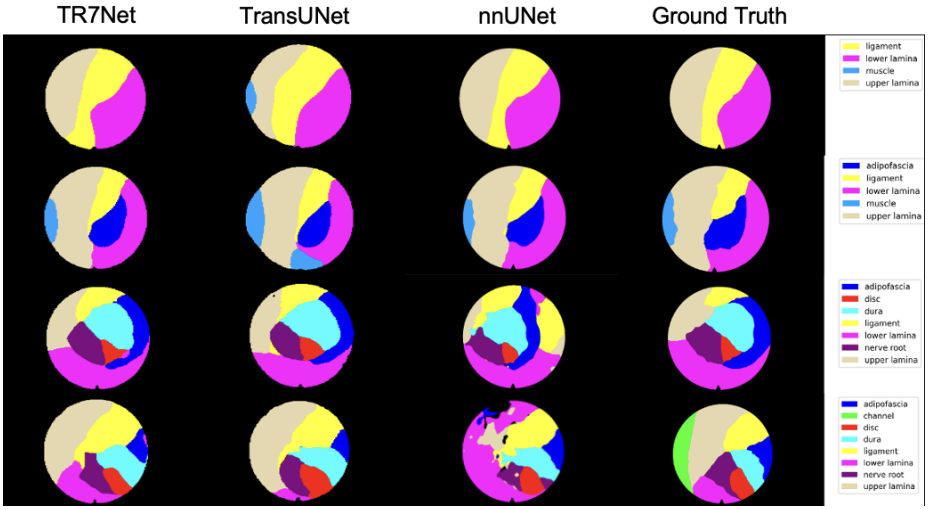

Fig. 4. Comparison of TR7Net, TransUNet and nnUNet

Observations from the images indicate that the TR7Net's segmentation outcomes closely align with the Ground Truth in most instances, demonstrating superior segmentation accuracy. Although TransUNet's overall performance is acceptable, there are slight discrepancies in segmentation results compared to the Ground Truth in certain areas. The nnUNet, however, exhibits a significant segmentation error in the last case, with its segmentation results showing a noticeable deviation from the actual conditions.

Table 1. Performance Result

	TR7Net	TransUNet	nnUNet
Precision	**0.6770**	0.6503	0.6304
Recall	**0.6456**	0.6339	0.6256
Dice	**0.6150**	0.5946	0.5871

The experimental results, as presented in **Fig. 4** and **Table 1**, indicate that TR7Net consistently outperforms both TransUNet and nnUNet across all evaluated metrics. This superior performance can be attributed to integrating TransUNet's global context modeling capabilities effectively with RSU7's robust feature extraction strengths. Particularly noteworthy is TR7Net's improvement in the Dice coefficient, which is approximately 2.04% higher than TransUNet and 2.79% higher than nnUNet. These improvements are clinically significant in spinal endoscopic surgery, where precise segmentation can directly impact surgical decision-making and patient outcomes. The balanced performance in precision and recall further indicates that TR7Net provides more reliable segmentations with fewer false positives and negatives, which is crucial for minimizing surgical risks. These results validate our hypothesis that a hybrid architecture that combines transformer-based global attention with recursive CNN-based local

feature extraction can effectively address the challenges of endoscopic image segmentation.

5 Conclusion and Future Work

5.1 Conclusion

In this paper, we present TR7Net, a novel hybrid deep learning framework that integrates the strengths of TransUNet and RSU7 architectures for medical image segmentation in spinal endoscopic surgery. Our experimental results demonstrate that TR7Net consistently outperforms state-of-the-art models, including TransUNet and nnUNet, across all evaluation metrics. The superior performance of TR7Net can be attributed to its unique architecture that combines the global context modeling capabilities of transformers with the robust feature extraction strengths of recursive CNN structures.

The proposed framework effectively addresses the challenges of endoscopic image segmentation, particularly in capturing complex anatomical details and long-range contextual information. By achieving higher precision, recall, and Dice coefficient scores, TR7Net provides more accurate and reliable segmentations that can significantly improve surgical planning, navigation, and real-time guidance in spinal endoscopic procedures. This improvement in segmentation performance has the potential to reduce surgical risks and improve patient outcomes in minimally invasive spinal surgeries.

5.2 Future Work

While TR7Net demonstrates promising results, several directions for future research can further enhance its capabilities:

- **Multi-modal Integration:** Incorporating additional imaging modalities, such as MRI or CT scans alongside endoscopic images, could provide complementary information and further improve segmentation accuracy.
- **Real-time Optimization:** Further optimization of the model architecture and implementation to achieve real-time performance on standard surgical hardware, making it more practical for intraoperative use.
- **Domain Adaptation:** Developing techniques to adapt the model to different endoscopic procedures beyond spinal surgery, such as gastrointestinal or gynecological endoscopy, to expand clinical applicability.
- **Uncertainty Estimation**: Integrating uncertainty estimation mechanisms to provide confidence levels for segmentation results, which could help surgeons make more informed decisions during procedures.
- **Clinical Validation:** Conducting comprehensive clinical validation studies to assess the impact of TR7Net on surgical outcomes and workflow efficiency in real-world clinical settings.

These future directions aim to enhance the technical capabilities and clinical utility of TR7Net, potentially establishing it as a valuable tool in the growing field of computer-assisted minimally invasive surgery.

Funding. This work was supported by the Project of the Educational Commission of Guangdong Province of China (No. 2022ZDJS113), the Shenzhen Science and Technology Program (No. KJZD20240903095605007), and the Higher Education Stability Support Program General Project (No. 20220715114836001).

Acknowledgments. We sincerely thank the ISICDM2024 organizing committee for providing the spinal endoscopy dataset from Challenge Project 3, which has greatly supported this research.

References

1. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ...& Zhou, Y. (2021). Tran-sUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv preprint arXiv:2102.04306.
2. Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., & Fu, H. (2023). Transformers in medical imaging: A survey. Medical Image Analysis, 85, 102802
3. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., & Jagersand, M. (2020). U²-Net: Going deeper with nested U-structure for salient object detection. Pattern Recognition, 106, 107404.
4. Liu, Z., Zhang, Z., & Chen, X. (2021). U²-Net-Based Segmentation for Medical Images. IEEE Access, 9, 123451–123461.
5. Ahn, Y., & Lee, S. H. (2018). Endoscopic spine surgery: Review of current trends and future directions. World Neurosurgery, 109, 134-141.
6. Wang, Y., Dou, Q., Heng, P. A., Liu, Y., & Wang, Y. (2021). Automatic segmenta-tion and tracking of surgical instruments in endoscopic spinal surgery. IEEE Trans-actions on Medical Imaging, 40(10), 2891–2901.
7. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015), Springer, 234–241.
8. Hesamian, M. H., Jia, W., He, X., & Kennedy, P. (2019). Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. Journal of Digital Imaging, 32(4), 582–596.
9. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2023). Swin-Unet: Unet-like pure transformer for medical image segmentation. European Conference on Computer Vision (ECCV), 205–218.
10. Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., ... & Rueckert, D. (2018). Attention U-net: Learning where to look for the pancreas. Medical Imaging with Deep Learning (MIDL).
11. Hatamizadeh, A., Yang, D., Roth, H., & Xu, D. (2022). UNETR: Transformers for 3D Medical Image Segmentation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 574–584.
12. Yang, J., Liu, Y., Zhang, Y., & Zhao, X. (2022). Improved U²-Net for Medical Image Segmentation. IEEE Access, 10, 10234–10245.
13. Zhang, Z., Liu, Q., & Wang, Y. (2022). Embedded RSU-Net: A Lightweight U²-Net-Based Architecture for Real-Time Medical Image Segmentation. Computers in Biology and Medicine, 141, 105125.
14. Wang, J., Yang, X., & Guo, Y. (2022). Real-time Surgical Instrument Segmentation and Tracking in Endoscopic Spine Surgery. Computer Methods and Programs in Biomedicine, 224, 107004.

15. Li, X., Zhang, H., & Wang, L. (2022). Artificial intelligence in endoscopic spinal surgery: current applications and future directions. European Spine Journal, 31(5), 1205–1216.
16. Ma, Y., Lei, B., Wang, S., Hayat, M., & Liu, Y. (2022). Deep Learning in Medical Image Segmentation: Advances and Challenges. Neurocomputing, 493, 199–214.
17. Li, Y., Jiang, Y., Li, J., Chen, X., Li, Y., & Xu, Z. (2023). Transformers in Medical Image Analysis: A Systematic Review. Medical Image Analysis, 87, 102834.
18. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., ... & Maier-Hein, K. H. (2020). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods, 18(2), 203-211