Biomedical Data Science
2025, VOL. 5, NO. 3, 82-85
DOI: 10.47297/wspbdsWSP2752-630516.20250503

Transformer-Based Lung CT Image Segmentation

Jie Cao

Faculty of Science and Technology, Beijing Normal-Hong Kong Baptist University (BNBU), Zhuhai, Guangdong, 519087, China

ABSTRACT
Transformer-based models have emerged as a breakthrough in lung CT image segmentation, offering superior performance over traditional CNNs by effectively capturing long-range dependencies through self-attention mechanisms. These models excel in segmenting complex lung structures and pathologies, including pulmonary nodules, COVID-19 infection regions, and ground-glass opacities, demonstrating particular robustness in analyzing scattered lesions that challenge conventional approaches. Current implementations primarily follow two paradigms: hybrid architectures that integrate CNN's local feature extraction with Transformer's global modeling (e.g., TransUNet, Swin UNETR), and pure Transformer models (e.g., UNETR, ViT), both achieving state-of-the-art results on benchmark datasets like LUNA16 and MosMed. Despite these advances, several challenges persist, including high computational demands, dependence on large annotated datasets, and occasional limitations in capturing fine local details. Researchers are addressing these issues through innovations in sparse attention mechanisms, efficient model architectures, and data augmentation techniques. Looking ahead, key development areas include lightweight model deployment for clinical settings (e.g., MobileViT adaptations), multimodal fusion strategies combining CT with other imaging modalities, and privacy-preserving federated learning approaches. As 3D Transformer architectures mature and incorporate more domain-specific knowledge, they are poised to transform clinical applications ranging from early lung cancer detection to quantitative COVID-19 assessment. However, achieving optimal balance between computational efficiency, segmentation accuracy, and clinical interpretability remains an ongoing challenge that will shape future research directions in this rapidly evolving field.

KEYWORDS
Transformer; Medical image; Image segmentation; Lung CT image

1 Introduction

1.1 Research background

Lung CT image segmentation is a crucial research direction in the field of medical image analysis. Accurate segmentation of lung images plays a pivotal role in disease diagnosis, treatment planning, and prognosis evaluation. Particularly after the COVID-19 pandemic, the importance of lung image analysis has become even more prominent, as assessing the severity of lung lesions is critical for treatment decisions.

In recent years, Transformer models have achieved groundbreaking progress in natural language processing (e.g., large language models) and computer vision, gradually being introduced into the medical field, including CT image segmentation.

However, COVID-19 and other lung diseases remain among the leading causes of morbidity and mortality worldwide. Especially after the COVID-19 pandemic, rapid and accurate analysis of lung images has become a clinical necessity.

Current lung CT image segmentation still faces several technical challenges, and the application of Transformer models can contribute significantly to addressing these issues:

Data annotation difficulties and subjective manual labeling: Lung CT images require annotations by professional physicians, which is time-consuming. Moreover, different physicians may exhibit subjective variability when processing multiple CT images, leading to segmentation errors and inaccurate diagnoses. Transformer models can mitigate this issue through self-supervised learning and pre-training, reducing reliance on annotated data while significantly improving medical diagnostic efficiency [1].

Inaccurate spatial alignment and insufficient local feature capture in existing models: Hybrid architectures combining CNN and Transformer can enable automatic scoring and classification of multi-region lung lesions, enhancing diagnostic efficiency [2].

1.2 Research objectives

This paper aims to systematically review the research progress of Transformer-based lung CT image segmentation, highlight the advantages of Transformer in medical image analysis, analyze its limitations, and provide insights for future improvements in medical image efficiency. The study also offers theoretical support and guidance for future research.

1.3 Research scope

This paper examines the advancements of Transformer models in lung image analysis over the past five years, including the development of the models themselves, their applications in the medical field, and the integration of these two aspects.

2 Main content

2.1 Development of transformer technology

In 2017, the Transformer architecture, based entirely on self-attention mechanisms, was first proposed. This architecture abandoned traditional recurrent and convolutional structures, adopting an encoder-decoder framework with positional encoding to capture dependencies between different positions [3].

From 2018 to 2019, BERT and GPT were developed based on Transformer. BERT utilized the Transformer encoder for bidirectional pre-training through masked language modeling (MLM) and next sentence prediction (NSP), while GPT employed the Transformer decoder for unidirectional pre-training via autoregressive language modeling [4-5].

In 2021, Vision Transformer (ViT) was introduced, which divided images into patches and processed them as sequences [6]. ViT demonstrated that pure Transformer structures could surpass traditional CNNs in image classification tasks, showcasing global modeling capabilities and scalability, albeit with challenges such as high data dependency and computational costs.

From 2021 to 2022, hybrid architectures combining Transformer with RNN/CNN emerged. Today, large-scale models like ChatGPT and DeepSeek, based on OpenAI's technology and leveraging trillion-parameter Transformers optimized through reinforcement learning from human feedback (RLHF), have become prominent.

2.2 Technical principles and mechanisms of transformer

Transformer is a neural network architecture based on self-attention mechanisms[3]. It has revolutionized natural language processing (NLP) and computer vision (CV), replacing traditional RNN and CNN structures.

The Transformer architecture follows an encoder-decoder framework, with its core modules being Multi-Head Self-Attention (MHSA) and Feed-Forward Network (FFN). The input includes word embeddings and positional encoding. The encoder consists of N identical layers, each containing MHSA, FFN, residual connections, and layer normalization. The decoder also comprises N identical layers but includes an additional masked self-attention mechanism to ensure that only preceding tokens are visible during decoding. The final output layer generates the results[6].

Key features of the Transformer model include parallel computation and long-range dependency modeling. Parallel computation is enabled by positional encoding, allowing self-attention to process all positions simultaneously, significantly improving training speed. Long-range dependency modeling is achieved through self-attention, which directly captures relationships between any positions, overcoming the limitations of CNN's local receptive fields and RNN's gradient vanishing issues.

2.3 Comparison between transformer (ViT) and other technologies

In lung CT image segmentation, Transformer technology, particularly ViT, is often compared with traditional CNN and RNN.

For CNN, representative models like U-Net and V-Net excel in local feature extraction and computational efficiency, making them suitable for texture and edge detection in medical images. However, their limitations include insufficient long-range dependency modeling and spatial invariance assumptions. For instance, in lung CT scans, lesions may be scattered across different regions (e.g., bilateral ground-glass opacities in COVID-19), but CNN's limited receptive fields may miss global information. Additionally, the variability in lesion location and shape in medical images may reduce segmentation accuracy due to CNN's translation invariance.

For RNN, processing 3D CT data requires layer-by-layer computation, which cannot be parallelized. This becomes particularly problematic for high-resolution CT scans, potentially leading to information loss.

Transformer technology, especially ViT, offers unique advantages in CT image segmentation.

ViT's self-attention mechanism directly computes relationships between all image regions without expanding receptive fields. For example, in COVID-19 segmentation, ViT can simultaneously capture lesion correlations in both lungs, whereas CNNs are limited to local regions. Models like TransUNet, which combine ViT and U-Net, have outperformed pure CNN methods on multiple medical datasets.

3D ViT variants, such as Swin UNETR, extend Transformer to 3D volumetric data, surpassing traditional CNN and RNN.

These models can uniformly process axial, coronal, and sagittal views of CT data, excelling in lung CT image analysis.

2.4 Applications of transformer

Transformer has been widely applied across various domains since its inception in 2017, expanding from natural language processing to computer vision, speech, multi-modal learning, bio-medicine, and reinforcement learning. It has become the core architecture for many State-of-the-Art (SOTA) models, which is why this paper focuses on Transformer-based lung CT image segmentation.

In natural language processing, OpenAI's GPT series is the most well-known example, enabling tasks like text generation, code completion, and dialogue, significantly improving productivity. Other applications include translation, text summary, and comprehension.

In computer vision, ViT has demonstrated superior performance in image classification, surpassing traditional CNNs. For object detection, Swin Transformer employs hierarchical window attention, making it suitable for high-resolution images. In image segmentation, models like TransUNet are specifically designed for medical image segmentation, with notable applications in lung CT analysis.

Beyond these, Transformer is also widely used in speech processing (e.g., multilingual speech recognition, translation, and generation) and multi-modal tasks (e.g., image-text matching, video understanding, and multi-modal generation).

The ubiquity of Transformer stems from its versatility. Its self-attention mechanism can model any sequence, including text, images, and speech. Its parallel computation is more efficient than RNN, and its long-range dependency modeling overcomes the locality limitations of CNN and RNN. Additionally, its scalability allows continuous performance improvement by increasing parameters. In summary, Transformer is an efficient and superior model.

2.5 Current research on transformer-based lung CT image segmentation

In recent years, Transformer has demonstrated strong potential in medical image segmentation, particularly in lung CT image segmentation, where its global modeling capability significantly improves lesion and organ boundary recognition accuracy.

International research primarily focuses on hybrid architectures and 3D segmentation. In 2021, Johns Hopkins University proposed TransUNet, the first model combining Vision Transformer with U-Net, introducing a CNN-Transformer hybrid encoder. It outperformed traditional CNNs in tasks like multi-organ segmentation (Synapse dataset) and COVID-19 lung infection segmentation [7]. In 2022, NVIDIA developed UNETR, a pure Transformer-based 3D segmentation model capable of directly processing volumetric CT data without a CNN backbone [8]. The same year, NVIDIA also introduced Swin UNETR, a 3D medical image segmentation framework based on Swin Transformer, suitable for lung CT and brain MRI [9].

Domestic research emphasizes lightweight design and local-global feature fusion. In 2021, Tsinghua University embedded lightweight Transformer modules to optimize COVID-19 lung infection segmentation, achieving a Dice coefficient of 0.812 [10]. In 2022, the Chinese Academy of Sciences proposed the Dense-Transformer Network, combining dense connections with Transformer to enhance pulmonary nodule segmentation accuracy [11]. Shanghai Jiao Tong University designed a gated positional encoding Transformer for 3D medical images, effectively addressing spatial information loss [12].

Current research trends focus on efficient 3D modeling (e.g., hierarchical Transformer), small-sample generalization (self-supervised pre-training), and multi-modal fusion. However, computational efficiency and data privacy remain key challenges. Future directions may expand into federated learning and edge device deployment (e.g., MobileViT variants).

2.6 Limitations and challenges

Despite its strengths, Transformer-based medical image segmentation faces several limitations:

High computational complexity: The self-attention mechanism's computational cost scales quadratically with input image size, slowing training and inference. High-resolution 3D CT data demands significant GPU memory, making it impractical for clinical settings requiring rapid segmentation.

Dependence on large-scale annotated data: Medical data is scarce, and Transformer models require vast datasets for pre-training. Annotating lung CT images is time-consuming and costly, often leading to overfitting and poor generalization with limited data.

Weak local detail capture: While Transformer excels at global dependency modeling, it may overlook local structures (e.g., textures, edges), which are critical for medical image segmentation.

To address these challenges, researchers are exploring sparse attention mechanisms and hybrid CNN-Transformer architectures. Although Transformer holds great promise for lung CT segmentation, further optimization tailored to

medical image characteristics is necessary to achieve optimal results.

3 Conclusion

In recent years, Transformer-based lung CT image segmentation has become a vital research direction in medical image analysis. Its core advantage lies in modeling global contextual information through self-attention, significantly improving the segmentation accuracy of lung lesions (e. g., pulmonary nodules, COVID-19 infections) and organ boundaries. This technology overcomes the limitations of traditional CNNs in long-range dependency modeling, particularly excelling in handling scattered lesions (e.g., bilateral ground-glass opacities).

Current research has developed two main approaches: hybrid architectures (e.g., TransUNet, Swin UNETR) combining CNN's local features with Transformer's global modeling, and pure Transformer models (e. g., UNETR, ViT). Both have outperformed traditional methods on public datasets (e. g., LUNA16, MosMed). However, challenges such as high computational complexity, reliance on large-scale annotated data, and weak local detail capture persist, driving research toward sparse attention and hybrid architectures.

Future research will focus on lightweight deployment (e. g., MobileViT variants), multimodal fusion (e. g., PET-CT + Transformer), and federated learning to address real-time performance, data scarcity, and privacy issues. As 3D Transformer architectures evolve and integrate more medical prior knowledge, this technology is expected to play a greater role in clinical applications like early lung cancer diagnosis and COVID-19 assessment. Nevertheless, balancing efficiency with accuracy and enhancing interpretability for clinical needs remain critical for future advancements.

References

[1] Chen T., Kornblith S., Norouzi M., & Hinton G. (2020). A Simple Framework for Contrastive Learning of Visual Representations (SimCLR). Proceedings of the 37th International Conference on Machine Learning (ICML).

[2] Lee J. B., Kim J. S., & Lee H. G. 2024. COVID-19 to pneumonia: Multi-region lung severity classification using CNN transformer position-aware feature encoding network. In M. G. Linguraruet al. Eds., Medical image computing and computer-assisted.

[3] Vaswaniet al., "Attention Is All You Need", NeurIPS 2017.

[4] Devlinet al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", NAACL 2019.

[5] Radfordet al., "Improving Language Understanding by Generative Pre-training", OpenAI Blog 2018.

[6] Dosovitskiyet al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", ICLR 2021.

[7] Chen et al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation", arXiv 2021.

[8] Hatamizadehet al., "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images", MICCAI 2022.

[9] Hatamizadehet al., "UNETR: Transformers for 3D Medical Image Segmentation", CVPR 2022.

[10] Wang et al., "A Hybrid Transformer Network for COVID-19 Infection Segmentation in CT Images", IEEE TMI 2021.

[11] Li et al., "Dense-Transformer for Medical Image Segmentation", Medical Image Analysis 2022.

[12] Zhang et al., "Gated Positional Transformer for 3D Medical Image Segmentation", AAAI 2023.