Review Article

Transformers in medical image segmentation: a narrative review

Rabeea Fatma Khan[1#], Byoung-Dai Lee[1#], Mu Sook Lee[2]

[1]Department of Computer Science, Graduate School, Kyonggi University, Suwon, Republic of Korea; [2]Department of Radiology, Keimyung University Dongsan Hospital, Daegu, Republic of Korea

Contributions: (I) Conception and design: BD Lee, MS Lee; (II) Administrative support: MS Lee; (III) Provision of study materials or patients: All authors; (IV) Collection and assembly of data: RF Khan, BD Lee; (V) Data analysis and interpretation: RF Khan, BD Lee; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Mu Sook Lee, MD, MS. Department of Radiology, Keimyung University Dongsan Hospital, 1035, Dalgubeol-daero, Sindang-dong, Daegu 24601, Republic of Korea. Email: musukilee@kmu.ac.kr.

Background and Objective: Transformers, which have been widely recognized as state-of-the-art tools in natural language processing (NLP), have also come to be recognized for their value in computer vision tasks. With this increasing popularity, they have also been extensively researched in the more complex medical imaging domain. The associated developments have resulted in transformers being on par with sought-after convolution neural networks, particularly for medical image segmentation. Methods combining both types of networks have proven to be especially successful in capturing local and global contexts, thereby significantly boosting their performances in various segmentation problems. Motivated by this success, we have attempted to survey the consequential research focused on innovative transformer networks, specifically those designed to cater to medical image segmentation in an efficient manner.

Methods: Databases like Google Scholar, arxiv, ResearchGate, Microsoft Academic, and Semantic Scholar have been utilized to find recent developments in this field. Specifically, research in the English language from 2021 to 2023 was considered.

Key Content and Findings: In this survey, we look into the different types of architectures and attention mechanisms that uniquely improve performance and the structures that are in place to handle complex medical data. Through this survey, we summarize the popular and unconventional transformer-based research as seen through different key angles and analyze quantitatively the strategies that have proven more advanced.

Conclusions: We have also attempted to discern existing gaps and challenges within current research, notably highlighting the deficiency of annotated medical data for precise deep learning model training. Furthermore, potential future directions for enhancing transformers' utility in healthcare are outlined, encompassing strategies such as transfer learning and exploiting foundation models for specialized medical image segmentation.

Keywords: Transformers; medical imaging; deep learning; artificial intelligence (AI); image segmentation

Submitted May 11, 2023. Accepted for publication Sep 14, 2023. Published online Oct 07, 2023.
doi: 10.21037/qims-23-542
View this article at: https://dx.doi.org/10.21037/qims-23-542

Introduction

One of the critical aspects of medical image analysis wherein artificial intelligence (AI) is used to boost performance is image segmentation, which is designed to segment out important organs and abnormal objects of the human body such as lungs, nodules, tumors, etc. A good segmentation result is highly useful when performing medical operations such as surgical planning, as well as in the diagnosis and prognosis of diseases, since it can help with outlining and pinpointing the exact location of the object, along with the determination of other properties such as size, volume, etc. The use of these AI-based solutions can significantly and efficiently reduce the time taken for these procedures (1).

Traditional handcrafted approaches based on image processing techniques such as simple thresholding of Hounsfield unit (HU) values and template matching have exhibited poor results due to a lack of robustness to perpetually varying medical images (2). The main performance boost arrived in the form of deep learning, such as convolutional neural networks (CNNs), which have achieved considerably improved results and such deep learning methods have led to a whole new domain of approaches for the segmentation problem. The ability of deep learning algorithms to learn and overcome data variations in medical data by generalizing has increased the quality of modern AI-based medical imaging systems (3). However, CNN-based approaches generally have certain limitations when modeling of long-range dependencies that are present in an image (4).

Recently, the transformer technique (5) has emerged as a prominent approach in sequence modeling. Originally introduced in natural language processing (NLP), transformers employ self-attention to overcome memory constraints and capture long-range dependencies. Unlike previous methods like gated neural networks (6), recurrent neural networks (RNNs), and long-short term memory (7), which faced limitations in memory and long-range dependencies, transformers revolutionized sequence modeling. By utilizing self-attention and discarding RNNs, transformers enable global dependency modeling and parallelization. This not only eliminates the impact of distance between input and output sequences but also significantly enhances computational efficiency.

After the transformer approach gained popularity in language processing, it quickly extended its reach to other AI domains such as audio (8) and vision (9). In the medical domain, transformers have been employed for various tasks including classification and detection (10,11), extracting information from clinical notes (12), and segmentation (13). This has posed a significant challenge to existing CNN-based solutions (14), which are currently the state-of-the-art in this field. Extensive research has focused on improving the precision and reliability of transformer-based solutions, leading to several surveys that summarize the current research and future directions of these networks (9,15). However, there is a lack of surveys specifically tailored to the medical imaging domain, despite its unique complexities and differences from other vision transformer networks. This significant difference arises from the many complexities that come with medical data.

Medical data is distinct in several aspects, including its high level of pixelation resulting from advanced imaging techniques like computed tomography (CT) (16), magnetic resonance imaging (MRI) (17), or X-ray (18). However, acquiring medical data poses challenges, particularly related to patient privacy regulations and the need for extensive and accurate annotation (19,20). Furthermore, the three-dimensional (3D) nature of medical imaging requires specialized expertise to extract cross-plane contextual information efficiently. In addition to the scarcity of data, computational efficiency is crucial in medical imaging to achieve real-time segmentation for prompt diagnosis. Moreover, medical images contain valuable metadata beyond pixel values, such as contrast and manufacturing information. Consequently, models and networks developed for medical imaging differ significantly from generalized camera imaging models (21).

Furthermore, transparency and reliability are crucial in medical applications. Transparency ensures the interpretability of computational models, enabling healthcare professionals to understand their reasoning and make informed decisions. Reliability involves rigorous validation procedures, adherence to standards, and transparent documentation, fostering confidence in the technology's dependability, safety, and ethical implications. Therefore, medical applications require heightened accountability, regulatory compliance, and ethical considerations to prioritize patient safety.

The unique set of challenges present in medical imaging has necessitated distinct approaches and solutions, setting it apart from other vision applications. Transformer-based networks have emerged as formidable competitors to CNNs as the state-of-the-art in the medical domain, primarily driven by some limitations inherent in CNNs. CNNs struggle with capturing long-range dependencies due to their reliance on local receptive fields and hierarchical

Table 1 Methods of research

Items	Specification
Date of search	23rd March 2023
Databases and other sources searched	Google Scholar, arXiv, ResearchGate, Microsoft Academic, and Semantic Scholar
Search terms used	"Medical" and "transformers" and "segmentation" and "imaging" or "vision"
Timeframe	2021 to 2023
Inclusion criteria	Medical image segmentation networks that include transformers, published manuscripts, pre-print articles, English language
Exclusion criteria	Classification networks, non-medical image segmentation networks, networks where transformers are not a prominent part of the model architecture, unpublished manuscripts, conference abstracts only
Selection process	Authors Khan RF and Lee BD performed joint selection

feature extraction. In contrast, transformers excel in modeling global interactions between image regions by utilizing self-attention mechanisms. They leverage attention mechanisms to simultaneously process all positions in the input, facilitating a holistic understanding of image context and comprehending contextual relationships. Transformers also incorporate positional encoding, which provides explicit spatial information to the model, enabling it to better handle spatial relationships between image elements. This positional encoding is lacking in CNNs, where spatial information is implicitly encoded through convolutional operations. Additionally, transformers exhibit adaptability in adjusting receptive fields, allowing them to selectively attend to relevant regions and incorporate contextual information adaptively (22).

Despite the advantages offered by transformers in certain aspects, they also exhibit some limitations. Particularly, transformers tend to face challenges when dealing with small datasets, where CNNs currently maintain their status as the state-of-the-art approach. Consequently, it becomes essential to investigate network architectures that incorporate both convolutional layers and transformers, combining the strengths of both paradigms to potentially achieve improved performance and address the limitations of each individual model (23).

In response to the abundance of research in the field of medical image segmentation, it becomes imperative to comprehensively examine the various transformer-based models and strategies. Prior research in the domain of medical imaging has explored transformers, as evident in references (24,25). However, these studies have taken a broad approach, encompassing various aspects of medical imaging such as classification, detection, and segmentation. In contrast, our current paper uniquely concentrates exclusively on the topic of segmentation. This specialization enables us to conduct a comprehensive investigation into the application of transformers specifically for medical image segmentation, delving deeply into this area. It investigates diverse medical data types, architectural designs, smart self-attention strategies employed as well as multi-scale connections. Additionally, a quantitative assessment of selected networks is conducted to facilitate a precise analysis of their techniques. The findings of this study contribute to a deeper understanding of transformer-based methodologies and offer insights for further enhancing performance in medical image segmentation. We present this article in accordance with the Narrative Review reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-23-542/rc).

Methods

ViT (26) is the first vision transformer and has been the basis for further research in transformers in vision since 2021. Many vision transformers have been designed to explore their potential in medical image segmentation, and a survey of research papers from 2021 to 2023 was conducted to investigate the use of transformers in this area. The survey used various search engines such as Google Scholar, arxiv, ResearchGate, Microsoft Academic, and Semantic Scholar along with the following keywords: medical, transformers, segmentation, imaging and vision to find relevant papers. The relevant details related to paper acquisition and research are mentioned in *Table 1*.

Workings of a transformer

In this part of the section, we describe in detail the

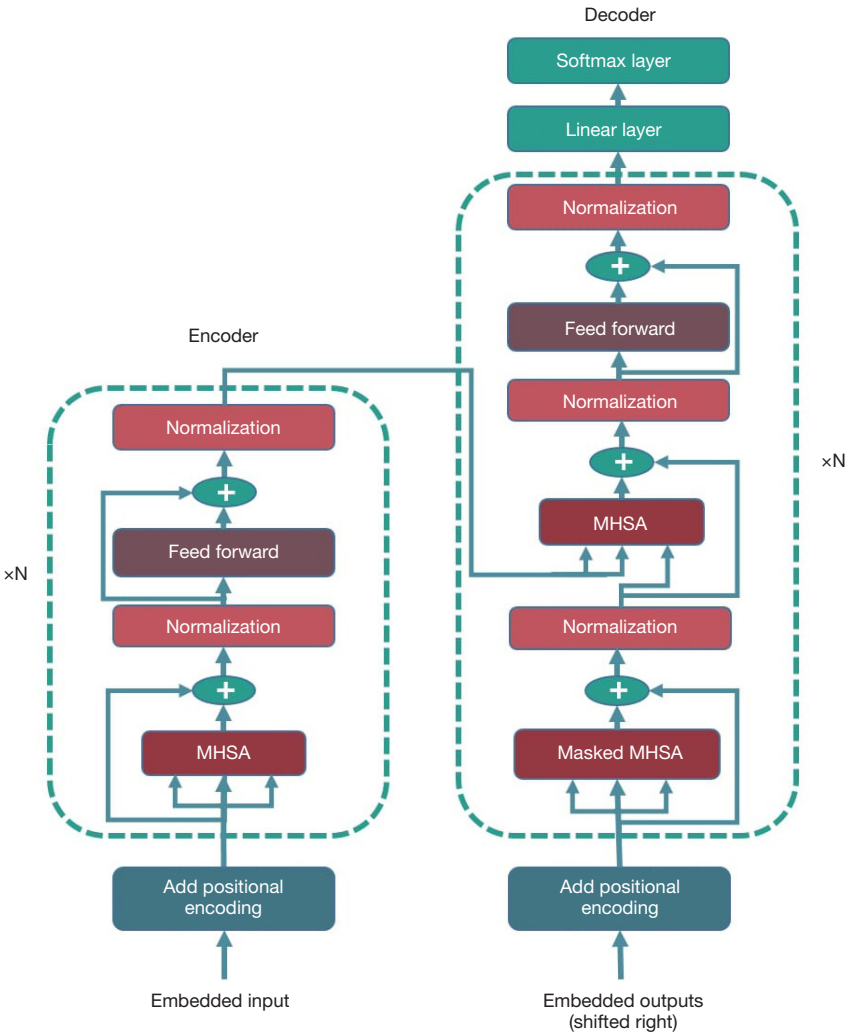

Figure 1 Structure of a transformer. MHSA, multi-head self-attention.

workings of the transformer (5). It comprises an encoder and a decoder, as can be seen in *Figure 1*. Before the data can reach the encoder or decoder, flattened tokens are generated, and their position information is added to keep the model spatially aware.

Encoder

The encoder has multiple stacks of layers, where each layer is composed of two sub-layers, one of which is the multi-head self-attention (MHSA) layer (see *Figure 2A*) and the other is a fully connected feed forward network. The transformer also utilizes residual connections after each sub-layer, thereby providing an alternate path for data to travel, while skipping a few layers. To this end, the dimensions of the input and output to the sub-layers are kept the same so element wise addition can be performed at the residual block. Each sub-layer is succeeded by a normalization layer.

Decoder

The decoder employs similar stacks of sub-layers to the encoder; however, it has an extra MHSA layer that performs self-attention over the result of the encoder block. The decoder also performs a masking technique in the first sub-layer to prevent positions from attending to subsequent positions. The goal here is to prevent predictions for position i to depend only on predictions for positions that are less than i. This technique is very useful in NLP where the prediction of a word should depend solely on its history.

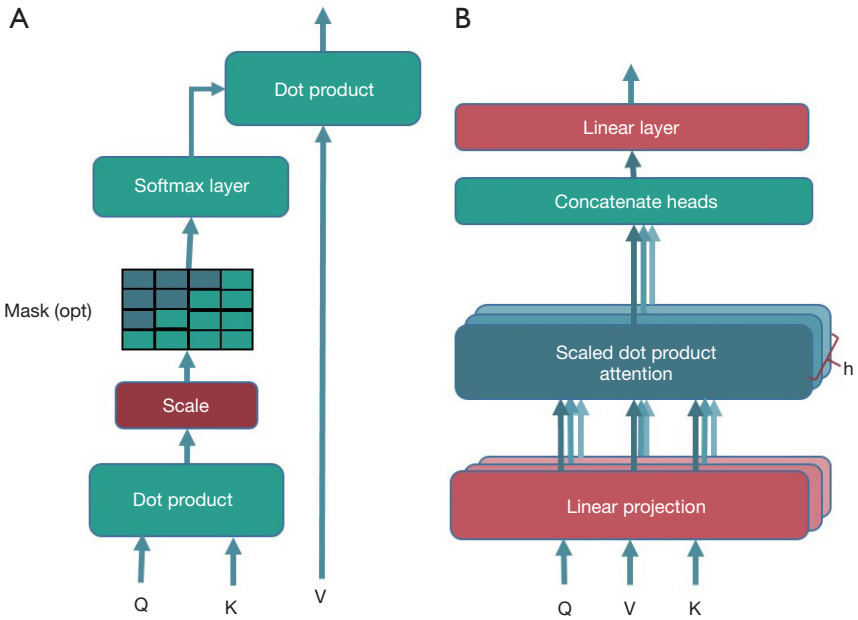

Figure 2 Attention mechanism of a transformer. (A) Scaled dot product. (B) MHSA layer of a transformer. MHSA, multi-head self-attention.

Scaled dot product attention

The attention mechanism of the transformer works by mapping a set of queries and key-value pairs of vectors into an output vector. To compute the output, a weighted sum of the value vector is used wherein the weights are calculated from the compatibility function between the query and its corresponding key.

This 'scaled dot product' attention is computed simultaneously on a set of projected queries, keys, and values that make up the matrices Q, K, and V with their respective vector dimensions being d_q, d_k, and d_v. These vectors have been projected into their respective dimensions from the positionally-embedded token vector of length d_{model}. To compute the query and key weights, a dot product of the query and key vector sets is calculated and divided by d_k, the square root of the dimension of key vector, d_k. Next, the softmax of the resulting matrix is computed. Once we have the query-key weights, their dot product with the set of value vectors provides us with the set of output vector or attention vector. This process is depicted in *Figure 2A*, and the mathematical equation of the complete attention mechanism is as follows:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad [1]$$

MHSA

In the MHSA layer, as can be seen in *Figure 2B*, instead of a single projection, the query, key, and value are projected H number of times with different, learned linear projections. The scaled dot product attention is then computed in parallel for all the heads, with the output vectors being concatenated. The concatenated vectors are then projected into final values. In this manner, the transformer jointly attends to the feature representation from different sub-spaces for all positions. MHSA can be expressed as follows:

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h) \quad [2]$$

$$head_i = Attention(Q_i, K_i, V_i) \quad [3]$$

where Q_i, K_i, and V_i are the query, key, and value matrices, respectively, of the i^{th} subspace or head.

Significance of MHSA

MHSAs have both advantages and disadvantages when it comes to generalizability. On one hand, they flatten loss landscapes, leading to improved performance and generalization. This is because flatter loss landscapes enhance accuracy and robustness, especially in scenarios with large amounts of data (22). On the other hand, MHSAs can have

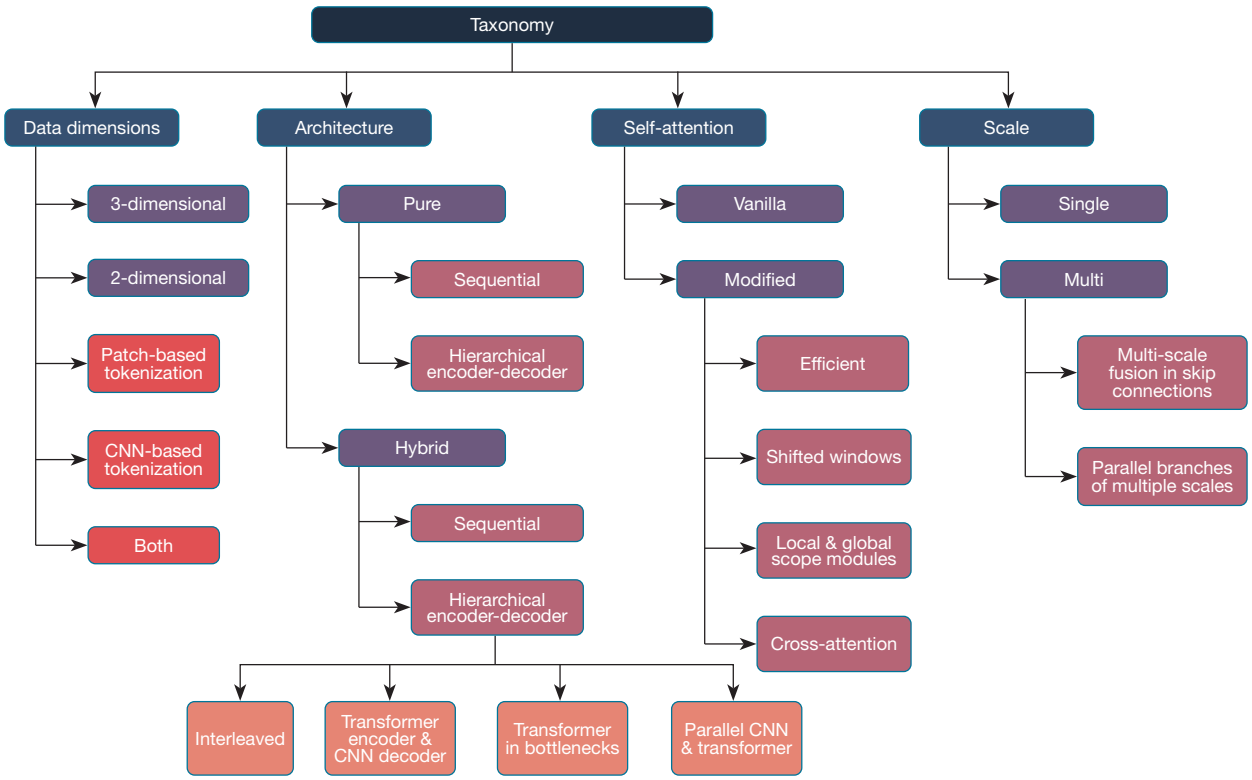

Figure 3 Main categories for transformer-network classification. CNN, convolutional neural network.

negative Hessian eigenvalues in situations with small amounts of data. This non-convexity of loss landscapes can disrupt neural network optimization. However, the presence of a substantial training dataset suppresses negative eigenvalues and makes the losses more convex (27).

MHSAs and convolutions exhibit contrasting behaviors in the sense that while MHSAs aggregates feature maps, convolutions diversify them. Additionally, Fourier analysis of feature maps reveals that MHSAs reduce high-frequency signals, whereas convolutions amplify high-frequency components. Essentially, MHSAs act as low-pass filters, while convolutions function as high-pass filters. Understandably, convolutions are susceptible to high-frequency noise, unlike MSHAs, which are more robust (23).

A survey of transformer-based medical image segmentation

Key contents & findings

In this paper, we survey diverse transformer models and strategies tailored for medical data. We examine structural modifications, self-attention enhancements, and multi-scale correlations in place to boost performance and generalizability. This is followed by conducting a quantitative analysis on select networks mentioned in our survey. Specifically, we focus on networks that were evaluated using benchmark datasets, which are commonly employed for the evaluation of medical image segmentation. Lastly, we conclude with a performance comparison of transformer-based networks with other networks for a number of different tasks to show how the transformer outperforms state of the art methodologies.

As shown in *Figure 3*, we introduce a taxonomy to categorize different transformer models based on medical data types, architectural designs, attention strategies and the presence of multi-scale correlations. Key characteristics of individual models were described based on this taxonomy. The papers included in this survey and their respective designs are listed in *Table 2*.

Dimensions

3D

In this category, we focus on transformer-based networks

Table 2 Main features of the transformer networks

Transformer network	Dimension	Modality	Segmentation tasks	Architecture	Patch representation	Use of vanilla self-attention	Multi-scale correlations
Convolution-free transformer (13)	3D	CT & MRI	Pancreas, hippocampus & brain cortical plate	Pure, sequential	Yes	Yes	No
UNETR (28)	3D	CT & MRI	Multi-organ & brain tumor	Hybrid, hierarchical transformer encoder & convolution decoder	Yes	Yes	No
TransBTS (29)	3D	MRI	Brain tumor	Hybrid, hierarchical encoder-decoder, transformer in bottleneck	No	Yes	No
Medical transformer (30)	3D	MRI	Brain tumor	Hybrid, sequential encoder-decoder	No	Yes	No
D-Former (31)	3D	CT & MRI	Multi-organ & cardiac chamber	Pure, hierarchical encoder-decoder	Yes	No	No
CoTr (32)	3D	CT	Multi-organ	Hybrid, hierarchical encoder-decoder	No	No	Yes
nnFormer (33)	3D	CT	Multi-organ & cardiac chamber	Hybrid, hierarchical encoder-decoder, Interleaved	Yes	No	No
TransAttUnet (34)	2D	CT & X-ray	Gland, nuclei & lesion	Hybrid, hierarchical encoder-decoder, transformer in bottleneck	No	Yes	Yes
TransUNet (35)	2D	CT & MRI	Multi-organ & cardiac chamber	Hybrid, hierarchical encoder-decoder, transformer in bottleneck	Yes	Yes	No
TransFuse (36)	2D	Colonoscopy	Polyp, lesion & hip	Hybrid, hierarchical encoder-decoder, parallel convolution and transformer-based encoders	Yes	Yes	No
SwinUNet (37)	2D	CT	Multi-organ & cardiac chamber	Pure, hierarchical encoder-decoder	Yes	No	No
DS-TransUNet (38)	2D	Colonoscopy	Polyp, gland, nuclei & lesion	Pure, hierarchical encoder-decoder separate local and global encoders	Yes	No	Yes
MissFormer (39)	2D	CT	Multi-organ & cardiac chamber	Pure, hierarchical encoder-decoder, interleaved	Yes	No	Yes
Mixed transformer (40)	2D	CT	Multi-organ & cardiac chamber	Hybrid, hierarchical encoder-decoder, transformer in bottleneck	Yes	No	No
UTNet (41)	2D	CT	Cardiac chamber	Hybrid, hierarchical encoder-decoder, interleaved	No	No	No
TUNet (42)	2D	CT	Pancreas	Hybrid, hierarchical encoder-decoder, parallel convolution and transformer-based encoders	Yes	Yes	No
TransBTSV2 (43)	3D	CT & MRI	Brain, liver tumors & kidney tumors	Hybrid, hierarchical encoder-decoder, transformer in bottleneck	No	No	No
Swin UNETR (44)	3D	MRI	Brain tumor	Hybrid, hierarchical transformer encoder & convolution decoder	Yes	No	No
VT-UNet (45)	3D	CT	Brain tumor	Hybrid, hierarchical encoder-decoder	Yes	No	No
LeVit-UNet (46)	2D	CT	Multi-organ & cardiac chamber	Hybrid, hierarchical encoder-decoder, transformer in bottleneck	No	Yes	No
TransClaw-UNet (47)	2D	CT	Multi-organ	Hybrid, hierarchical encoder-decoder, transformer in bottleneck	Yes	Yes	No
Segtran (48)	2D & 3D	MRI, fundus & colonoscopy	Brain tumor & polyp	Hybrid, sequential	No	No	No
GT-UNet (49)	2D	X-ray & fundus	Tooth root & retinal vessel	Hybrid, hierarchical encoder-decoder, interleaved	Yes	Yes	No

Table 2 *(continued)*

Table 2 *(continued)*

Transformer network	Dimension	Modality	Segmentation tasks	Architecture	Patch representation	Use of vanilla self-attention	Multi-scale correlations
Pyramid medical transformer (50)	2D	Microscopy	Gland, head & neck tumors, & nuclei	Hybrid, hierarchical encoder-decoder, parallel convolution and transformer-based encoders	No	No	No
Medformer (51)	2D & 3D	CT & MRI	Multi-organ, cardiac chamber, liver tumors & kidney tumors	Hybrid, hierarchical encoder-decoder	Yes	No	Yes
HybridCTrm (52)	3D	MRI	Brain tissue	Hybrid, sequential, parallel convolution & transformer-based encoders	Yes	Yes	No
HyLT (53)	2D	Histology & microscopy	Gland & nuclei	Hybrid, hierarchical encoder-decoder, interleaved	Yes	No	No
PHTrans (54)	3D	MRI	Multi-organ & cardiac chamber	Hybrid, hierarchical encoder-decoder, transformer in bottleneck	No	No	No

3D, three-dimensional; CT, computed tomography; MRI, magnetic resonance imaging; 2D, two-dimensional.

designed for 3D segmentation of medical images obtained from modalities like MRI, CT, and ultrasound (US). These networks enable direct anatomical segmentation without the need for slice-by-slice two-dimensional (2D) segmentation. Due to the impracticality of flattening and encoding 3D volumes in terms of computational and memory resources, network redesign is necessary to handle volumetric medical scans efficiently.

One way to deal with volumetric images is to divide them into smaller 3D volumes and encode features for each volume through a transformer, as performed by convolution-free transformer (13). Another more common approach is to treat the smaller 3D volumes as individual single tokens, which can then be embedded and fed to a transformer. Doing this significantly reduces the spatial dimension depending on the number of volumes, thus enabling prediction at very low computational costs. The global scope of the transformer is also maintained, since self-attention can then be applied to all the tokens combined. This is the approach utilized by various networks (28,31,33,44,45,51,52) and the generic architecture can be seen in *Figure 4A*.

3D computations can also be reduced significantly by using convolution layers in the early stages of the network to transform the volume into smaller feature maps. The transformer architecture can then be applied to these feature maps without the need to reduce spatial dimensions, as can be seen in (29,43,48,54). This process is depicted in *Figure 4B*. Medical transformer (30) introduced another method wherein the 3D volume was divided into 2D slices of three different views: coronal, sagittal, and axial. Then, each view slice was regarded as a separate image on which the transformer self-attention (TSA) was applied.

2D

In medical imaging, 2D images encompass various modalities such as X-rays and colonoscopy images, as well as individual slices extracted from 3D scans like CT and MRI. When employing models designed for 2D medical images in 3D segmentation tasks, the training typically occurs on individual 2D slices, resulting in a loss of volumetric information and patterns inherent in the 3D scan.

As is the case for 3D volumetric images, tokens can be generated from 2D images by first breaking the image or feature map into smaller patches, where each patch is converted into a single token (see *Figure 4A*). Patch embedding like this has been performed in a number of networks like (36-39,42,49). Similar to the 3D networks, another technique to reduce the spatial dimensions of the input is to begin by employing convolution layers to transform the image into lower dimension feature maps followed by transformer layers (see *Figure 4B*). Networks such as (34,41,46,48) follow this method to enable the computationally efficient modelling of transformer-based networks on medical images.

Another technique that many networks follow to further reduce computations is to use convolution in the early levels of the network as well as patch embedding immediately before transformer blocks. This method aims to make the network more optimized by substantially downsizing

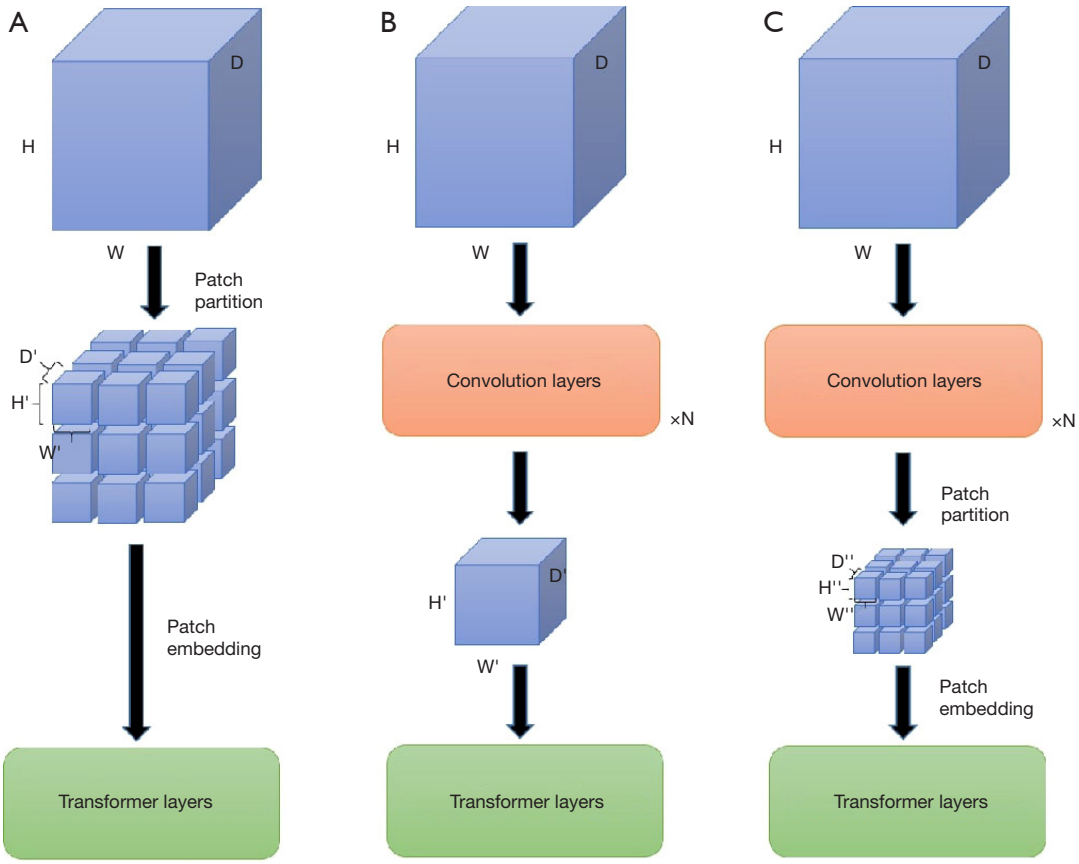

Figure 4 Popular feature sub-space reduction techniques for 3D and 2D medical data. (A) Patch partitioning (B) Convolution layers to spatially reduce the feature dimensions. (C) Convolution layers and patch partition to significantly minimize the feature subspace. 3D, three-dimensional; 2D, two-dimensional.

the tokens before performing self-attention through transformer blocks. This technique was performed in (35,40,47,50,51,53) and the general process is portrayed in *Figure 4C*.

Architecture

Pure transformer architecture

Pure transformer architectures exclusively rely on transformers, without incorporating CNNs. These models employ tokenization and self-attention to learn features and capture long-range dependencies. They stack transformer blocks or MHSA modules in serial or encoder-decoder configurations to establish relationships between different embedding spaces.

Patch-wise architectures are commonly used in pure transformer models, where image patches replace pixels as units of information. Self-attention is applied to explore the connections among embedded information. Examples of pure transformer networks, such as (31,37-39,45) adopt a U-net-like (55) encoder-decoder architecture (*Figure 5A*). Transformers learn local and global dependencies through feature representation and down-sampling, followed by up-sampling for pixel-wise prediction. MissFormer (39) further utilizes transformers in skip connections between encoder and decoder levels to generate meaningful multi-scale features. An alternative approach, although less common, is to construct a sequential transformer architecture as demonstrated in convolution-free transformer (13). In this setup, the incoming image is divided into patches and processed by stacked transformers for segmentation.

Hybrid transformer

As is suggested by the name, these are models that join transformers with CNNs in a hybrid architecture to attain the key features of both, the local information from

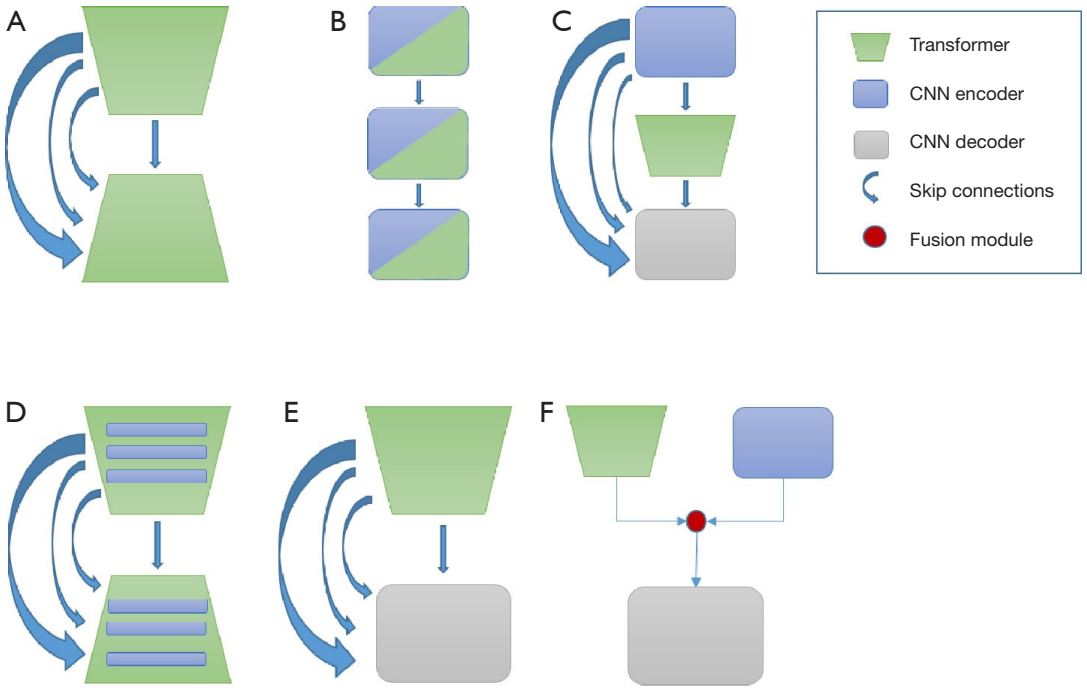

Figure 5 Popular hierarchical encoder-decoder techniques involving transformers. (A) Transformer encoder-decoder. (B) Sequential network. (C) CNN encoder-decoder with transformer in bottleneck. (D) Interleaved CNN and transformer blocks within the encoder and decoder. (E) Transformer encoder with CNN decoder. (F) Parallel branches of CNN encoder and transformer encoder followed by a fusion module before the CNN decoder. CNN, convolutional neural network.

CNNs and the global information from transformers. The transformer and convolutions layers are placed strategically to attain the most out of the network.

Consider medical transformer (30), a hybrid architecture utilized for the 3D segmentation task among sequential networks (depicted in *Figure 5B*). It incorporates parallel encoder branches, each acquiring 2D slices from distinct views (axial, coronal, and sagittal). The sequential network commences with a convolution encoder followed by a transformer encoder for each branch. Subsequently, the features are fused according to each view and then processed by the prediction network.

Another sequential network, HybridCTrm (52), focuses on multi-modal image segmentation and proposes two architectures. The single-path strategy combines both modalities into a multi-channel image, which is then split into separate branches: a CNN-based encoder and a transformer-based encoder. The features are fused and decoded using convolution to generate the segmentation mask. Alternatively, the multi-path strategy employs parallel CNN and transformer encoders for each modality. The features from each path are then integrated for subsequent decoding.

In medical-specific networks, the prevalent technique involves employing a U-net-like encoder-decoder architecture for feature extraction at multiple scales. However, instead of utilizing convolution layers at every level, these hybrid architectures strategically position convolution blocks early on to extract high-level local features, while transformer blocks are placed deeper or at the bottleneck to extract more global features, effectively reducing the size of the feature map. This process is illustrated in *Figure 5C*. Several segmentation networks, such as (29,32,34,35,40,43,46-48,54) employ this approach. Notably, CoTr (32) incorporates skip connections from the encoder to the transformer bottleneck, enabling multi-level feature extraction.

Another intelligent approach, depicted in *Figure 5D*, involves interleaving convolution and transformer blocks at each level of the U-net-like encoder and decoder. Networks like (33,41,49,53) adopt this method to extract both local and global features using transformers and convolutions at each scale throughout the encoder and decoder. This enables the model to learn spatial and temporal features

from each resolution in the network. Additionally, some methods solely employ transformers in the encoder levels, while the decoder levels exclusively consist of convolutions. *Figure 5E* presents these strategies, assuming that the transformer in the encoder can sufficiently extract both low and high-level features, with convolutions reserved for the prediction network. Networks such as UNETR (28) and Swin UNETR (44) follow this strategy.

Among the unconventional network approaches, TransFuse (36) and TUnet (42) stand out. In both cases, the image is processed by parallel branches consisting of a convolution-based encoder and a transformer-based encoder, while a single convolution-based decoder predicts the segmentation mask. However, there is a distinction between the two in the manner that in TransFuse, features of the same scale from both encoder branches are fused together, whereas this fusion does not occur in TUnet. Additionally, TUnet employs skip connections from the CNN-based encoder to the decoder, while the parallel transformer branch provides global features to the decoder input. The general architecture of such networks is depicted in *Figure 5F*.

Another network, pyramid medical transformer (50), deviates from the standard architecture which incorporates a branch comprising a simple convolution-based encoder, along with three hybrid branches consisting of convolution and transformer-based encoders. Each hybrid branch focuses on different dependency ranges (short, medium, and long), with the input image resized to various dimensions corresponding to large, medium, and small-scale image sizes for the respective branches. The features of different dimensions are fused with the pure convolution-based encoder, considering the feature map size, and are then passed as skip connections to the decoder.

Medformer (51) introduces a network that directly divides the image into large patches as tokens. These local image features undergo embedding into a token map using multiple convolution and down-sampling layers to reduce spatial size before being fed into the transformer for global feature extraction. Similar to the previous networks, it adopts a U-net-like architecture with transformer blocks at each level to extract meaningful information.

Attention

Vanilla self-attention

This refers to the attention style of the original transformer architecture, where a set of tokens is projected in a linear manner into query, key, and value matrices. The query and the set of key-value pairs are mapped to an output as detailed in the "Scaled dot product attention" section, where it is stated that all the queries, keys, and values come from the same sources. In this manner, self-attention is conducted while utilizing information from a single space. Medical segmentation networks like (13,28-30,35,36,42,46,47,49,52) all employ vanilla self-attention to extract global information.

Modified self-attention

In this category, the transformer's self-attention mechanism is modified to enhance specific aspects: (I) efficiency: computation is optimized to reduce complexity, making the architecture more efficient for training and inference. (II) Shifted mechanisms: self-attention is computed on separate and overlapping windows to expand the receptive field of transformer blocks. (III) Local and global scope modules: high and low-level features are captured separately to retain maximum information and improve generalization. (IV) Cross-attention: self-attention involves queries, keys, and value matrices from different subspaces, enhancing the model's robustness. (V) Other approaches: novel strategies for self-attention that don't fit into specific categories, offering unique ways to improve performance.

Efficiency

Efficiency-focused networks include (32,39,41,48). In CoTr (32) and UTNet (41), the token space is sub-sampled to reduce the vector space, with UTNet sub-sampling only key and value vectors. Conversely, MissFormer (39) reduces the spatial dimension before projecting query, key, and value vectors, aiding multi-scale feature handling with transformers in skip connections. The network Segtran (48) introduces efficiency via squeezed attention blocks (SABs), performing a more efficient NxM computation (M << N) instead of NxN, achieved by learning an M-dimensional embedding notebook to 'squeeze' the attention matrix, reducing time and complexity.

Shifted mechanism

These networks, inspired by the Swin transformer (56), divide the feature space into local windows of fixed patch count. Within each transformer block, self-attention is computed once within these windows and once after shifting them vertically and horizontally with a fixed stride. This captures both local and global dependencies by correlating neighboring windows. Networks such as (33,37,38,44,45,54) incorporate these shifted window mechanisms in each transformer block.

Local and global scope modules

These networks go beyond the traditional hierarchical encoder-decoder architecture by incorporating separate local and global scope modules within the self-attention module of the transformer block. In D-Former (31), attention is computed within local windows of neighbouring patches as part of the local scope module, and within new windows of patches at a fixed distance as the global scope module. This approach captures both local and global dependencies at each hierarchical level. In mixed transformer (40), the local scope module computes self-attention within local windows, generating a single token for each window. The global self-attention module performs axial self-attention with a learned Gaussian mask on these tokens to capture global dependencies. This reduces computational cost by effectively enhancing the perception of each query's neighborhood. Furthermore, the network also incorporates external attention (57) to learn correlations between different data samples. Memory units are used for key and value matrices, capturing essential information from the entire dataset to calculate attention scores for each sample.

Cross-attention

VT-Unet (45) incorporates shifted self-attention (SA) and cross SA. In the decoder transformer blocks of the hierarchical U-net-like architecture, SA is computed in parallel pairs. Each pair consists of a simple SA module and one with shifted windows. One pair receives key and value projections from deeper decoder levels, while the other pair receives projections from the encoder level of the same resolution. Value fusion between the SA types allows joint attention to spatial context from the encoder and semantic information from the decoder. HyLT (53) also utilizes cross SA in a hybrid network with a CNN encoder path and a transformer encoder path. Self-attention includes a vanilla MHSA block sandwiched between two cross MHSA blocks in each encoder level. The cross MHSA blocks merge local and global information from both paths.

Others

In TransAttUnet (34), the hybrid network employs TSA and global spatial attention (GSA) in the bottleneck. TSA is a simple MHSA module, while GSA combines convolutions, matrix multiplication, and softmax to generate a position attention map, enhancing global feature aggregation in a selective manner.

Another unique approach is seen in Medformer (51), which utilizes Bi-directional transformer blocks in both encoding and decoding paths. This reduces computational complexity from quadratic to linear while enabling multi-scale fusion of features. By projecting a concise semantic token map, each level of the hierarchical encoder-decoder is summarized holistically, reducing redundancy.

Scale

Single-scale transformer

Another interesting feature to look for in transformer-based architectures is the presence or absence of multi-scale correlation. Networks that have no multi-scale correlations extract meaningful information from features of the same resolution. These methods include networks like (13,28-31,33,35-37,40-50,52-54) where a multi-scale interrelationship is not given precedent to extract local and global context.

Multi-scale correlations

These networks enhance performance by incorporating multi-scale feature correlation. They leverage different strategies to efficiently integrate datasets of varying resolutions, improving segmentation models (58). For instance, MissFormer (39), CoTr (32) and Medformer (51) utilize multi-scale transformer-based fusion modules in skip connections. MissFormer and CoTr flatten and concatenate feature maps from each encoder level as embedded tokens, processed by multiple transformer blocks for multi-scale correlations. In contrast, Medformer employs semantic maps for fusion to prevent redundant tokens. There is also DS-TransUNet (38), a pure transformer network with hierarchical encoder branches at different scales. The transformer interactive fusion (TIF) module in this network performs cross-attention by generating tokens based on one branch's feature map and computing self-attention with the reshaped token sequence from the other branch. Other than this, TransAttUNet (34) utilizes a bi-linear up-sampling module for multi-scale skip connections, incorporating cascade connection, residual connection, and dense connection techniques to aggregate features of varying semantic scales in the decoder.

Quantitative analysis

In this section we have conducted a thorough analysis of various networks, focusing on their performance in different tasks within the realms of CT and MRI modalities. Specifically, for CT, we have examined the multi-organ segmentation task by utilizing the widely recognized Synapse multi-organ segmentation (59) dataset, which

Table 3 Network performances on CT tasks reported in Dice score (%)

Networks	Average	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
TransUNet† (35)	77.48	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
TransClaw-Unet (47)	78.09	85.87	61.38	84.83	79.36	94.28	57.65	87.74	73.55
LeVit-Unet (46)	78.53	87.33	62.23	84.61	80.25	93.11	59.07	88.86	72.76
Mixed Transformer (40)	78.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
Swin-UNet† (37)	79.13	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
MISSFormer (39)	81.96	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
UNetR (28)	79.42	88.92	69.80	81.38	79.71	94.28	58.93	86.14	76.22
VT-UNet† (45)	79.02	88.54	70.07	84.43	87.14	94.79	71.47	91.12	82.16
MedFormer (51)	84.52	92.43	77.36	92.40	91.21	95.85	81.92	93.15	90.23
Swin UNETR (44)	85.78	92.78	76.55	85.25	89.12	96.91	77.22	88.70	79.72
CoTr (32)	86.33	92.10	81.47	85.33	86.41	96.87	80.20	92.21	76.08
nnFormer† (33)	87.40	92.04	71.09	87.64	87.34	96.53	82.49	92.91	89.17
PHTrans (54)	88.55	92.54	80.89	85.25	91.30	97.04	83.42	91.20	86.75
D-Former (31)	88.83	92.12	80.09	92.60	91.91	96.99	76.67	93.78	86.44

†, the model was pre-trained on ImageNet. Top section reports 2D networks with bottom section reporting 3D networks. CT, computed tomography; L, left; R, right; 2D, two-dimensional; 3D, three-dimensional.

offers valuable annotations for a range of organs. Moving on to MRI, we concentrated our evaluation on a single, highly popular task—the Automated Cardiac Diagnosis Challenge (ACDC) (60). This challenge provides annotated data for three critical regions: left ventricle (LV), right ventricle (RV), and left ventricular myocardium (Myo). By assessing the networks' performances on this task, we were able to gain insights into their efficacy in automated cardiac diagnosis.

The selection of these tasks and datasets was intentional, as they collectively offer diverse shape, size, and tissue characteristics, encompassing a wide range of real-world scenarios. This diversity allows us to effectively approximate the strengths and weaknesses of different models, enabling a comprehensive analysis of their capabilities.

Table 3 in our study presents the performance evaluation of various networks on the Synapse dataset challenge presenting results across eight crucial organs as well as the average performance. The evaluation metric used to assess the performance is the Dice similarity coefficient, presented as a percentage. The results presented in *Table 3* are sourced from the original papers (31,39,40,51,54), representing the highest reported scores achieved by each respective network on the specific task. Upon examination, we observe that D-Former (31) emerges as the top-performing network for the multi-organ segmentation task closely followed by PHTrans (51). These networks demonstrate superior performance according to the reported results, showcasing their effectiveness in the respective tasks.

Similarly, *Table 4* in our study presents the evaluation of various networks specifically designed for MRI in the context of the ACDC task. The evaluation metric used is, again, the Dice similarity coefficient, expressed as a percentage. The table provides an overview of the scores obtained by each network on four different segmentation areas: LV, RV, Myo, and the average Dice score across all three areas. Upon analyzing the results, it becomes evident that D-Former (31) exhibits the highest average Dice score, positioning it as the leading network. MedFormer (51) closely follows with competitive performance. However, in terms of Myo segmentation specifically, MedFormer outperforms other networks, claiming the lead.

These findings, based on the compiled scores (31,51,54) highlight the effectiveness of D-Former and MedFormer in the ACDC task, showcasing their superior performance across multiple segmentation areas. Such insights enable us to make informed assessments regarding the suitability and potential applications of these networks in the field of MRI-based medical imaging.

Table 4 Network performances on the MRI ACDC challenge in Dice score (%)

Networks	Average	RV	Myo	LV
UTNet (41)	88.30	90.41	89.15	94.39
TransUNet† (35)	89.71	88.86	84.54	95.73
Swin-UNet† (37)	90.00	88.55	85.62	95.83
LeVit-Unet (46)	90.32	89.55	87.64	93.76
Mixed Transformer (40)	90.43	86.64	89.04	95.62
MISSFormer (39)	90.86	89.55	88.04	94.99
UNETR (28)	87.15	84.52	84.36	92.57
VT-UNet† (45)	91.13	89.44	88.42	95.53
nnFormer† (33)	91.78	90.22	89.53	95.59
PHTrans (54)	91.79	90.13	89.58	95.76
MedFormer (51)	92.14	90.95	89.71	95.76
D-Former (31)	92.29	91.33	89.60	95.93

†, the model was pre-trained on ImageNet. MRI, magnetic resonance imaging; ACDC, Automated Cardiac Diagnosis Challenge; RV, right ventricle; Myo, left ventricular myocardium; LV, left ventricle.

Transformers vs. other networks: performance

In our investigation, we have also conducted a comprehensive performance comparison between transformer-based networks and other benchmark CNN networks in various medical segmentation tasks. This analysis has revealed the notable superiority of transformer-based models over the state-of-the-art networks across different modalities and problem domains. Specifically, we present the performance comparison in *Table 5*, focusing on three significant tasks: brain tumor segmentation (BraTS 2019 challenge) (72) in MRI, multi-organ segmentation (Synapse) (59) in CT, and gland segmentation (GlaS) (73) in Microscopy. These tasks encompass a wide range of target scale, form, and structure attributes, thereby providing a comprehensive evaluation of the networks' capabilities. The evaluation metric employed in this comparison is, again, Dice similarity score in percentage. The table presents the performance scores for each task based on the area of expertise of the networks involved.

Starting with the BraTS challenge, we observe from experiment reports (43) that transformer-based networks, such as TransUNet (35), SwinUNet (37), and TransBTSV2 (43), demonstrate performance scores comparable to or even surpassing those of the best CNN network performances. Notably, TransBTSV2 achieves the highest score, highlighting the effectiveness of transformer-based architectures in brain tumor segmentation. Moving to the Synapse dataset for multi-organ segmentation, reported by Zhou *et al.* (33), we find that nnFormer (33), a transformer-based network, outperforms state-of-the-art CNN networks by a significant margin. This performance superiority is consistent across different organs, indicating the robustness and versatility of transformer-based models in handling varied organ contours, proportions, and radiographic densities. Additionally, results on the GlaS dataset, reported from Chen *et al.* (34) and Luo *et al.* (53), show a remarkable improvement in performance when transformer-based networks like HyLT (53) are employed. This demonstrates the ability of transformer architectures to capture complex glandular structures and enhance segmentation accuracy.

In summary, through a comprehensive performance comparison on these diverse medical segmentation tasks, transformer-based networks consistently demonstrate their superiority over other state-of-the-art CNN networks. The robust performance of transformers in these tasks highlights their effectiveness in capturing intricate details, modelling complex relationships, and achieving better segmentation accuracy in medical image analysis.

Conclusions

Through this survey of transformers in medical image

Table 5 Performance in Dice score (%) of CNN-based and transformer-based networks across three different datasets

Network	BraTS 2019 (avg.)	Synapse (avg.)	GlaS
U-Net (55)	76.90	76.85	75.73
Att-Unet (61)	80.65	77.77	81.59
Tunet (62)	83.29	–	–
3D KiU-Net (63)	78.24	–	83.25
V-Net (64)	79.72	–	–
TransUNet (35)	82.18	–	–
SwinUNet (37)	82.20	79.13	86.70
TransBTS (29)	83.62	–	–
TransBTSV2 (43)	85.17	–	–
DualNorm-UNet (65)	–	80.37	–
ENet (66)	–	77.63	–
R50-DeepLabv3+ (67)	–	75.73	–
EDANet (68)	–	75.43	–
LeVit-UNet-384s (46)	–	78.53	–
nnFormer (33)	–	87.40	–
UNet++ (69)	–	–	81.83
ResUNet (70)	–	–	80.88
FANet (71)	–	–	84.67
TransAttUNet (34)	–	–	89.11
HyLt (53)	–	–	90.86

CNN, convolutional neural network; BraTS, brain tumor segmentation; avg., average; GlaS, gland segmentation.

segmentation applications, we have elucidated the different approaches that can be taken to efficiently boost the performance of deep learning in medical AI. We started by highlighting the different medical applications that can be ameliorated through image segmentation via AI. We then moved on to the state-of-the-art networks already in use in several application areas and their limitations, with CNNs representing the state of the art before transformers were introduced in the vision domain.

In this paper, we have explored the different aspects of transformer-based networks, ranging from network architecture designs to modified self-attention mechanisms. Additionally, we have conducted quantitative analyses on significant tasks to compare transformer-based networks with each other, as well as with other state-of-the-art networks that do not incorporate transformers. Our findings reveal that hybrid networks combining Transformers and CNNs, such as PHTrans (54), HyLt (53), TransBTSV2 (43), and MedFormer (51), demonstrate impressive results across most tasks. However, even pure transformer-based networks like D-Former (31) exhibit excellent performance in 3D tasks. Notably, our analysis highlights the significance of modified self-attention mechanisms as a key factor contributing to the enhanced performance of these networks across all examined tasks.

However, there are a number of important research directions for improving the performance of the transformer in the medical domain. While transformer-based approaches have shown promise, there is room for further enhancement. Transfer learning is a potential direction to explore, especially in the medical domain where annotated images are scarce for training accurate deep learning models from scratch. By utilizing pre-trained transformer-based models trained on large datasets from related domains, such

as ImageNet (74) and Microsoft COCO (75), or specifically medical imaging datasets like RadImageNet (76), NIH (77), and CheXpert (78), which include diverse modalities (CT, MRI, X-ray, etc.), the models can be fine-tuned on limited medical image datasets of interest. This approach has the potential to enhance the accuracy and efficiency of segmentation models.

Semi-supervised learning and self-supervised learning are utilized to address the scarcity of medical data in training transformer-based models. Semi-supervised learning combines labeled and unlabeled data, augmenting the labeled data to enhance accuracy and robustness of segmentation models. Promising results have been achieved by leveraging multiple segmentation tasks and their consistency (79). Another approach involves a combination of supervised and unsupervised losses to utilize unlabeled data, as demonstrated in cardiac MR image segmentation (80). Self-supervised learning trains models to solve pretext tasks, like restoring image context (81), or predicting anatomical position (82), leading to notable advancements in medical image segmentation. However, further investigation is required to comprehensively evaluate these techniques and their applicability to other medical imaging tasks.

Medical imaging technologies like CT and MRI produce scans with very high resolutions that can even reach up to three times the image resolution we now see in models for natural camera images. However, these are then processed and down-sampled greatly to fit the state-of-the-art model networks, which causes the loss of a lot of vital information. If these images could somehow be kept at their original resolutions, it would greatly improve current segmentation solutions.

Lastly, analogous to how domain-general large language models such as Bidirectional Encoder Representations from Transformers (BERT) (83) and ChatGPT (84) have provided a strong foundation for diverse successful domain-specific large langue models, an intriguing direction for future research in medical image segmentation involves leveraging foundation models such as segment anything model (SAM) (85). Specifically, there is a need for further exploration in areas including fine-tuning, embedding and in-context learning, among others, to effectively adapt these foundation models for general medical image segmentation while considering unique characteristics of medical images.

Strengths & limitations

This section outlines the limitations and strengths of our research survey on transformers in medical image segmentation.

Limitations:

(I) The survey primarily focuses on analyzing and comparing transformers specifically tailored for medical image segmentation, potentially overlooking the effectiveness of general transformer-based architectures in this domain.

(II) Despite efforts to include innovative and pioneering research, there remains a possibility of omission of certain medical segmentation transformer networks. Given the continuously evolving and demanding nature of research in transformers for medical image segmentation, it is important to acknowledge that new developments and advancements may continue to emerge beyond the scope of this survey.

(III) The quantitative analysis of transformer-based networks is cantered on widely known medical segmentation tasks and challenges, encompassing only a subset of medical data. The assessment relies on data reported in respective papers rather than conducting new experiments.

Strengths:

(I) The survey introduces a unique taxonomy for evaluating transformer-based networks, incorporating distinct assessment categories such as self-attention types and multi-scale strategies.

(II) The taxonomy facilitates coherent grouping of networks sharing similar characteristics, enhancing comprehension of underlying architectural principles.

(III) The survey delves into understanding architectures that deviate from prevalent approaches within the taxonomy, offering a comprehensive insight into various network traits.

(IV) A quantitative assessment of numerous transformer-based networks is provided across diverse medical segmentation datasets, covering a wide spectrum of segmentation tasks. These tasks encompass diverse modalities, prediction volumes, sizes, and contrast, ensuring a comprehensive performance evaluation of the networks.

Acknowledgments

Funding: This work was supported by the Kyonggi University's Graduate Research Assistantship 2023 and

the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1A2C1007169).

Footnote

Reporting Checklist: The authors have completed the Narrative Review reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-23-542/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-23-542/coif). The authors have no conflict of interest to share.

Ethical Statement: The authors are accountable for all aspects of the work ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Norouzi A, Rahim MS, Altameem A, Saba T, Rad AE, Rehman A, Uddin M. Medical image segmentation methods, algorithms, and applications. IETE Tech Rev 2014;31:199-213.
2. Pham DL, Xu C, Prince JL. Current methods in medical image segmentation. Annu Rev Biomed Eng 2000;2:315-37.
3. Kayalibay B, Jensen G, van der Smagt P. CNN-based segmentation of medical imaging data. arXiv:1701.03056 [Preprint]. 2017. Available online: https://arxiv.org/abs/1701.03056
4. Li F, Zhou L, Wang Y, Chen C, Yang S, Shan F, Liu L. Modeling long-range dependencies for weakly supervised disease classification and localization on chest X-ray. Quant Imaging Med Surg 2022;12:3364-78.
5. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017.
6. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555 [Preprint]. 2014. Available online: https://arxiv.org/abs/1412.3555
7. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-8.
8. Lu WT, Wang JC, Won M, Choi K, Song X. SpecTNT: A time-frequency transformer for music audio. arXiv:2110.09127 [Preprint]. 2021. Available online: https://arxiv.org/abs/2110.09127
9. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in vision: A survey. ACM Comput Surv 2022;54:1-41.
10. Dai Y, Gao Y, Liu F. TransMed: Transformers Advance Multi-Modal Medical Image Classification. Diagnostics (Basel) 2021;11:1384.
11. Gheflati B, Rivaz H. Vision Transformers for Classification of Breast Ultrasound Images. Annu Int Conf IEEE Eng Med Biol Soc 2022;2022:480-3.
12. Li M, Liu F, Zhu J, Zhang R, Qin Y, Gao D. Model-based clinical note entity recognition for rheumatoid arthritis using bidirectional encoder representation from transformers. Quant Imaging Med Surg 2022;12:184-95.
13. Karimi D, Vasylechko S D, Gholipour A. Convolution-free medical image segmentation using transformers. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part I 24. Cham: Springer International Publishing; 2021:78-88.
14. O'Shea K, Nash R. An introduction to convolutional neural networks. arXiv:1511.08458 [Preprint]. 2015. Available online: https://arxiv.org/abs/1511.08458
15. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao A, Xu C, Xu Y, Yang Z, Zhang Y, Tao D. A Survey on Vision Transformer. IEEE Trans Pattern Anal Mach Intell 2023;45:87-110.
16. Jiang H. Computed tomography: principles, design, artifacts, and recent advances. Bellingham: SPIE; 2009.
17. Vlaardingerbroek MT, Boer JA. Magnetic resonance imaging: theory and practice. 3rd ed. Berlin: Springer Science & Business Media; 2013.
18. Pasveer B. Knowledge of shadows: the introduction of X-ray images in medicine. Sociol Health Illn

1989;11:360-81.
19. Razzak MI, Naz S, Zaib A. Deep Learning for Medical Image Processing: Overview, Challenges and the Future. In: Dey N, Ashour A, Borra S. editors. Classification in BioApps. Lecture Notes in Computational Vision and Biomechanics. Cham: Springer; 2018:323-50.
20. Cardoso MJ, Arbel T, Lee SL, Cheplygina V, Balocco S, Mateus D, Zahnd G, Maier-Hein L, Demirci S, Granger E, Duong L. Intravascular Imaging and Computer Assisted Stenting, and large-scale annotation of biomedical data and expert label synthesis. In: CVII-STENT and Second International Workshop, LABELS. Cham: Springer; 2017.
21. Panayides AS, Amini A, Filipovic ND, Sharma A, Tsaftaris SA, Young A, Foran D, Do N, Golemati S, Kurc T, Huang K, Nikita KS, Veasey BP, Zervakis M, Saltz JH, Pattichis CS. AI in Medical Imaging Informatics: Current Challenges and Future Directions. IEEE J Biomed Health Inform 2020;24:1837-57.
22. Chen X, Hsieh CJ, Gong B. When vision transformers outperform resnets without pre-training or strong data augmentations. arXiv:2106.01548 [Preprint]. 2021. Available online: https://arxiv.org/abs/2106.01548
23. Park N, Kim S. How do vision transformers work? arXiv:2202.06709 [Preprint]. 2022. Available online: https://arxiv.org/abs/2106.01548
24. He K, Gan C, Li Z, Rekik I, Yin Z, Ji W, Gao Y, Wang Q, Zhang J, Shen D. Transformers in medical image analysis. Intelligent Medicine 2023;3:59-78.
25. Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, Fu H. Transformers in medical imaging: A survey. Med Image Anal 2023;88:102802.
26. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 [Preprint]. 2020. Available online: https://arxiv.org/abs/2010.11929
27. Dauphin YN, Pascanu R, Gulcehre C, Cho K, Ganguli S, Bengio Y. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In: Advances in Neural Information Processing Systems 27 (NIPS 2014). 2014.
28. Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, Roth HR, Xu D. Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022:574-84.
29. Wang W, Chen C, Ding M, Yu H, Zha S, Li J. TransBTS: Multimodal Brain Tumor Segmentation Using Transformer. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part I. Cham: Springer; 2021;24:109-19.
30. Jun E, Jeong S, Heo DW, Suk HI. Medical transformer: Universal brain encoder for 3D MRI analysis. arXiv:2104.13633 [Preprint]. 2021. Available online: https://arxiv.org/abs/2104.13633
31. Wu Y, Liao K, Chen J, Wang J, Chen DZ, Gao H, Wu J. D-former: A u-shaped dilated transformer for 3d medical image segmentation. Neural Comput Appl 2023;35:1931-44.
32. Xie Y, Zhang J, Shen C, Xia Y. CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part III. Cham: Springer; 2021;24:171-80.
33. Zhou HY, Guo J, Zhang Y, Yu L, Wang L, Yu Y. nnformer: Interleaved transformer for volumetric segmentation. arXiv:2109.03201 [Preprint]. 2021. Available online: https://arxiv.org/abs/2109.03201
34. Chen B, Liu Y, Zhang Z, Lu G, Kong AW. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. IEEE Trans Emerg Top Comput Intell 2023. doi: 10.48550/arXiv.2107.05274.
35. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. Transunet: Transformers make strong encoders for medical image segmentation. arXiv:2102.04306 [Preprint]. 2021. Available online: https://arxiv.org/abs/2102.04306
36. Zhang Y, Liu H, Hu Q. Transfuse: Fusing transformers and cnns for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part I 24. Cham: Springer; 2021:14-24.
37. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M. Swin-unet: Unet-like pure transformer for medical image segmentation. In: Karlinsky L, Michaeli T, Nishino K. editors. European conference on computer vision. Cham: Springer Nature Switzerland; 2022:205-18.
38. Lin A, Chen B, Xu J, Zhang Z, Lu G, Zhang D. Ds-transunet: Dual swin transformer u-net for medical image segmentation. IEEE Trans Instrum Meas 2022;71:1-15.

39. Huang X, Deng Z, Li D, Yuan X. Missformer: An effective medical image segmentation transformer. arXiv:2109.07162 [Preprint]. 2021. Available online: https://arxiv.org/abs/2109.07162
40. Wang H, Xie S, Lin L, Iwamoto Y, Han XH, Chen YW, Tong R. Mixed Transformer U-Net for Medical Image Segmentation. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE; 2022:2390-4.
41. Gao Y, Zhou M, Metaxas DN. UTNet: a hybrid transformer architecture for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part III. Cham: Springer; 2021;24:61-71.
42. Sha Y, Zhang Y, Ji X, Hu L. Transformer-unet: Raw image processing with unet. arXiv:2109.08417 [Preprint]. 2021. Available online: https://arxiv.org/abs/2109.08417
43. Li J, Wang W, Chen C, Zhang T, Zha S, Yu H, Wang J. Transbtsv2: Wider instead of deeper transformer for medical image segmentation. arXiv:2201.12785 [Preprint]. 2022. Available online: https://arxiv.org/abs/2201.12785
44. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin UNetr: Swin transformers for semantic segmentation of brain tumors in MRI images. In: International MICCAI Brainlesion Workshop. Cham: Springer International Publishing; 2021:272-84.
45. Peiris H, Hayat M, Chen Z, Egan G, Harandi M. A robust volumetric transformer for accurate 3D tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland; 2022:162-72.
46. Xu G, Wu X, Zhang X, He X. Levit-unet: Make faster encoders with transformer for medical image segmentation. arXiv:2107.08623 [Preprint]. 2021. Available online: https://arxiv.org/abs/2107.08623
47. Yao C, Hu M, Zhai G, Zhang XP. Transclaw u-net: Claw u-net with transformers for medical image segmentation. arXiv:2107.05188 [Preprint]. 2021. Available online: https://arxiv.org/abs/2107.05188
48. Li S, Sui X, Luo X, Xu X, Liu Y, Goh R. Medical image segmentation using squeeze-and-expansion transformers. arXiv:2105.09511 [Preprint]. 2021. Available online: https://arxiv.org/abs/2105.09511
49. Li Y, Wang S, Wang J, Zeng G, Liu W, Zhang Q, Jin Q, Wang Y. Gt u-net: A U-Net like group transformer network for tooth root segmentation. In: Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12. Cham: Springer International Publishing; 2021:386-95.
50. Zhang Z, Zhang W. Pyramid medical transformer for medical image segmentation. arXiv:2104.14702 [Preprint]. 2021. Available online: https://arxiv.org/abs/2104.14702
51. Gao Y, Zhou M, Liu D, Yan Z, Zhang S, Metaxas DN. A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark. arXiv:2203.00131 [Preprint]. 2022. https://arxiv.org/abs/2203.00131
52. Sun Q, Fang N, Liu Z, Zhao L, Wen Y, Lin H. HybridCTrm: Bridging CNN and Transformer for Multimodal Brain Image Segmentation. J Healthc Eng 2021;2021:7467261.
53. Luo H, Changdong Y, Selvan R. Hybrid ladder transformers with efficient parallel-cross attention for medical image segmentation. In: International Conference on Medical Imaging with Deep Learning. PMLR; 2022:808-19.
54. Liu W, Tian T, Xu W, Yang H, Pan X, Yan S, Wang L. Phtrans: Parallelly aggregating global and local representations for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland; 2022:235-44.
55. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Cham: Springer International Publishing; 2015:234-41.
56. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:10012-22.
57. Guo MH, Liu ZN, Mu TJ, Hu SM. Beyond Self-Attention: External Attention Using Two Linear Layers for Visual Tasks. IEEE Trans Pattern Anal Mach Intell 2023;45:5436-47.
58. Alber M, Buganza Tepole A, Cannon WR, De S, Dura-Bernal S, Garikipati K, Karniadakis G, Lytton WW, Perdikaris P, Petzold L, Kuhl E. Integrating machine learning and multiscale modeling-perspectives, challenges, and opportunities in the biological, biomedical, and

behavioral sciences. NPJ Digit Med 2019;2:115.
59. Landman B, Xu Z, Igelsias J, Styner M, Langerak T, Klein A. MICCAI multi-atlas labeling beyond the cranial vault-workshop and challenge. In: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. 2015;5:12.
60. Bernard O, Lalande A, Zotti C, Cervenansky F, Yang X, Heng PA, et al. Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? IEEE Trans Med Imaging 2018;37:2514-25.
61. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, Glocker B. Attention u-net: Learning where to look for the pancreas. arXiv:1804.03999 [Preprint]. 2018. Available online: https://arxiv.org/abs/1804.03999
62. Vu MH, Nyholm T, Löfstedt T. TuNet: End-to-end hierarchical brain tumor segmentation using cascaded networks. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I 5. Cham: Springer International Publishing; 2020:174-86.
63. Valanarasu JMJ, Sindagi VA, Hacihaliloglu I, Patel VM. KiU-Net: Overcomplete Convolutional Architectures for Biomedical Image and Volumetric Segmentation. IEEE Trans Med Imaging 2022;41:965-76.
64. Milletari F, Navab N, Ahmadi SA. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). IEEE; 2016:565-71.
65. Xiao J, Yu L, Xing L, Yuille A, Zhou Y. DualNorm-UNet: Incorporating global and local statistics for robust medical image segmentation. arXiv:2103.15858 [Preprint]. 2021. Available online: https://arxiv.org/abs/2103.15858
66. Paszke A, Chaurasia A, Kim S, Culurciello E. Enet: A deep neural network architecture for real-time semantic segmentation. arXiv:1606.02147 [Preprint]. 2016. Available online: https://arxiv.org/abs/1606.02147
67. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Trans Pattern Anal Mach Intell 2018;40:834-48.
68. Lo SY, Hang HM, Chan SW, Lin JJ. Efficient dense modules of asymmetric convolution for real-time semantic segmentation. In: Proceedings of the ACM multimedia Asia. 2019:1-6.
69. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. Deep Learn Med Image Anal Multimodal Learn Clin Decis Support (2018) 2018;11045:3-11.
70. Diakogiannis FI, Waldner F, Caccetta P, Wu C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS J Photogramm Remote Sens 2020;162:94-114.
71. Tomar NK, Jha D, Riegler MA, Johansen HD, Johansen D, Rittscher J, Halvorsen P, Ali S. FANet: A Feedback Attention Network for Improved Biomedical Image Segmentation. IEEE Trans Neural Netw Learn Syst 2023;34:9375-88.
72. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Trans Med Imaging 2015;34:1993-2024.
73. Sirinukunwattana K, Pluim JPW, Chen H, Qi X, Heng PA, Guo YB, Wang LY, Matuszewski BJ, Bruni E, Sanchez U, Böhm A, Ronneberger O, Cheikh BB, Racoceanu D, Kainz P, Pfeiffer M, Urschler M, Snead DRJ, Rajpoot NM. Gland segmentation in colon histology images: The glas challenge contest. Med Image Anal 2017;35:489-502.
74. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE conference on Computer Vision and Pattern Recognition. IEEE; 2009:248-55.
75. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft coco: Common objects in context. In: Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Cham: Springer International Publishing, 2014:740-55.
76. Mei X, Liu Z, Robson PM, Marinelli B, Huang M, Doshi A, Jacobi A, Cao C, Link KE, Yang T, Wang Y, Greenspan H, Deyer T, Fayad ZA, Yang Y. RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning. Radiol Artif Intell 2022;4:e210315.
77. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers R. Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: IEEE CVPR. 2017;7:46.
78. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskaya K, Seekins J. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence

2019;33:590-7.
79. Luo X, Chen J, Song T, Wang G. Semi-supervised medical image segmentation through dual-task consistency. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2021; 35:8801-8809.
80. Bai W, Oktay O, Sinclair M, Suzuki H, Rajchl M, Tarroni G, Glocker B, King A, Matthews PM, Rueckert D. Semi-supervised learning for network-based cardiac MR image segmentation. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20. Cham: Springer International Publishing; 2017:253-60.
81. Chen L, Bentley P, Mori K, Misawa K, Fujiwara M, Rueckert D. Self-supervised learning for medical image analysis using image context restoration. Med Image Anal 2019;58:101539.
82. Bai W, Chen C, Tarroni G, Duan J, Guitton F, Petersen SE, Guo Y, Matthews PM, Rueckert D. Self-supervised learning for cardiac MR image segmentation by anatomical position prediction. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part II 22. Cham: Springer International Publishing; 2019:541-9.
83. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 [Preprint]. 2018. Available online: https://arxiv.org/abs/1810.04805
84. Liu Y, Han T, Ma S, Zhang J, Yang Y, Tian J, He H, Li A, He M, Liu Z, Wu Z. Summary of ChatGPT/gpt-4 research and perspective towards the future of large language models. arXiv:2304.01852 [Preprint]. 2023. Available online: https://arxiv.org/abs/2304.01852
85. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo WY, Dollár P. Segment anything. arXiv:2304.02643 [Preprint]. 2023. Available online: https://arxiv.org/abs/2304.02643

Cite this article as: Khan RF, Lee BD, Lee MS. Transformers in medical image segmentation: a narrative review. Quant Imaging Med Surg 2023;13(12):8747-8767. doi: 10.21037/qims-23-542

