Attention Mechanisms in Medical Image Segmentation: A Survey

Yutong Xie[a,1], Bing Yang[b,1], Qingbiao Guan[b], Jianpeng Zhang[c], Qi Wu[a,*], Yong Xia[b,*]

[a] *University of Adelaide, Adelaide, Australia*
[b] *Northwestern Polytechnical University, Xi'an, China*
[c] *Alibaba DAMO Academy, China*

arXiv:2305.17937v1 [eess.IV] 29 May 2023

ARTICLE INFO

Article history:

Medical Image Segmentation; Attention Mechanism; Transformer; Deep Learning

ABSTRACT

Medical image segmentation plays an important role in computer-aided diagnosis. Attention mechanisms that distinguish important parts from irrelevant parts have been widely used in medical image segmentation tasks. This paper systematically reviews the basic principles of attention mechanisms and their applications in medical image segmentation. First, we review the basic concepts of attention mechanism and formulation. Second, we surveyed over 300 articles related to medical image segmentation, and divided them into two groups based on their attention mechanisms, non-Transformer attention and Transformer attention. In each group, we deeply analyze the attention mechanisms from three aspects based on the current literature work, *i.e.*, the principle of the mechanism (what to use), implementation methods (how to use), and application tasks (where to use). We also thoroughly analyzed the advantages and limitations of their applications to different tasks. Finally, we summarize the current state of research and shortcomings in the field, and discuss the potential challenges in the future, including task specificity, robustness, standard evaluation, *etc*. We hope that this review can showcase the overall research context of traditional and Transformer attention methods, provide a clear reference for subsequent research, and inspire more advanced attention research, not only in medical image segmentation, but also in other image analysis scenarios.

© 2023

1. Introduction

Medical image segmentation, as an important and difficult part of computer-aided diagnosis (CAD), has attracted much attention in recent studies. Its purpose is to differentiate anatomical or pathological structures in various medical images, such as computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), X-ray, ultrasound imaging (UI), and common RGB images like microscopy and fundus retinal images. Accurate segmentation of medical images is advantageous for diagnosis, treatment, and prognosis. The automation of this task, however, is extremely challenging due to three reasons: (1) the low soft tissue contrast results in fuzzy object boundaries; (2) anatomical or pathological structures may vary greatly in shape, size, and location; and (3) it is difficult to obtain sufficient annotated medical images to train segmentation models constrained by the labor cost and expertise. This makes it difficult to model the semantic relationship between different objects and backgrounds properly.

In the human visual cognition system, we are naturally skilled at focusing on the area of interest and ignoring the interference of other background information, which helps us to identify and judge more accurately and efficiently. Imitating this, attention mechanisms are proposed to adaptively assign weights to different regions in an image, enabling neural networks to focus on the important regions related to the target task

*Corresponding authors; yxia@nwpu.edu.cn; qi.wu01@adelaide.edu.au
[1]Co-first authors; yutong.xie678@gmail.com; yangbing@mail.nwpu.edu.cn

and disregard irrelevant areas, as shown in Figure 1. This powerful capability is well-suited for capturing complex semantic relationships in medical image segmentation. Furthermore, the attention mechanism can be utilized to explain the correlation between input and output data, illustrating what the model has learned, thus providing us with an interpretability insight into the black box of neural networks.

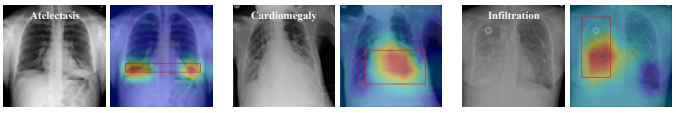

Fig. 1: Visualization of three chest X-ray samples with their attention maps obtained by Hu et al. (2018b). The red rectangles indicate the ground truth bounding box including disease regions.

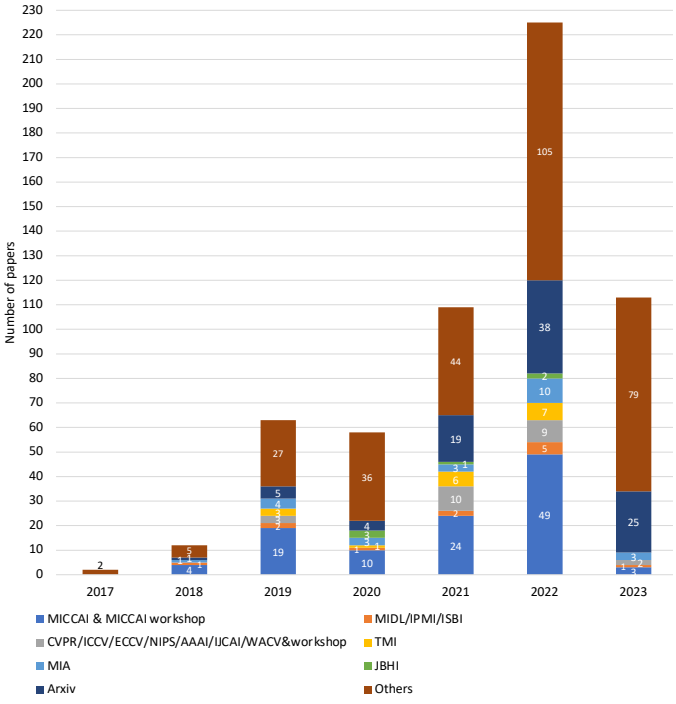

Fig. 2: Number of medical image segmentation papers from 2017 to 2023 whose titles include the word "Attention" or "Transformer". Different colors represent different sources of selected papers.

Over the past few years, attention mechanisms have played an increasingly important role in medical image segmentation. Figure 2 illustrates that the number of papers focused on attention mechanisms in medical image segmentation has seen a steady annual growth from 2017 to 2023 and lists the source of published papers. Convolutional neural networks (CNNs) and their variants have shown advantages in medical image segmentation tasks due to their ability to learn hierarchical features. Attention mechanisms were first introduced as a plug-in to CNNs, allowing the network to adjust weights dynamically based on features. For example, works in Wang et al. (2019e); Oktay et al. (2018) improved the segmentation performance of abdominal organs by integrating the attention mechanism into U-Net. To further model long-distance dependencies, the attention mechanism of the Transformer was proposed(Vaswani et al. (2017)), as demonstrated in the natural language processing field. Since the emergence of Vision Transformer in 2020, the Transformer attention mechanism has made significant breakthroughs in many visual tasks, drawing the attention of the medical community. As a result, there has been a rapid increase in Transformer-based papers related to medical image segmentation.

To keep up with the rapid increase in attention-based medical segmentation research, a survey of existing relevant works is urgently needed to provide the latest and comprehensive view of new research. The purpose of this paper is to summarize and categorize the current attention-based methods of medical image segmentation, while offering thoughtful commentary on the current state of the field and making suggestions for future research. Here, we classify the research into two types: traditional attention (Non-Transformer-based in this survey) and Transformer-based attention, as Transformer-based methods are increasingly recognized as a distinct and mainstream category in medical image segmentation research. Notably, Transformer blocks can serve as basic network blocks, while traditional attention should be combined with convolutional layers to form a block, which is also one of the bases for our classification.

In addition, we briefly compare this paper to various existing surveys which have reviewed attention methods and Transformer in medical image analysis. Shamshad et al. (2023), He et al. (2022), Azad et al. (2023b) and Li et al. (2023) provided a survey of Transformer-based applications in medical image analysis. In contrast, our work reviews the entire domain of attention mechanisms beyond just Transformer-based methods. Gonçalves et al. (2022) surveyed attention mecha-

nisms in medical image analysis tasks. However, segmentation is a challenging and critical task in medical image analysis, and the number of methods for this task exceeds that of the other tasks combined. Therefore, we have focused our summary of attention-based methods specifically on medical image segmentation on providing a deeper, task-specific understanding.

The rest of the paper is organized as follows (see Figure 3). Section 2 is the notation definition. Section 3 and Section 4 are the main parts of our survey, in which we revisit the basic principles of attention mechanism separately and extensively review the existing Non-Transformer-based and Transformer-based medical segmentation methods following our taxonomies. We summarize the main conclusion of this survey in and highlight several future challenges in the study of attention-based algorithms for medical image segmentation tasks in Section 5.

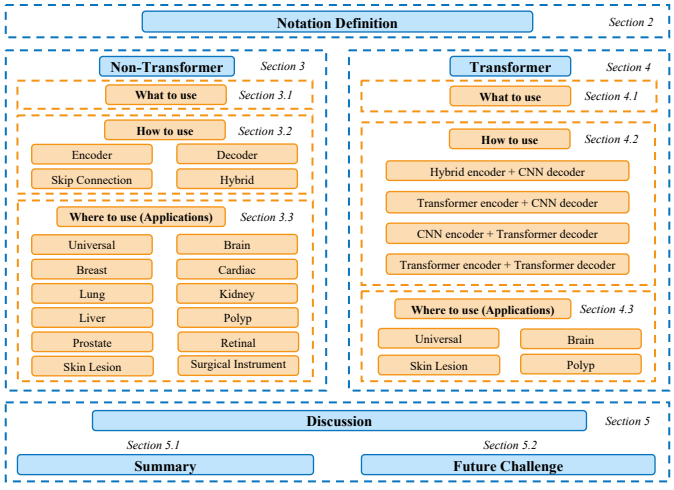

Fig. 3: Schematic structure of attention mechanisms in the medical segmentation and the relationship between the adjacent sections. The body of this survey mainly contains the notation definitions, Non-Transformer applications, Transformer applications, discussions, and future challenges.

2. Notation Definition

Table 1 gives some notation definitions mentioned in the following chapters to aid understanding.

3. Non-Transformer in Medical Segmentation

To provide researchers with a more comprehensive and clear understanding of the applications of Non-Transformer attention

Table 1: Key notations in this paper.

Type	Notation	Meaning
Universal	NM	Not mentioned
	NaN	Not a number
	A+B	Join the two datasets A and B together to train or test
	Enc	Encoder
	Dec	Decoder
	Skip.	Skip connection
Metric	P	Precision
	F1	F1-score
	HD	Hausdorff distance
	JC	Jaccard coefficent
	JI	Jaccard index
	OE	Overlapping error
	SE	Sensitivity
	SP	Specificity
	ACC	Accuracy
	AUC	Area under receiver operating characteristic (ROC)
	FDR	False discovery rate
	MSD	Mean surface distance
	OE	Overlapping error
	mIoU	Mean intersection-over-union

mechanisms in medical image segmentation, we will progressively categorize these studies into three aspects: (1) the types of attention mechanisms used in medical image segmentation (what to use), (2) the locations in the network where attention mechanisms are utilized (how to use), and (3) the specific clinical tasks where attention mechanisms are applied (where to use).

3.1. What to Use

3.1.1. Definition

The attention mechanism is inspired by the recognition process of the human visual system, allowing networks to focus on specific objects while ignoring irrelevant areas. This provides localized classification information, which is often desirable in image-processing networks. We will briefly introduce the theoretical basis of the attention mechanism and then illustrate common attention mechanisms in medical image segmentation tasks.

Following the description and taxonomy by Guo et al. (2022), almost all existing attention mechanisms can be formulated as

$$\text{Attention} = f(g(x), x) \tag{1}$$

where $g(x)$ is a specific kind of generated attention. $f(g(x), x)$ represents processing input vector x based on the attention $g(x)$

which is consistent with processing critical regions and getting information. Take self-attention as an example, we first transformed x into three distinct matrix representations: queries $Q \in \mathbb{R}^{n\times d_q}$, keys $K \in \mathbb{R}^{n\times d_k}$, and values $V \in \mathbb{R}^{n\times d_v}$, all with dimensions $d_q = d_k = d_v = d_{model}$. $g(x)$ and $f(g(x), x)$ can be represented as

$$g(x) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}}) \\ f(g(x), x) = g(x)V \tag{2}$$

where QK^T computes the relevance score between different entities, d_k is the scaling factor, softmax operation translates the score into probability and multiplying with V is to obtain the weighted matrix.

3.1.2. Non-Transformer attention

Commonly used Non-Transformer attention mechanisms can be categorized into three main types: channel, spatial, and temporal attention, as shown in Figure 4. These attention mechanisms can be used individually or in combination to enhance the network's performance in medical image segmentation tasks.

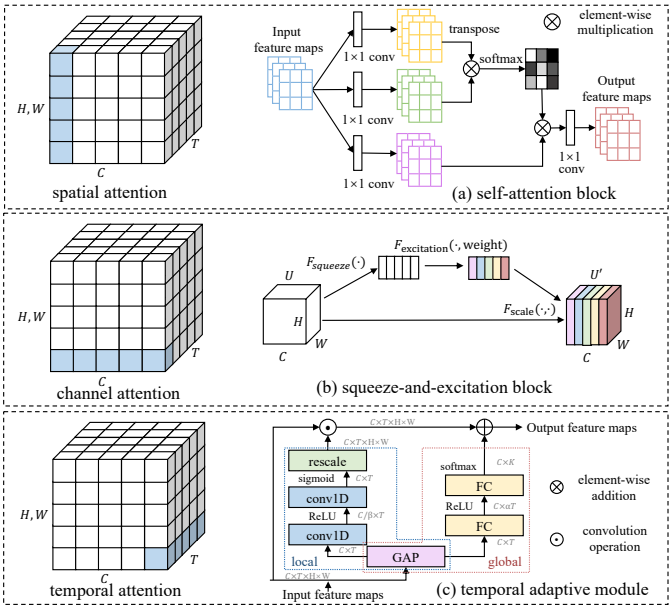

Fig. 4: Spatial, channel and temporal attentions, which follow Guo et al. (2022). Figure (a)(b)(c) follow Zhang et al. (2019a), Hu et al. (2018b) and Liu et al. (2021d).

Channel attention adaptively reweights each channel, treating them as representing different objects (Chen et al. (2017)). The concept of channel attention was first introduced in squeeze-and-excitation networks (SENet) by Hu et al. (2018b), using a squeeze-and-excitation (SE) block to capture channel-wise relationships. The squeeze module converts each channel into a single value using global average pooling, while the excitation module outputs an attention vector through fully-connected and non-linear layers. Subsequent works aimed to improve the squeeze or excitation process. For example, Qin et al. (2021) treated global average pooling as a case of the discrete cosine transform in the squeeze module and proposed the frequency channel attention net (FcaNet) based on information compression. Wang et al. (2020c) proposed an efficient channel attention (ECA) block, which takes the direct interaction between each channel and its k-nearest neighbors into account to improve the excitation module with less complexity. Meanwhile, Lee et al. (2019) adopted style pooling in the squeeze process and inserted a channel-wise fully-connected layer in the excitation module to reduce the computational cost in both steps.

In the medical image segmentation, SE block (Hu et al. (2018b)) has been applied by Yang and Qiu (2021). To avoid the loss of information in the SE block, Xie et al. (2021b) proposed introducing multiscale context information via parallel dilated convolution with different dilation rates. Other channel attention methods Wang et al. (2021h, 2019a); Guo et al. (2021a) have also been designed for specific tasks, focusing on channels with ample information while restraining irrelevant channels.

Spatial attention helps to identify the important regions in an image by assigning importance scores to different spatial regions in the feature map (bounded by width and height). Attention gates (Oktay et al. (2018)) utilize additive attention between the input and a gate signal collected at a coarse scale to obtain the gating coefficient for building a spatial attention weight map. It provides a modular and light paradigm highlighting important areas and suppressing features in unrelated regions. GENet (Hu et al. (2018a)) introduces a spatial recalibration function called a gather-excite module, similar to SENet (Hu et al. (2018b)), and then uses interpolation to form an attention map. To increase the receptive field, self-attention

has also been introduced. Wang et al. (2018) proposed the Non-local network, which augments each pixel of the convolutional features with contextual information (the weighted sum of the whole feature map) to encode the correlated patches in a long-range fashion.

We observe that attention gates (Oktay et al. (2018)), and Non-local networks (Wang et al. (2018)) often appear in medical image segmentation. Oktay et al. (2018) proposed the attention gate to guide the model's attention on targeted regions with a gating signal collecting from a course scale for abdominal multi-organ segmentation, and the attention gate is widely used and adapted in medical image segmentation (Duran et al. (2022); Kearney et al. (2019)). While in the attention map comes from the resampling of the prostate prediction instead of the preceding convolutional block in Duran et al. (2022). And Kearney et al. (2019) substitute the spatial attention coefficients for the global gating signal as opposed to previous study (Oktay et al. (2018)) to obtain more sensitive spatial information.

Besides, Mostayed et al. (2019) applied self-attention multiplication with decoder feature maps to reduce the number of parameters and learn the object boundaries. Ding et al. (2020a) proposed high-order attention, which allows each pixel to build its own global attention map and then constructs the attention map through graph transduction, thus capturing relevant context information at high order to enhance relevant pixels. Xu et al. (2021c) proposed a vector self-attention layer for long-range spatial reasoning with geometric priors and multi-scale calibration.

Furthermore, some customized spatial attentions are designed to solve specific problems, *e.g.*, blur boundaries. Some may develop the edge attention (Zhang et al. (2019c); Wang et al. (2020b)), while others regard the shape as prior knowledge to guide a shape attention (Li et al. (2020d); Zhang et al. (2021b)).

Temporal attention is usually seen as a dynamic time frame selection mechanism when the data has a temporal dimension, *e.g.*, a video. Li et al. (2019a) propose the global-local representation (GLTR) for temporal relation modeling through self-attention. While Liu et al. (2021d) propose the temporal adaptive module (TAM), adopting an adaptive kernel instead of self-attention with a local branch and global branch to capture complex temporal relationships with lower computational costs. Temporal attention may be applied in multi-frame images or videos clinically. Jin et al. (2019) utilize the inherent temporal clues from the instrument motion as prior. Ahn et al. (2021) adopt spatio-temporal attention in multi-time frames to obtain interframe consistency among the images.

Besides, spatial attention may combine with channel or temporal attention to obtain more comprehensive information. The convolutional block attention module (CBAM, Woo et al. (2018)) calculates channel and spatial attention in serial separately, in which the channel attention module utilizes two parallel branches via max-pool, and avg-pool operations and the spatial attention module applies a convolution layer with a larger kernel to generate the attention map. Thus, CBAM can emphasize useful channels and enhance informative local regions. The dual attention network (DANet, Fu et al. (2019)) also computes channel and spatial attention separately via self-attention and fuses them for final results. Moreover, the spatial attention in CBAM (Woo et al. (2018))(with max-pooling and average-pooling) is also introduced with the original version (Guo et al. (2021b)) and improved version(*e.g.,* Guo et al. (2020); Jiang et al. (2021)), in which Guo et al. (2020) apply max-pooling and average-pooling operations along the channel axis as CBAM Woo et al. (2018) and concatenate them to produce an efficient feature descriptor, while Jiang et al. (2021) utilize CBAM mechanism on fused multi-scale to re-distribute the scale-wise weights.

3.2. How to Use

In Non-Transformer-based methods, the attention mechanism is typically used as a plug-in sublayer inserted into the convolutional block. The plug-in location can be divided into three categories: in the encoder, in the decoder, in the skip connection, and hybrid, based on where the attention layer occurs.

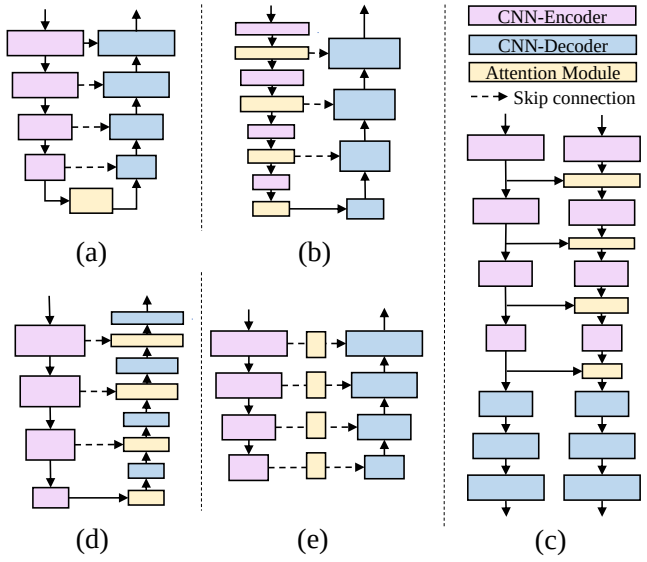

Fig. 5: Locations where the attention layer occurs in Non-Transformer-based methods, including (a) bottleneck (top of the encoder), (b)(c) each stage of the encoder in a network or between two networks, (d) decoder and (e) skip connection.

3.2.1. In the Encoder

The attention mechanism is often introduced into the encoder to enlarge the receptive field and extract richer encoded information.

In the bottleneck. Some researchers add the attention mechanism at the encoder's bottleneck, as shown in Fig 5 of (a). With respect to the channel-based methods, Jiang *et al.* propose a scale-selection mechanism on concatenated multi-scale features, utilizing channel attention to calculate the correlation weight for each feature map with different scales to supplement the information lost in down-sampling. Yang et al. (2022a) propose an improvement on ECANet (Wang et al. (2020c)) by introducing a new branch that captures the distinct interactions between the current channel and its k nearest channels after channel shuffling, thus enlarging the receptive fields. Differently, to highlight useful information and suppress redundant features, some adopt the spatial attention module(Cheng et al. (2022b); Guo et al. (2021b); Hu et al. (2019b); Tang et al. (2019)), *e.g.*, CBAM (Woo et al. (2018)) and the criss-cross attention (Huang et al. (2019)). Liu et al. (2019, 2021a) create a pyramid-like feature structure Li et al. (2018) at the final output of the encoder, where Liu et al. (2019) utilizes different convolution kernels and Liu et al. (2021a) employs a pyramid local attention module to capture supporting information from compact and sparse (*i.e.*, different distance) neighboring contexts. Moreover, since the convolution operations can only capture local information, self-attention is chosen to model global relations to solve this problem (Ding et al. (2019, 2020a); Mou et al. (2019); Fan et al. (2020b)). In particular, Ding et al. (2019) propose high-order attention and its improved version Ding et al. (2020a) to reduce computational costs, setting a threshold in the similarity matrix to reduce the noise of weak correlations. While Mou et al. (2019) apply channel-wise and spatial-wise self-attention in parallel to better aggregate features. Qu et al. (2022) propose a dual-path network to utilize the complementation of arterial and venous phases visual information of CT and the attention module is inserted into the bottleneck of two networks to generate cross-phase reliable feature correlations as well as the channel dependencies.

At each stage in a network. To take advantage of multi-scale information, some works integrate the attention module into each stage in the encoder (Fig 5 (b)). For example, Zhang et al. (2021a) propose a plug-and-play and lightweight attention module in each layer, which aggregates spatial information from x and y directions into channel attention through an adaptive global pooling. Singh et al. (2019) adopt a multi-scale input strategy (Van Noord and Postma (2017)), where the input images are resized into three different scales and the corresponding feature maps are aggregated as the encoder layer input. They also propose a factorized channel attention module with a 1-D kernel factorized convolution at each scale. Differently, Wang et al. (2019c) construct a 2.5D network by combining the target slice with neighboring slices and apply both channel attention and spatial attention serially at each scale to obtain the corresponding response.

At each stage between two networks. Additionally, attention mechanisms are sometimes introduced to facilitate information transfer between two networks in the encoder, as illustrated in Fig 5 (c). Those networks receive the same inputs. Min et al. (2019) design a two-stream mutual attention

network in semi-supervised segmentation to render the network robust in unclean data, whose input is the mixture of labeled data with manual labels and unlabeled data with pseudo labels. The features are exchanged in the mutual attention modules at each stage to discover incorrect labels and weaken their influence during parameter updating. Wang et al. (2019b) and Chen et al. (2019a) propose multi-task hybrid supervised networks to improve segmentation performance. Specifically, Wang et al. (2019b) corrects the segmentation feature distribution using weak annotations from a detection network through attention modules, and Chen et al. (2019a) leverages foreground and background information from the segmentation branch to build the auxiliary reconstruction task.

It is worth noting that Ahn et al. (2021) have proposed a novel spatiotemporal attention module in the encoder that can leverage the inter-frame consistency in echocardiography images. This module operates on a target slice and its neighboring frames and represents the first instance of spatiotemporal attention being applied to medical image segmentation.

3.2.2. In the Decoder

The decoder's primary task is to reconstruct the latent representation into an image that satisfies the specified criteria. Incorporating attention mechanisms in the decoder assists in generating dense feature maps by aggregating additional information, as demonstrated in (e) of Fig 5. These mechanisms can be classified based on their type of attention.

In the channel-wise applications, Ni et al. (2019a) incorporate the attention module at each layer in the decoder to fuse multi-level features by utilizing the global context of high-level features as guidance information. Additionally, both Tomar et al. (2021) and Jia et al. (2019) utilize a dual-decoder architecture with one of the decoders acting as an auxiliary decoder to enhance segmentation. The attention mechanism is inserted in the main decoder to extract semantically discriminative intra-slice features in 3D prostate MRI segmentation (Jia et al. (2019)) or between two decoders to provide an attention map (Tomar et al. (2021)).

Spatial attention-based networks often benefit from incorporating edge/shape information to improve the reconstruction of deep semantic features into a high-resolution segmentation map. For instance, Karthik et al. (2022) use an attention module to capture contextual information from the contour feature maps in its spatial neighborhood. Wei et al. (2021a) use deep features to filter out background noise in shallow features and preserve edge information. Li et al. (2020d) use the left atrial boundary as an attention mask on scar features to perform shape attention. Qin et al. (2020) propose an attention distillation technique to pass fine-grained details down to lower-resolution attention maps, improving their performance.

The channel-spatial attention modules are typically integrated into convolutional blocks in the decoder at each layer. Especially, Fang and Han (2020) found that combining the channel and spatial attention using element-wise addition achieved the best results in their experiments. Gu et al. (2020) add another scale attention on the concatenated feature to achieve comprehensive attention based on the spatial-channel attention network.

Here, as experiments in Gu et al. (2020) and Qin et al. (2020) have compared that the decoder-side attention block performs better than the encoder-side attention block. This may be because the encoders learn general feature, which is probably not strongly related to the target description. Thus applying an attention mechanism in the encoder may suppress some important information. However, this assumption needs to be further validated on a wider range of datasets with different modalities and under various scenarios to understand better the impact of attention mechanism placement on segmentation performance.

3.2.3. In the Skip Connection

The skip connection plays a crucial role in restoring the full spatial resolution at the network's output, improving medical image segmentation performance. The addition of the attention module (as shown in Fig 5 (e)) can further enhance this process by mitigating the semantic gap between the encoder and decoder. Many examples in medical image segmentation demonstrate this.

The attention gate (Oktay et al. (2018)) is the most com-

mon form of attention mechanism in the skip connection, which is incorporated to selectively amplify informative features and suppress noise and irrelevant responses. Some researchers adopt the original spatial attention gate in their networks(*e.g.,* Zuo et al. (2021); Li et al. (2020b); Kearney et al. (2019); Wu et al. (2019); Li et al. (2020a)), while others keep the form with different attention computations. For instance, Khanh et al. (2020) improve it with channel-spatial attention instead. Inspired by the focal loss, Yeung et al. (2021) add a focal parameter to reduce the contributions of easy examples for harder examples learning. Li et al. (2019c) also incorporate the loss function into their network design by designing a topology-aware loss. This loss function aims to model the structure properties by encouraging the probability of connectivity in the neighboring region to be high.

The attention module can also be inserted into the skip connection at each stage in a different way than the attention gate. At first, well-known attention methods are applied to the skip connection(Yang and Qiu (2021); Hu et al. (2021); Sinha and Dolz (2020); Cheng et al. (2020); Zhou et al. (2021b); Guo et al. (2020, 2021a)), *e.g.*, SE (Hu et al. (2018b)), CBAM (Woo et al. (2018)), Efficient channel attention(ECA, Wang et al. (2020c)), Concurrent Spatial and Channel Squeeze & Excitation (scSE, Roy et al. (2018)), self-attention (Vaswani et al. (2017)). Besides, the attention module is improved for various purposes. To respond to the different directions of the features for better information conservation, Tong et al. (2021a) design a side attention block through different shapes of convolution kernels. To bring in multi-scale property for its richer semantic information, Xia et al. (2022) input the multi-scale encoded feature to the skip connection and then apply residual attention to highlight salient areas for each stage. Differently, Wang et al. (2021h) and Lyu et al. (2020) design multi-scale attention directly to solve the problem. Wang et al. (2021h) applies channel attention at each layer and only adds additional hybrid dilated attention layers with different dilation rates at the bottleneck. Lyu et al. (2020) inserts the multi-scale attention module at each layer of their network, using a multi-branch architecture with different numbers of convolutional operations to capture semantic information without reducing dimensions. Additionally, some researchers consider that features from different levels should be handled differently due to their characteristics. Tong et al. (2019) perform spatial attention on low-level features and channel attention on high-level features. Fan et al. (2020a) and Lou et al. (2021) only apply the attention mechanism to high-level features to reduce computational costs. Reverse attention is designed in the former to establish the relationship between areas and boundary cues, and axial reverse attention is designed in the latter to analyze localization information.

3.2.4. Hybrid

To combine the advantages of the methods mentioned above and apply various attention to the network, some researchers integrate attention into multiple locations in the architecture design.

Among them, the most common form is inserting the attention module into both the encoder and decoder in each layer with different attention modules. The attention modules, like CBAM (Guo et al. (2020)), SE block (Yin et al. (2022)) and its varieties (Xie et al. (2021b); Li and Rahardja (2021); Wang et al. (2022b)), are usually adopt. Moreover, since SE block and its varieties recalibrate spatial information for the same single weight, Lu et al. (2022) propose double group attention modules to extract multigroup weighs of the feature maps to strengthen the spatial information. Furthermore, Xu et al. (2021c) consider introducing long-term dependency to exploit context prior. Some works also take the functional differences between the encoder and decoder and apply distinct attention mechanisms. The encoder is typically designed to extract features from images while preserving as much detail as possible, while the decoder is responsible for recovering enough information from the features to enable accurate segmentation. As a result, it is common practice to apply channel attention to the encoder to enhance the representation of target features, while using spatial attention in the decoder to emphasize the position of useful areas when fusing low-level and high-level features.

This approach has been demonstrated in various studies, such as Li et al. (2021c) and Guo et al. (2021c). The attention module can also be applied between two frameworks for information transactions in the encoder and decoder. Some works adopt the same inputs but different networks design. Due to the task relationship brought inherently by the data, Xu et al. (2019) utilize the properties of the progressive inclusion of subregions to guide the current task with the outputs of previous tasks as salient regions on the BraTS dataset. In Xu et al. (2020), the connection between the prostate bed and the bladder and rectum segmentation task is considered, and the feature from the auxiliary network is transferred to the target prostate bed segmentation network through the attention module. Besides, the attention mechanism is also integrated into the skip connection based on the encoder-decoder attention architecture (Yao et al. (2022)).

However, there is currently no experiment that demonstrates the individual effects of adding attention at different locations on performance gain.

3.3. Where to use (Applications)

After discussing the basic types and the embedded locations of attention mechanisms used in medical image segmentation, we now introduce the specific application tasks. These tasks mainly include brain, breast, cardiac, kidney, liver, lung, polyp, prostate, eye, skin lesion, and surgical instrument segmentation. Some universal attention-based medical segmentation models may not be designed for specific tasks but for multiple organs/lesions segmentations on different modalities datasets as shown in Table 2 In addition, we also give a series of tables to illustrate applications on specific organs with attention types (What to use), locations (How to use), and their performance on the common dataset.

Brain. The brain segmentation applications focus on the brain tissue (Sun et al. (2019)) and tumor segmentation task (Noori et al. (2019); Yuan et al. (2019); Islam et al. (2019); Chen et al. (2019a); Zhou et al. (2018, 2020a); Akil et al. (2020); Zhou et al. (2020b); Xu et al. (2019); Zhang et al. (2020); Maji et al. (2022); Fang et al. (2022)), here we list the applications on glioma segmentation as it is the mainstream due to the muli-modal brain tumor segmentation benchmark (BraTS, Menze et al. (2014)) (shown in Table 3). Especially, most of the attention designs are based on SE blocks (Hu et al. (2018b)) and integrated into the decoder. Some works consider that unrelated information such as the contour as well as internal structures of the brain may be preserved in the encoder and insert the attention layer into the encoder to emphasize tumor-related features (Yuan et al. (2019); Noori et al. (2019)). Others take into account the potential subregion correlations between the whole tumor (WT), enhanced tumor (ET), and tumor core (TC) and propose a cascaded attention mechanism in the decoder (Xu et al. (2019); Zhou et al. (2020a)). In a semi-supervised approach proposed by Chen et al. (2019a), labels generated using an attention mechanism are used to help the reconstruction task separate foreground and background, thus aiding in the segmentation of brain tumors and white matter hyperintensities due to the shared encoder.

Breast. The segmentation tasks on breast focus on breast anatomy (Lei et al. (2020a)), breast cancer (Lee et al. (2020); Zhuang et al. (2019); Vakanski et al. (2020); Punn and Agarwal (2022b)) and breast mass (Li et al. (2019b); Sun et al. (2020)) as shown in Table 4. Through this literature, we can conclude that most attention methods used in this domain are directly transplanted, *e.g.*, SE blocks (Hu et al. (2018b)), attention gate (Oktay et al. (2018)). Some works also attempt to introduce domain knowledge. For instance, Vakanski et al. (2020) obtain the saliency map (Xu et al. (2021a)) through prior knowledge constraints (the information of the degree of connectedness and confidence of connectedness between the image regions) and utilize precomputed saliency maps in the attention module that point out to target spatial regions.

Cardiac. The segmentation tasks surrounding the heart mainly include left ventricle (Ge et al. (2019); Ahn et al. (2021); Lu et al. (2022)), left atrial (Li et al. (2020d)), and cardiac anatomical structure segmentation (Tong et al. (2019); Guo et al. (2021c); Liu et al. (2021a,b); Wang et al. (2022d); Ding et al. (2020b)) and we list the cardiac anatomical structure seg-

Table 2: An overview of non-Transformer methods for universal segmentation. 'How/What' means 'How/What to use'.

Author	How	What	Datasets	Metric	Data split (Train:Val:Test)	Modality (Type)
Wang et al. (2019b)	Skip.	Channel	1.Lung nodule dataset (NM) 2.Inner ear dataset (NM)	Dice:0.8490 Dice:0.8873	160:80:80 66:40:40	CT (2D)
Kaul et al. (2019)	Skip.	Spatial	1.ISBI 2017 (Codella et al. (2018)) 2.Finding and Measuring Lungs in CT Data	ACC:0.9214 JI:0.7562 Dice:0.8315 SE:0.7673 ACC:0.9932 JI:0.9965 SE:0.9757	2000:150:600 80%:20%:NM	Dermoscopy (2D) CT (2D)
Zhang et al. (2019c)	Skip.	Spatial	1.REFUGE (Orlando et al. (2020)) 2.Drishti-GS (Sivaswamy et al. (2014)) 3.DRIVE (Staal et al. (2004)) 4.MC (Jaeger et al. (2014)) 5.Finding and Measuring Lungs in CT Data	Dice:0.8912(OC) 0.9529(OD) Dice:0.9314(OC) 0.9752(OD) ACC:0.9560 mIoU:0.7744 ACC:0.9865 mIoU:0.9420 ACC:0.9868 mIoU:0.9623	400:400:NaN 50:51:NaN 20:NaN:20 80:NaN:58 214:NaN:53	Retinal fundus (2D) Retinal fundus (2D) Retinal fundus (2D) X-Ray (2D) CT (2D)
Lin et al. (2020)	Skip.	Spatial	1.ETIS (Silva et al. (2014)) 2.CVC-ColonDB (Bernal et al. (2012)) 3.ISBI 2016 (Gutman et al. (2016)) 4.ISBI 2017 (Codella et al. (2018))	Dice:0.9602 JI:0.8133 ACC:0.9936 Dice:0.9386 JI:0.8234 ACC:0.9823 Dice:0.9172 JI:0.8508 ACC:0.9823 Dice:0.8775 JI:0.7538 ACC:0.9321	156:NaN:40 304:NaN:76 900:NaN:379 2000:NaN:600	Endoscopy (2D) Endoscopy (2D) Dermoscopy (2D) Dermoscopy (2D)
Ding et al. (2020a)	Skip.	Spatial	1.REFUGE (Orlando et al. (2020)) 2.Finding and Measuring Lungs in CT Data 3.DRIVE (Staal et al. (2004))	mDice:0.9302 ACC:0.9945 Sen:0.9879 mIoU:0.9849 ACC:0.9712 F1:0.8300 SE:0.8297 SP:0.9843 mIoU:0.7094	400:400:NaN 214:NaN:53 20:NaN:20	Retinal fundus (2D) CT (2D) Retinal fundus (2D)
Zuo et al. (2021)	Skip.	Spatial	1.CVC- 2018 (Tschandl et al. (2018)) 2.DRIVE (Staal et al. (2004)) 3.Finding and Measuring Lungs in CT Data	ACC:0.9277 F1:0.8660 AUC:0.8957 ACC:0.9555 F1:0.8213 AUC:0.9790 ACC:0.9950 F1:0.9868 AUC:0.9921	1815:259:520 20:NaN:20 134:54:79	Dermoscopy (2D) Retinal fundus (2D) CT (2D)
An and Liu (2021)	Skip.	Spatial	1.IBSR (of Massachusetts General Hospital) 2.JSRT (Rikitake et al. (2019))	Jaccard similarity:0.9725 Jaccard similarity:0.9836	NM NM	MRI (2D) CT (2D)
Xia et al. (2022)	Skip.	Spatial	1.Finding and Measuring Lungs in CT Data 2.Bladder dataset (Private dataset) 3.Tipdm Cup Challenge 4.KiTs (Heller et al. (2019))	ACC:0.9960 AUC:0.9947 Dice:0.9897 ACC:0.9946 AUC:0.9679 Dice:0.9679 ACC:0.9981 AUC:0.9367 Dice:0.9977 ACC:0.9924 AUC:0.9356 Dice:0.9729	80%:NaN:20%	CT (2D)
Cheng et al. (2022b)	Skip.	Spatial	1.COVID-19 CT segmentation dataset (of Medical and Radiology) 2.CVC-ClinicDB (Bernal et al. (2015)) 3.CVC- 2018 (Tschandl et al. (2018)) 4.Lung nodule competition 2017 (Liao et al. (2019))	F1:0.7516 AUC:0.8382 F1:0.7998 AUC:0.8804 F1:0.8627 Dice:0.9041 F1:0.9868 AUC:0.9929	45:5:50 414:85:113 1815:259:520 571:143:307	CT (2D) Endoscopy (2D) Dermoscopy (2D) CT (2D)
Min et al. (2019)	Skip.	Spatial and Channel	1.HVMSR 2016 (Pace et al. (2015)) 2.BRATS-2015 (Kistler et al. (2013))	Overall score:-0.024 Dice:0.792	10:NaN:10 244:NaN:30	MRI (3D)
Sinha and Dolz (2020)	Skip.	Spatial and Channel	1.CHAOS (Kavur et al. (2021b)) 2.HSVMR 2016 (Pace et al. (2015)) 3.Brain segmentation dataset of MSD (Antonelli et al. (2022))	Dice:0.8675 MSD:0.66 Dice:0.8320 MSD:1.19 Dice:0.8037 MSD:0.90	13:2:5 60%:20%:20% 388:48:48	MRI (3D)
Khanh et al. (2020)	Skip.	Spatial and Channel	1.CVC-ClinicDB (Bernal et al. (2015)) 2.VIP-VUP18 (Video and Cup) 3.TCGA (Buda et al. (2019))	Dice:0.7331 Dice:0.5626 Dice:0.8583	80%:20%:NaN	Endoscopy (2D) CT (2D) MRI (2D)
Cheng et al. (2020)	Skip.	Spatial and Channel	1.Chest X-ray collection (Demner-Fushman et al. (2016)) 2.Kaggle 2018 data science bowl (Hamilton (2018)) 3.Herlev (Zhao et al. (2020); Zhang et al. (2017))	P:0.9860 IoU:0.9669 Dice:0.9832 P:0.9022 IoU:0.8209 Dice:0.8989 P:0.9413 IoU:0.8862 Dice:0.9321	80%:10%:10%	X-ray (2D) Retinal fundus (2D) Mircoscopy (2D)
Fang and Han (2020)	Dec	Spatial and Channel	1.LUNA 2.ISIC (Tschandl et al. (2018))	F1:0.9841 AUC:0.9897 F1:0.8700 AUC:0.9310	70%:NaN:30% 1815:259:520	CT (2D) Dermoscopy (2D)
Gu et al. (2020)	Dec	Spatial and Channel	1.ISIC 2018 (Tschandl et al. (2018)) 2.Placenta and Fetal Brain Segmentation (Private dataset)	Dice:0.9208 Dice:0.8708 (placenta), 0.9588 (brain)	1816:260:518 1050:150:300	Dermoscopy (2D) MRI (2D)
Gao et al. (2021a)	Skip.	Spatial and Channel	1.BUS (Yap et al. (2017)) 2.CVC-ClinicDB (Bernal et al. (2015)) 3.ISBI-2014 (Lu et al. (2016, 2015))	Dice:0.8539 IoU:0.7706 P:0.8831 Dice:0.8535 IoU:0.7861 P:0.8950 Dice:0.9082 IoU:0.8331 P:0.8822	NM NM 45:NaN:90	Ultrasound (2D) Endoscopy (2D) Microscopy (2D)
Xu et al. (2021c)	Enc & Dec	Spatial and Channel	1.Fetal A4C (Private dataset) 2.Fetal head (van den Heuvel et al. (2018))	Dice:0.849 HD:3.421 Dice:0.971 HD:3.234	70%:NaN:30%	Ultrasound (2D)
Yao et al. (2022)	Enc, Dec & Skip.	Spatial and Channel	1.Vestibular schwannoma segmentation dataset (Shapey et al. (2021)) 2.BCV (Landman et al. (2015)) + CHAOS (Kavur et al. (2021a))	mDice:0.8233 mDice:0.8704 (MRI-CT), 0.9125 (CT-MRI)	NM	MRI (3D) CT, MRI (3D)

mentation applications in Table 5. Automatic cardiac segmentation is by no means a trivial task, as some parts of the cardiac borders are not obvious due to the low contrast to the surrounding tissue. Thus, the attention blocks are designed pixel-wise or consistent (*e.g.*, neighbor consistent (Liu et al. (2021a)), inter-frame consistent (Ahn et al. (2021)), category consistent (Ding et al. (2020b))) to identify the boundary. Moreover, the attention mechanism is also implemented in task transferring and domain transferring in the cardiac segmentation and qualification analysis tasks Ge et al. (2019); Li et al. (2020d).

Lung. The lung-related segmentation task includes lung anatomical segmentation (Tang et al. (2019)), lung airways segmentation (Qin et al. (2020); Tan et al. (2022)), and lesion/disease segmentation (Chen et al. (2020); Zhang et al. (2021b); Xie et al. (2021b); Zhou et al. (2021b); Karthik et al. (2022); Hu et al. (2022); Punn and Agarwal (2020); Yin et al. (2022)), most of which integrate spatial attention into the network (as shown in Table 6) as there are strong connections between the locations of lesions and lung. It is worth noting that adding a sub-network to offer the lungs contour maps is an efficient and common way for COVID-19 lesion segmentation (Yin et al. (2022); Punn and Agarwal (2020); Xie et al.

Table 3: An overview of non-Transformer methods for brain segmentation, in which WT, TC, ET, WM, GM, and CSF is the whole tumor, tumor core, enhanced tumor, white matter, gray matter, and cerebrospinal fluid seperately.

Author	How	What	Datasets	Metric			Data split(Train:Val:Test)	Modality (Type)
Noori et al. (2019)	Dec	Channel	1.BRATS 2017 (Menze et al. (2014); Bakas et al. (2017))	Dice:0.791 (ET)	0.885 (WT)	0.783 (TC)	NM	MRI (2D)
			2.BRATS 2018 (Menze et al. (2014); Bakas et al. (2017))	Dice:0.813 (ET)	0.895 (WT)	0.823 (TC)		
Zhou et al. (2018)	Enc & Dec	Channel	BRATS2018 (Menze et al. (2014); Bakas et al. (2017))	Dice:0.7775 (ET)	0.8842 (WT)	0.7960 (TC)	NM	MRI (3D)
				HD95:2.9366 (ET)	5.4681 (WT)	6.8773 (TC)		
Sun et al. (2019)	Enc & Dec	Spatial	1.MRBrainS13 (Mendrik et al. (2015))	Dice:0.8656 (GM)	0.8886 (WM)	0.8553 (CSF)	80%:NaN:20%	MRI (3D)
			2.MALC12 (Landman and Warfield (2012))	Dice:0.8482			20:NaN:10	
Yuan et al. (2019)	Dec	Spatial	Brain Tumors Task of MSD (Antonelli et al. (2022))	Dice:0.5361 (T1Gd)	0.8155 (FLARI)	0.7654 (T2)	50%:NaN:50%	MRI (2D)
Chen et al. (2019a)	Skip.	Spatial	1.BRATS 2018 (Menze et al. (2014); Bakas et al. (2017))	Dice:0.7702			120:50:50	MRI (2D)
			2.WMH17 (Kuijf et al. (2019))	Dice:0.7204			30:10:20	
Islam et al. (2019)	Dec	Spatial and Channel	BRATS 2019 (Menze et al. (2014); Bakas et al. (2017, 2018))	Dice:0.7780 (ET)	0.8689 (WT)	0.7771 (TC)	NM	MRI (3D)
Xu et al. (2019)	Dec	Spatial	BRATS 2018 (Menze et al. (2014); Bakas et al. (2017))	Dice: 0.8171 (ET)	0.9118 (WT)	0.8619 (TC)	NM	MRI (3D)
Zhou et al. (2020a)	Dec	Channel	1.BRATS 2018 (Menze et al. (2014); Bakas et al. (2017))	Dice:0.8111 (ET)	0.9078 (WT)	0.8575 (TC)	NM	MRI (3D)
			2.BRATS 2017 (Menze et al. (2014); Bakas et al. (2017))	Dice:0.7852 (ET)	0.9071 (WT)	0.8422 (TC)		
			3.BRATS 2015 (Menze et al. (2014))	Dice:0.65 (ET)	0.87 (WT)	0.75 (TC)		
Maji et al. (2022)	Skip.	Spatial	BRATS2019 (Menze et al. (2014); Bakas et al. (2017, 2018))	Dice:0.801 (ET)	0.911 (WT)	0.876 (TC)	4700:775:1000	MRI (2D)
				mIoU:0.668 (ET)	0.838 (WT)	0.781 (TC)		

Table 4: An overview of non-Transformer methods for breast segmentation.

Author	How	What	Datasets	Metric	Data split (Train:Val:Test)	Modality (Type)
Zhuang et al. (2019)	Skip.	Spatial	1.ultrasoundcases	ACC:0.9791 Dice:0.8469 F1:0.8478 IoU:0.8067 AUC:0.9227	NaN:NaN:100%	Ultrasound (2D)
			2.Private dataset		NaN:NaN:100%	
			3.Breast ultrasound dataset (Yap et al. (2017))		730:127:NaN	
Li et al. (2019b)	Skip.	Spatial	DDSM (PUB et al.)	F1:0.8224 SE:0.7789	66%:17%:17%	manmography (2D)
Vakanski et al. (2020)	Enc	Spatial	BUSIS (Xian et al. (2018))	Dice:0.901 JI:0.832 ACC:0.979 AUC:0.955	80%:20%:NaN	Ultrasound (2D)
Lee et al. (2020)	Enc	Channel	Breast ultrasound dataset (Yap et al. (2017))	ACC:0.97794 F1:0.7658 IoU:0.6226	146/147:NaN:16/17	Ultrasound (2D)
Lei et al. (2020a)	Enc	Spatial and Channel	1.Private dataset	Dice:0.866 P:0.954 ACC:0.922	NM	Ultrasound (2D)
			2.Private dataset	Dice:0.911 P:0.912 ACC:0.963		
Sun et al. (2020)	Dec	Channel	1.CBIS-DDSM (PUB et al.; Lee et al. (2017))	Dice:0.818 SE:0.849	690:168:NaN	Mammography (2D)
			2.Inbreast (Moreira et al. (2012))	Dice:0.791 SE:0.808	80%:20%:NaN	
Punn and Agarwal (2022b)	Skip.	Spatial	1.BUSIS (Xian et al. (2018))	ACC:0.990 Dice:0.937 mIoU:0.910	70% : NaN : 30%	Ultrasound (2D)
			2.BUSI (Al-Dhabyani et al. (2020))	ACC:0.970 Dice:0.914 mIoU:0.899		

Table 5: An overview of non-Transformer methods for cardiac segmentation, in which Endo, Epi, ED, ES, LV, RV and MYO refers to endocardium, myocardium, end diastole, end systole, left ventricle, right ventricle and myocardium separately.

Author	How	What	Datasets	Metric	Data split (Train:Val:Test)	Modality (Type)
Tong et al. (2019)	Skip.	Spatial	ACDC 2017 (Bernard et al. (2018))	Dice:0.966 (LV ED), 0.918 (LV ES), 0.948 (RV ED), 0.898 (RV ES), 0.905 (MYO ED), 0.915 (MYO ES)	100:NaN:50	MRI (2D)
Li et al. (2020d)	Dec	Spatial	MICCAI 2018 LA challenge (Xiong et al. (2021))	ACC:0.867 Dice:0.543 (Scar)	40:NaN:20	MRI (3D)
Liu et al. (2021a)	Enc	Spatial	1.CAMUS (Leclerc et al. (2019))	Dice:0.951 (Endo.ED), 0.931 (Endo.ES), 0.962 (Epi.ED), 0.956 (Epi.ES)	NM	Echocardiographic (2D)
			2.Sub-EchoNet-Dynamic (Ouyang et al. (2020))	Dice:0.942 (Endo.ED), 0.918 (Endo.ES), 0.951 (Epi.ED), 0.943 (Epi.ES)	1600:400:500	
Wang et al. (2022d)	Dec	Spatial	MyoPS 2020 (Li et al. (2022b))	Dice:0.658 (Scar), 0.720 (Scar+Edema) JC:0.535 (Scar), 0.577 (Scar+Edema) HD:14.13 (Scar), 14.12 (Scar+Edema)	20:2:NaN	CMR (2D)

(2021b)). Another noteworthy is that Zhang et al. (2021b) proposed multi-layer attention specifically for COVID-19 lesion segmentation. It includes an edge attention module to learn semantic edge information from low-level features, a shape at-

tention module with a circular shape filter to enhance attention ability for round or quasi-round pulmonary nodules, and a local attention module to learn from the adjacent region. This attention design has significantly improved the segmentation dice from 46.8% to 56.3%.

Kidney. Since renal tumors in CT or MRI images can appear similar to their parenchyma and other nearby tissues, accurately segmenting them can be challenging (Hu et al. (2019b); Myronenko and Hatamizadeh (2019); Sabarinathan et al. (2019); Jia et al. (2022); Xuan et al. (2022)). Table 7 shows that spatial attention is the most commonly used attention mechanism for kidney tumor segmentation, as there is a regional inclusive relationship between the kidney and tumors (Xuan et al. (2022)) similar to the lung lesion segmentation and the information of edge is added through the spatial attention mechanism.

Liver. Most liver-related segmentation tasks focus on liver (Haseljić et al. (2023)), liver vessel (Yan et al. (2020); Kuang et al. (2023)), and tumors segmentation (Jin et al. (2020); Chen et al. (2019b); Jiang et al. (2019a); Li et al. (2020a); Fan et al. (2020b); Xu et al. (2021d); Zhang et al. (2021a); Bi et al. (2022); Zhang et al. (2022a)). The heterogeneous and diffusive shape, as well as the relatively small sizes of most lesions, makes the segmentation a rough task. It is noted that the spatial attention modules are usually integrated into the skip connection in those applications (shown in Table 8). Some researchers may prefer to perform liver segmentation before the tumor segmentation (Jin et al. (2020); Jiang et al. (2019a)).

Polyp. The challenge of polyp segmentation is the diversity of size, color, and texture in the same type of polyps and the blurred boundary between a polyp and its surrounding mucosa (Fan et al. (2020a); Tomar et al. (2021); Wei et al. (2021a); Kim et al. (2021); Yeung et al. (2021); Yang et al. (2022a)). Spatial attention is the mainstream to obtain spatial features with deep information, and there are some public datasets for performance comparison (in Table 9). Interestingly, some researchers invariably choose to apply attention only to high-level features in parallel for different reasons. Fan et al. (2020a) argue that low-level feature demands more computational resource and less contributes to performance. Wei et al. (2021a) state that deep features are course in boundary but have a clear background and thus they utilize the clear feature to filter out the background noise in the shallow features and propose the shallow attention module. Kim et al. (2021) use the parallel axial attention on the outputs of encoder blocks and propose the uncertainty augmented context attention on the coarse-to-fine segmentation to build a bottom-up stream, which incorporates high-level semantic features for better performance.

Prostate. Developing automatic prostate segmentation remains challenging due to the missing/ambiguous boundary and inhomogeneous intensity distribution of the prostate, as well as the large variability in prostate shapes (Wang et al. (2019d); Liu et al. (2019); Kearney et al. (2019); Jia et al. (2019); Lei et al. (2020b); Xu et al. (2020); Duran et al. (2022)). In our knowledge, the mainstream attention types is spatial attention (as shown in Table 10). Moreover, the performance of these methods cannot be fairly compared since there are few publicly available dataset benchmarks.

Retinal. Retinal segmentation contains the iris segmentation (Lian et al. (2018)), disc and cups segmentation (Jiang et al. (2019b); Zhang et al. (2019c); Bhatkalkar et al. (2020)), and the retinal vessel segmentation (Mou et al. (2019); Zhang et al. (2019b); Li et al. (2019c); Luo et al. (2019); Wu et al. (2019); Wang et al. (2019a); Hu et al. (2019a); Li et al. (2020e); Guo et al. (2021b); Lyu et al. (2020); Lv et al. (2020); Guo et al. (2020); Li et al. (2020c); Tong et al. (2021a); Guo et al. (2021a); Li and Rahardja (2021); Wu et al. (2021a); Jiang et al. (2021); Wang et al. (2020a, 2022b); Liu et al. (2022a); Yang et al. (2022b)) , while the last one is the majority. We observe that the spatial-wise attention module is mostly inserted into the skip connection in the retinal vessel segmentation (as shown in Table 11) as the characteristic spatial information of different scales can help to better extract vessels due to the blurring of vessel boundary and the reflection of vessel centerline in fundus images (Liu et al. (2022a)). Especially, hard attention is first introduced into retinal vessel segmentation by Wang et al. (2020a) in HAnet, which is composed of one encoder and three decoder

Table 6: An overview of non-Transformer methods for lung segmentation.

Author	How	What	Datasets	Metric	Data split (Train:Val:Test)	Modality (Type)
Tang et al. (2019)	Enc	Spatial	1.JSRT (Shiraishi et al. (2000)) + Montgomery (Jaeger et al. (2014)) 2.NIH (Tang et al. (2019))	Dice:0.976 Dice:0.943	280:37:78 NaN:NaN:100	X-ray (3D)
Qin et al. (2020)	Dec	Spatial	LIDC (Armato III et al. (2011)) + EXACT'09 (Lo et al. (2012))	Branches detected:0.962 Tree-length detected:0.907 TPR:0.936 FPR:0.035 Dice:0.925	63:9:18	CT (2D)
Chen et al. (2020)	Skip.	Spatial	COVID-19 CT segmentation dataset (of Medical and Radiology)	Dice:0.83 ACC:0.79 P:0.82	90%:NaN:10%	CT (3D)
Zhou et al. (2021b)	Skip.	Spatial and Channel	COVID-19 CT segmentation datase (of Medical and Radiology) + Segmentation dataset nr. 2	Dice:0.831 HD:18.8	80%:NaN:20%	CT (3D)
Punn and Agarwal (2022a)	Skip.	Spatial	Segmentation dataset nr. 2 + COVID-19-CT-Seg dataset (Ma et al. (2021))	ACC:0.965 P:0.758 Dice:0.816 JI:0.791	70%:NaN:30%	CT (2D)
Yin et al. (2022)	Enc & Dec	Spatial and Channel	1.COVID-19 CT segmentation dataset + COVID-19-CT-Seg dataset (Ma et al. (2021)) 2.Segmentation dataset nr. 2	Dice:0.8696 ACC:0.9906 JI:0.7702 Dice:0.5936 ACC:0.9821 JI:0.4788	1373:196:391 258:38:77	CT (2D)

Table 7: An overview of non-Transformer methods for kidney tumor segmentation.

Author	How	What	Datasets	Metric	Data split (Train:Val:Test)	Modality (Type)
Myronenko and Hatamizadeh (2019)	Skip.	Spatial	KiTs (Heller et al. (2019))	Kidney Dice:0.9742 Tumor Dice:0.8103	210:NaN:90	CT (3D)
Sabarinathan et al. (2019)	Dec	Spatial	KiTs (Heller et al. (2019))	Kidney Dice:0.9535 Tumor Dice:0.8967	32175:13790:NaN	CT (2D)
Xuan et al. (2022)	Skip.	Spatial	KiTs (Heller et al. (2019))	Kidney Dice:0.961 IoU:0.926 HD:17.571 Tumor Dice:0.865 IoU:0.772 HD:33.839	134:34:42	CT (3D)

Table 8: An overview of non-Transformer methods for liver segmentation.

Author	How	What	Datasets	Metric	Data split (Train:Val:Test)	Modality (Type)
Chen et al. (2019b)	Skip.	Channel	LiTs (Bilic et al. (2019))	Tumor Dice:0.766	131:NaN:70	CT (2D)
Jiang et al. (2019a)	Skip.	Spatial	1.LiTs (Bilic et al. (2019))	NaN	110:NaN:NaN	CT (2D)
			2.3DIRACADb (Soler et al. (2010))	Liver Dice:0.959 Tumor Dice:0.734 MSD:6.271	NaN:NaN:20	
			3.Private dataset	Tumor Dice:0.591 MSD:7.538	NaN:NaN:117	
Jin et al. (2020)	Skip.	Spatial	1.LiTs (Bilic et al. (2019))	Liver Dice:0.961 JI:0.926 MSD:26.948 Tumor Dice:0.595 JI:0.611 MSD:6.775	130:NaN:70	CT (3D)
			2.3DIRACADb (Soler et al. (2010))	Liver: Dice:0.977 JI:0.977 MSD:18.617 Tumor: Dice:0.830 JI:0.744 MSD:53.324	NaN:NaN:20	
Li et al. (2020a)	Skip.	Spatial	LiTs (Bilic et al. (2019))	Dice:0.9815 P:0.98 R:0.99 IoU:0.9748	83%:NaN:17%	CT (2D)
Fan et al. (2020b)	Dec & Skip.	Spatial and Channel	LiTs (Bilic et al. (2019))	Liver Dice:0.960 Tumor Dice:0.749	90%:10%:70	CT (2D)
Yan et al. (2020)	Skip.	Spatial	1.Liver vessel segmentation (Yan et al. (2020))	Dice:0.805 P:0.789 SE:0.857	30:NaN:10	CT (3D)
			2.3DIRCADb (Soler et al. (2010))	Dice:0.904 P:0.990 SE:0.936	NaN:NaN:NaN	

sub-networks. Specifically, a basic decoder is expected to yield a coarse vessel segmentation result and provide a probabilistic map to automatically determine which region is "easy" or "hard" to segment. The attention mechanism is integrated into the "hard" decoder branch to effectively reinforce vessel features. HAnet achieves the highest segmentation accuracy and area under the receiver operating characteristic curve (AUC) on the public fundus datasets.

Skin Lesion. Due to the influence of color, boundaries, and shapes of melanoma as well as various artifacts, the segmentation of the skin lesion area is still a challenging problem (Kaul et al. (2019); Wei et al. (2019); Singh et al. (2019); Wu et al. (2020); Tong et al. (2021b); Ren et al. (2022); Arora et al. (2021); Wang and Wang (2022)). Here, we observe that spatial attention is the most commonly used mechanism in the skip connection for skin lesion segmentation, as shown in Table 12. This may be due to the importance of multi-scale information in handling skin lesions of varying sizes and shapes. By in-

Table 9: An overview of non-Transformer methods for polyp segmentation, in which mDice is the mean Dice.

Author	How	What	Datasets	Metric	Data split (Train:Val:Test)	Modality (Type)
Fan et al. (2020a)	Skip.	Spatial	1.ETIS (Silva et al. (2014))	mDice:0.628 mIoU:0.567 MAE:0.031	80%:10%:10%	Endoscopy (2D)
			2.CVC-ClinicDB (Bernal et al. (2015))	mDice:0.899 mIoU:0.849 MAE:0.009		
			3.CVC-ColonDB (Bernal et al. (2012))	mDice:0.709 mIoU:0.640 MAE:0.045		
			4.EndoScene (Vázquez et al. (2017))	mDice:0.871 mIoU:0.797 MAE:0.010		
			5.Kvasir (Jha et al. (2020))	mDice:0.898 mIoU:0.840 MAE:0.030		
Tomar et al. (2021)	Dec	Channel	Kvasir-SEG (Jha et al. (2020))	mDice:0.8576 mIoU:0.7800 SE:0.8880 P:0.8643	88%:NaN:12%	Endoscopy (2D)
Wei et al. (2021a)	Dec	Spatial	1.ETIS (Silva et al. (2014))	mDice:0.750 mIoU:0.654	80%:10%:10%	Endoscopy (2D)
			2.CVC-ClinicDB (Bernal et al. (2015))	mDice:0.916 mIoU:0.859		
			3.CVC-ColonDB (Bernal et al. (2012))	mDice:0.753 mIoU:0.670		
			4.EndoScene (Vázquez et al. (2017))	mDice:0.888 mIoU:0.815		
			5.Kvasir (Jha et al. (2020))	mDice:0.904 mIoU:0.847		
Kim et al. (2021)	Skip. & Dec	Spatial	1.ETIS (Silva et al. (2014))	mDice:0.766 mIoU:0.689 MAE:0.012	NaN:NaN:100%	Endoscopy (2D)
			2.CVC-ClinicDB (Bernal et al. (2015))	mDice:0.926 mIoU:0.880 MAE:0.006	90%:NaN:10%	
			3.CVC-ColonDB (Bernal et al. (2012))	mDice:0.783 mIoU:0.704 MAE:0.034	NaN:NaN:100%	
			4.CVC-300 (Vázquez et al. (2017))	mDice:0.910 mIoU:0.849 MAE:0.005	NaN:NaN:100%	
			5.Kvasir (Jha et al. (2020))	mDice:0.912 mIoU:0.859 MAE:0.025	90%:NaN:10%	

Table 10: An overview of non-Transformer methods for prostate segmentation, in which TZ is the prostatic transition zone, PZ is the peripheral zone and PB is the prostate bed.

Author	How	What	Datasets	Metric	Data split (Train:Val:Test)	Modality (Type)
Wang et al. (2019d)	Dec	Spatial	Private dataset	Dice:0.90 JI:0.82 HD95:8.37 P:0.90	75%:20%:NaN	Ultrasound (3D)
Liu et al. (2019)	Skip.	Spatial	1.PROSTATEX (Litjens et al. (2014a))	Dice:0.74 (PZ), 0.86 (TZ)	250:NaN:63	MRI (2D)
			2.Private dataset	Dice:0.74 (PZ), 0.79 (TZ)	NaN:NaN:46	
Kearney et al. (2019)	Skip.	Spatial	Private dataset	Dice:0.9002 (Prostate), 0.9312 (Bladder), 0.846 (Rectum), 0.7221 (Penile bulb)	80:20:20	CT (3D)
Jia et al. (2019)	Dec	Spatial	PROMISE12 (Litjens et al. (2014b))	Dice:0.9135 HD95:3.93 Score:90.34	50:NaN:30	MRI (3D)
Lei et al. (2020b)	Skip.	Spatial	Private dataset	Dice:0.91 HD:4.57 MSD:0.62	49:NaN:50	CT (3D)
Xu et al. (2020)	Skip.	Spatial	Private dataset	Dice:0.7567 (PB), 0.8840 (Bladder), 0.8035 (Rectum) ASD:0.42 (PB), 1.47 (Bladder), 2.60 (Rectum)	60%:20%:20%	CT (2D)

corporating spatial attention in the skip connection, the down-sampling path ensures the maximum flow of multi-scale information between layers, allowing for more accurate and effective segmentation of skin lesions, regardless of their size or shape. Besides, multi-resolution inputs are adopted to learn better specific and discriminative features in these applications (Singh et al. (2019); Wu et al. (2020); Wang and Wang (2022)). Based on that, Wu et al. (2020) propose an adaptive dual attention module, integrating global context and pixel-wise correlation, with different dilation rates for different sizes of skin lesions, and the model performs best on both ISBI2017 (Codella et al. (2018)) and ISIC2018 datasets (Tschandl et al. (2018)).

Surgical Instrument. Since robot-assisted surgery has gained increasing popularity, segmentation for tracking instruments has attracted more attention. However, the geometry change due to the pose changing, the partial occlusion caused by the narrow field of view, the specular reflection, and the serious class imbalance obstacles the automatic surgical instrument segmentation.

In the early works, the attention mechanism is introduced into the ResNet with channel/spatial attention Ni et al. (2019a,b). Then the complementary attention is integrated into the network Ni et al. (2022), in which a position attention block and a channel attention block are combined into a double attention module. Moreover, the inherent temporal clues from the instrument motion are also taken into account (Jin et al. (2019)). Specifically, the prediction mask of the previous frame can be regarded as temporal prior to the instrument's current location

Table 11: An overview of non-Transformer methods for retinal vessel segmentation.

Author	How	What	Datasets	Metric	Data split (Train:Val:Test)	Modality (Type)
Zhang et al. (2019b)	Skip.	Spatial	1.DRIVE (Staal et al. (2004))	ACC:0.9692 AUC:0.9856 IoU:0.6965	20:NaN:20	Retinal fundus (2D)
			2.CHASEDB1 (Owen et al. (2009))	ACC:0.9743 AUC:0.9863 IoU:0.6669	NaN:NaN:100%	
			3.ORIGA (Zhang et al. (2010))	OE:0.061 (Diac), 0.212 (Cup), 0.137 (Total)	325:NaN:325	
Li et al. (2019c)	Skip.	Spatial	1.DRIVE (Staal et al. (2004))	AUC:0.9807 ACC:0.9560	20:NaN:20	Retinal fundus (2D)
			2.STARE (Hoover et al. (2000))	AUC:0.9834 ACC:0.9673	10:NaN:10	
			3.HRF (Budai et al. (2013))	AUC:0.9867	15:NaN:30	
Mou et al. (2019)	Skip.	Spatial and Channel	1.DRIVE (Staal et al. (2004))	ACC:0.9632 SE:0.8215 AUC:0.9825	80%:NaN:20%	Retinal fundus (2D)
			2.STARE (Hoover et al. (2000))	ACC:0.9752 SE:0.8816 AUC:0.9932	75%:NaN:25%	Retinal fundus (2D)
			3.Private dataset	ACC:0.9183 SE:0.8631 AUC:0.9453	80%:NaN:20%	OCT-A (2D)
			4.CMM-1 (Intelligent Medical Imaging Research Group)	SE:0.8415 FDR:0.2521	80%:NaN:20%	Microscopy (2D)
			5.CMM-2 (Intelligent Medical Imaging Research Group)	SE:0.8345 FDR:0.2591	80%:NaN:20%	Microscopy (2D)
Li et al. (2020e)	Dec	Spatial	1.DRIVE (Staal et al. (2004))	ACC:0.9568 AUC:0.9806	20:NaN:20	Retinal fundus (2D)
			2.STARE (Hoover et al. (2000))	ACC:0.9678 AUC:0.9678	19:NaN:1	
			3.CHASEDB1 (Owen et al. (2009))	ACC:0.9635 AUC:0.9702	20:NaN:8	
			4.IOSTAR (Zhang et al. (2016))	ACC:0.9544 AUC:0.9623	20:NaN:10	
			5.RC-SLO (Abbasi-Sureshjani et al. (2015))	AUC:0.9696 AUC:0.8119	30:NaN:10	
Guo et al. (2021b)	Enc.	Spatial	1.DRIVE (Staal et al. (2004))	ACC:0.9698 AUC:0.9864	20:NaN:20	Retinal fundus (2D)
			2.CHASEDB1 (Owen et al. (2009))	ACC:0.9755 AUC:0.9905	20:NaN:8	
Lv et al. (2020)	Enc.	Spatial	1.DRIVE (Staal et al. (2004))	ACC:0.9558 AUC:0.9847	20:NaN:20	Retinal fundus (2D)
			2.STARE (Hoover et al. (2000))	ACC:0.9640 AUC:0.9824	10:NaN:10	
			3.CHASEDB1 (Owen et al. (2009))	ACC:0.9608 AUC:0.9865	20:NaN:8	
Li et al. (2020c)	Dec	Spatial	1.DRIVE (Staal et al. (2004))	ACC:0.9769 AUC:0.9895	20:NaN:20	Retinal fundus (2D)
			2.STARE (Hoover et al. (2000))	ACC:0.9797 AUC:0.9924	10:NaN:10	
			3.CHASEDB1 (Owen et al. (2009))	ACC:0.9803 AUC:0.9912	14:NaN:14	
Guo et al. (2021a)	Skip.	Channel	1.DRIVE (Staal et al. (2004))	ACC:0.9699 AUC:0.9852	20:NaN:20	Retinal fundus (2D)
			2.STARE (Hoover et al. (2000))	ACC:0.9743 AUC:0.9911	15:NaN:15	
			3.CHASEDB1 (Owen et al. (2009))	ACC:0.9751 AUC:0.9898	20:NaN:8	
Wang et al. (2020a)	Skip.	Spatial	1.DRIVE (Staal et al. (2004))	AUC:0.9823 ACC:0.9581	20:NaN:20	Retinal fundus (2D)
			2.STARE (Hoover et al. (2000))	AUC:0.9881 ACC:0.9673	19:NaN:1	
			3.CHASEDB1 (Owen et al. (2009))	AUC:0.9871 ACC:0.9670	20:NaN:8	
			4.HRF (Budai et al. (2013))	AUC:0.9837 ACC:0.9654	15:NaN:30	
Liu et al. (2022a)	Skip.	Spatial and Channel	1.DRIVE (Staal et al. (2004))	ACC:0.9699	20:NaN:20	Retinal fundus (2D)
			2.CHASEDB1 (Owen et al. (2009))	ACC:0.9751	14:NaN:14	

Table 12: An overview of non-Transformer methods for skin lesion segmentation.

Author	How	What	Datasets	Metric	Data split (Train:Val:Test)	Modality (Type)
Wei et al. (2019)	Skip.	Spatial	1.ISBI 2016 (Gutman et al. (2016))	ACC:0.9683 Dice:0.9536 JI:0.9142	900:NaN:379	Dermoscopy(2D)
			2.ISBI 2017 (Codella et al. (2018))	ACC:0.9329 Dice:0.8786 JI:0.8045	2000:150:600	
			3.PH2 (Mendonça et al. (2013))	Divergence Value:8.23	NaN:NaN:200	
Wu et al. (2020)	Skip.	Spatial and Channel	1.ISBI 2017 (Codella et al. (2018))	ACC:0.9570 Dice:0.8969 JI:0.8255	2000:150:600	Dermoscopy (2D)
			2.ISIC 2018 (Tschandl et al. (2018))	ACC:0.9470 Dice:0.9080 JI:0.8440	2000:594:NaN	
Arora et al. (2021)	Skip.	Spatial	ISIC 2018 (Tschandl et al. (2018))	ACC:0.95 Dice:0.91 JI:0.83	2000:594:NaN	Dermoscopy (2D)
Tong et al. (2021b)	Dec	Spatial and Channel	1.ISIC 2016 (Gutman et al. (2016))	ACC:0.9540 JI:0.8450	2000:150:600	Dermoscopy (2D)
			2.ISIC 2017 (Codella et al. (2018))	ACC:0.9260 JI:0.7420	900:379:NaN	
			3.PH2 (Mendonça et al. (2013))	ACC:0.9430 JI:0.8420	NaN:NaN:200	

and shape, thus integrated into the segmentation network as an initial attention gate signal. Existing medical surgical instrument dataset includes Cata7 (Ni et al. (2019b)), CataIS (Ni et al. (2019a)), EndoVis 2017 (Allan et al. (2019)), and ROBUST-MIS dataset (Maier-Hein et al. (2021)), in which Cata7 and CataIS are constructed by Ni et al. (2019a,b) for cataract surgical instrument segmentation from Beijing Tongren Hospital and will be public soon. EndoVis 2017 (Allan et al. (2019)) is from the MICCAI Endovis Challenge 2017, which is based on endoscopic surgery with 3000 images and ROBUST-MIS dataset is

comprised of a total of 10,040 annotated video frames from 30 minimally invasive daily-routine surgical procedures.

Conclusion. In short, attention mechanisms have been applied to various aspects of medical image segmentation, and the number of papers depends on the number of publicly available datasets in this domain. We also observe that spatial attention is more popular than channel attention, this may be because edge information is of significance for blurring boundaries (polyp segmentation, prostate segmentation and skin lesion segmentation. etc.) and there is a strong connection between lesions/tumors and the corresponding organs (e.g., the lung lesion segmentation, kidney tumor segmentation, liver tumor segmentation, etc.). What needs to be emphasized is that very few researchers design the attention module for organ-specific.

4. Transformer in medical Segmentation

To provide a comprehensive understanding of Transformer-based methods in medical segmentation, we have structured our survey to follow the format of the Non-Transformer part. We first introduce the core concept in Transformer (what to use), followed by the mainstream network architecture (how to use), and the specific application tasks (where to use).

4.1. What to Use

The basic Transformer layer (Vaswani et al. (2017)) comprises two main sub-layers: the multi-head self-attention and feedforward layers. Self-attention mechanism employs scaled dot-product attention to model interactions between all elements of a sequence, given by equation (2). The multi-head attention mechanism captures complex relationships between entities by allowing attention layers to focus on different representation subspaces. Given an input vector and the number of heads h, the input vector is transformed into three separate groups of vectors, each containing h vectors. The following equation can describe this process:

$$\begin{aligned} &Q = \{Q_i\}_{i=1}^{h}, K = \{K_i\}_{i=1}^{h}, V = \{V_i\}_{i=1}^{h} \\ &\text{head}_i = \text{Attention}(Q_i, K_i, V_i) \\ &\text{MultiHead}(Q, K, V) = Concat(\text{head}_1, \cdots, \text{head}_h)W^O \end{aligned} \tag{3}$$

where $W^O \in \mathbb{R}^{d_{model} \times d_{model}}$ is the linear projection matrix.

The feedforward layer applies a two-layer feedforward neural network to the attended input. It consists of two linear transformations with a ReLU activation function in between. This allows the model to transform the attended input into a different space and learn more complex representations. In addition to these two sub-layers, the Transformer layer includes residual connections and layer normalization. The residual connections allow the model to learn identity mapping and mitigate the vanishing gradient problem. Layer normalization helps to stabilize the training process by normalizing the inputs to each layer.

The Transformer layer can be stacked multiple times to form the encoder and decoder networks in the Transformer architecture. The encoder takes an input sequence and produces a sequence of hidden representations, while the decoder takes a sequence of hidden representations and produces an output sequence.

Some successful Transformer-based works have been proposed in computer vision. **Vision Transformer (ViT**, Dosovitskiy et al. (2020)) is the first model to introduce the Transformer encoder into computer vision for image classification, which applies the Transformer model on a sequence of image patches flattened as vectors directly with a 1D learnable positional encoding. ViT inserts a learned *[class]* embedding whose state at the output of the Transformer encoder serves as a representation to perform classification. ViT was shown to achieve state-of-the-art performance on the ImageNet benchmark dataset, demonstrating the effectiveness of the Transformer architecture for computer vision tasks. **DEtection TRansformer (DETR**, Carion et al. (2020)) is a Transformer-based model for object detection that treats object detection as a direct set prediction problem. It uses an encoder-decoder architecture based on the Transformer to predict all objects simultaneously. The Transformer module outputs N objects' embeddings in parallel after a CNN backbone for N final predictions (including box coordinates and class labels). **Deformable DETR** (Zhu et al. (2020)) is a variation of DETR that proposes the deformable attention module to mitigate the high computa-

tional cost issues in DETR. Deformable attention attends to a sparse set of elements from the whole feature map regardless of all elements. **SEgmentation TRansformer (SETR**, Zheng et al. (2021)) attempts to migrate the Transformer to image segmentation tasks, with a pure Transformer encoder and a CNN-based decoder. Three fashions of the encoder are employed for pixel-wise classification: naive up-sampling, progressive up-sampling, and multi-level feature aggregation for exploration. **Pyramid Vision Transformer (PVT**, Wang et al. (2021g)) introduces a progressive shrinking pyramid with pure Transformer block and adopts a spatial-reduction attention layer to reduce the computational costs in dense prediction tasks. Unlike the above transformer-based models that operate on fixed-sized image patches, **Swin Transformer** (Liu et al. (2021c)) introduces the shifted window operation to efficiently capture global image features while preserving fine-grained spatial information. Swin Transformer also incorporates several novel techniques, such as shifted window attention and local feature aggregation. With its hierarchical design, Swin Transformer has achieved impressive results on several benchmark datasets, demonstrating its effectiveness for various vision tasks.

In medical image segmentation tasks, Transformer and its improved versions (*e.g.*, Swin Transformer) are usually introduced into the U-Net architecture as a plug-in module or the basic block.

4.2. How to Use

Despite Transformer's ability to model the global contextual dependency, the self-attention induces missing inductive bias of locality (Dosovitskiy et al. (2020)). Meanwhile, convolution block with locality advances to deal with these features with preferable inductive bias (Simoncelli and Olshausen (2001)) (*e.g.*, translation invariance (Scherer et al. (2010))). Thus Transformer-based methods are more likely to combine with CNN by leveraging the locality of CNNs and the long-range dependency character of the Transformer for medical image segmentation with limited data. Here, we observe that Transformer-based models can be categorized by the main architecture design as shown in Fig 6, consisting of Hybrid encoder + CNN decoder, Pure Transformer encoder + CNN decoder, CNN encoder + Pure Transformer decoder, and Transformer encoder + Transformer decoder.

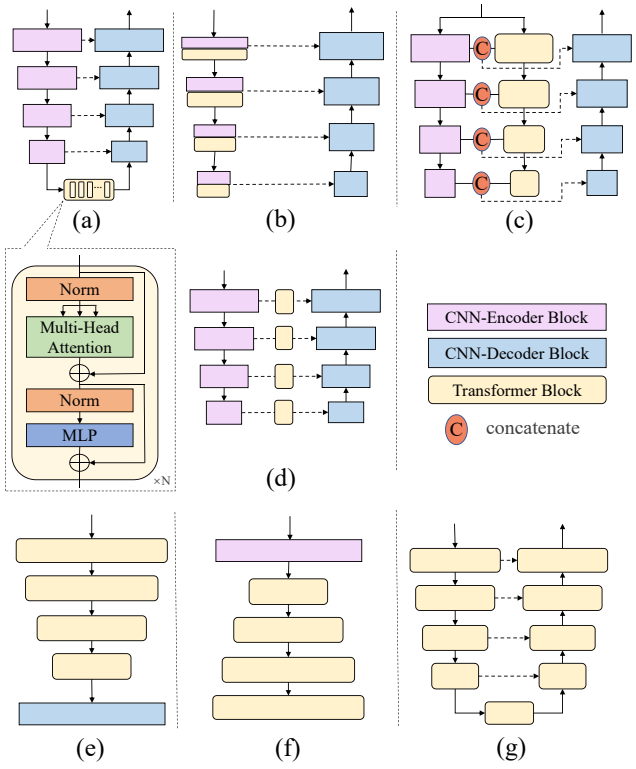

Fig. 6: Comparison of different Transformer-based architectures for medical image segmentation. The first two rows represent Hybrid encoder + CNN decoder cases: (a) Transformer used as the bottleneck, (b) Transformer combined with CNN in serial as encoder, (c) Transformer combined with CNN in parallel as a dual-path encoder, and (d) Transformer in the skip connection. The last row represents cases where Transformer is regarded as the main body in the encoder/decoder: (e) Transformer encoder + CNN decoder, where Transformer stacks are in pyramid style in the encoder, (f) CNN encoder + Transformer decoder, with Transformer stacks in the decoder, and (g) Transformer encoder + Transformer decoder. Please note that the decoder in (e) and the encoder in (f) are simplified due to limited space.

4.2.1. Hybrid encoder + CNN decoder

U-Net and its modified versions based on CNN have been widely used in medical image segmentation due to their efficient and precise feature extraction capabilities. To further improve the performance of U-Net and its variants, researchers have explored integrating the Transformer into these architectures to improve the long-term modeling ability. This is typically done by adding Transformer in the encoder to form a hybrid encoder leveraging both benefits of Transformer and CNN. Here we categorize these methods into four groups according to Transformer's location and proportion in the U-Net style network.

Transformer as the bottleneck. A most natural and source-saving form is introducing the Transformer as the bottleneck in the encoder (shown in Fig6 (a)). Chen et al. (2021b) propose TransUnet, which integrates ViT into the U-Net architecture to capture long-term dependencies in the CNN. TransUnet has shown promising results on 2D medical images and has inspired further research in integrating Transformers into U-Net for medical image segmentation. For multi-task learning, Cheng et al. (2022a) follow the TransUNet design for glioma segmentation and combine the features from CNN and Transformer for isocitrate dehydrogenase (IDH) genotyping. Zhang et al. (2021c) implement a vanilla Transformer encoder-decoder as a bottleneck module within their approach. After the U-Net processing, the resulting feature is divided into two branches: body and edge. This division promotes local consistency and provides edge position information. Wang et al. (2021c) also include the Transformer encoder-decoder as a bottleneck but introduce boundary-wise prior knowledge for guidance, enhancing the model's performance in tasks that require a focus on boundaries. To reduce the information loss during the upsampling operations, Yang and Tian (2022) add CBAM (Woo et al. (2018)) in upsampling operations to capture the region of interest. Chen et al. (2021a) design a novel Transformer module as the bottleneck. It consists of a Transformer Self-Attention module to jointly attend to semantic information from the global representation subspace and a Global Spatial Attention module to selectively aggregate global context to the learned features. Chang et al. (2021) insert the Transformer block in the bottleneck and introduce multi-scale properties by upsampling the Transformer's feature to different scales and concatenating them with each layer feature from the CNN encoder. LeViT-UNet (Xu et al. (2021b)) uses LeViT (Graham et al. (2021)) instead of the vanilla Transformer to reduce the computational cost. Meanwhile, Guo and Terzopoulos (2021) and Yan et al. (2022) try to take 3D information into account in the 2D network with the help of the Transformer bottleneck block, specifically, the former encodes dependencies between slices along the z-axis for the 3D anisotropy problem and the latter proposes an axial Transformer to leverage intra- and inter-slice contextual information. TransBTS (Wang et al. (2021f)) improve a 3D version of TransUNet (Chen et al. (2021b)). After that, TransBTSV2 version (Li et al. (2022a)) inserts an expansion module to expand the Transformer width. To boost the performance of TransBTS, Jia and Shu (2021) prefer to add another Transformer layer into the penultimate layer to get dense information, while Dobko et al. (2021) add SE blocks (Hu et al. (2018b)) to each layer of the encoder block in TransBTS (Wang et al. (2021f)), meanwhile, the positional encoding is replaced with a learnable MLP block.

Cross-scale dependency and consistency are crucial in segmenting objects, such as lesions, that experience significant size changes. To address this challenge, researchers have proposed various methods that involve concatenating encoded features of different scales and sending them to a Transformer block within the bottleneck. CoTr (Xie et al. (2021c)) applies an efficient deformable Transformer on the concatenated multi-layer feature to model multi-scale contextual features, paying attention to only a small set of key positions to reduce the computational and spatial complexities. Wang et al. (2021e) design a novel Transformer module and place it at the top of the encoder path, aiming to capture multi-scale non-local features with long-range dependencies from different layers of the encoder. UCTransUNet (Wang et al. (2021a)) designs a channel-wise cross fusion Transformer to capture local-channel interaction and solve the inconsistent semantic level problem, which replaces the simple skip connection. In this method, the concatenated features from the encoder act as the key and value, and the feature from each layer is the query in the cross fusion Transformer. Ji et al. (2021) embed the multi-scale convolutional features as a sequence of tokens and perform Transformer self-attention and cross-attention sequentially to capture the cross-scale dependencies. Additionally, they introduce a learnable proxy embedding to model semantic relationships and use it as the query in the cross-attention module.

CNN-Transformer in Serial as Encoder. Incorporating Transformers only as a bottleneck may not effectively cap-

ture long-term dependencies, thus researchers have introduced Transformer modules into more network stages, enhancing complementary benefits and improving performance.

As shown in Fig 6 (b), Transformers can be introduced after each convolutional layer to build long-range dependency on features of various scales. UTNet (Gao et al. (2021b)) adds a Transformer layer after each convolutional operation in the encoder and designs an efficient self-attention mechanism along with relative position encoding to reduce complexity. SpecTr (Yun et al. (2021)) employs a similar architecture, but with a sparsity scheme and spectral normalization strategy in the Transformer block for hyperspectral pathology image segmentation. Wang et al. (2021b) propose a Transformer with a SE operation at each scale and design a CCA block consisting of GCNet (Cao et al. (2019)) and ECA-Net (Wang et al. (2020c)) to optimize the model by combining both spatial and channel attention in the skip connection. RTNet (Huang et al. (2022)) introduces a global Transformer and a relation Transformer to capture inter- and intra-information between vessel and lesion features for retinopathy multi-lesion segmentation. Additionally, Wang et al. (2021d) integrate CNN-Transformer information through concatenation, rather than operating in sequence, and propose a 3D image position embedding that allows the local neighbourhood to facilitate role in global attention within multi-head self-attention.

CNN-Transformer in Dual-Path as Encoder. Some works adopt a dual-path encoder, combining Transformers and CNNs, to simultaneously learn global and local dependencies, as shown in Fig 6 (c). Sun et al. (2021) propose a dual-path CNN-Transformer encoder and concatenate the final outputs of both at the network's bottom as the decoder's input. Wu et al. (2022a) extend this model by adding a SE block (Hu et al. (2018b)) on the fused feature for channel information. Sha et al. (2021) also employ a dual-path encoder and concatenate feature maps from both encoders at each layer. To further combine the information from each layer, Zhang et al. (2021e) design three axial Transformer branches in different input scales and a CNN branch, integrating them at three levels to grasp associations of diverse scales within the image. Similarly, Zhang et al. (2021d) report Transfuse, where features with the same resolution are fused through self-attention and bilinear Hadamard product, acting as signal gates between a CNN and a Transformer branch. Besides, Reza et al. (2022) present a different fusion method, adding image-level contextual representation and regional importance coefficients from Transformers to CNNs for spatial normalization.

Indeed, the dual-path framework offers time-saving benefits, while the serial framework is more space-saving. However, there has not been a comprehensive experiment or study that thoroughly compares these two structures in terms of their advantages, disadvantages, time efficiency, local/global information interaction, and potential information loss. Such a study would provide valuable insights into the most effective approach for specific applications and guide future research in the medical image segmentation domain.

Transformer in the Skip Connection. Some approaches utilize Transformers in the skip connection to bring global information into CNN architectures, as shown in Fig 6 (d). Yu et al. (2022) add channel attention vision transformer (CAViT) in the skip connection, combining Transformers and ECA (Wang et al. (2020c)) blocks to leverage both channel attention and self-attention. They also introduce a deep adaptive gamma correlation to improve retinal segmentation. Petit et al. (2021) apply a multi-head cross-attention in the skip connection to filter non-semantic richness features as well as a multi-head self-attention at the top of U-Net encoder to leverage global interactions between semantic features. You et al. (2022) build a GAN to enhance performances, incorporating a class-aware Transformer module in the skip connection to learn regions of interest in the generator progressively. They also devise a ResNet-Transformer discriminator for improved performance.

4.2.2. Pure Transformer encoder + CNN decoder

To take advantage of the U-Net architecture design, Transformer layers can be stacked in a pyramid fashion inspired by PVT (Wang et al. (2021g)), as shown in the last row of Fig 6. The building blocks include multilayer perceptrons (MLP),

multi-head self-attention layers (MSA), and appropriate down-sampling layers. Thus they can learn hierarchical object concepts at different resolutions.

As an example, those methods introduce Transformer only into the encoder to replace the convolution block. Karimi et al. (2021) and UNETR (Hatamizadeh et al. (2022b)) applies ViT-like Transformer backbone as a substitute for the convolution block in the encoder part in a U-Net-like structure. Swin UNETR Hatamizadeh et al. (2022a) employs Swin Transformer (Liu et al. (2021c)) to reduce computational costs. Zhu et al. (2021) also use SegFormer encoder (Xie et al. (2021a)) as the backbone, adding region prior information for more accurate breast ultrasound tumor segmentation.

4.2.3. CNN encoder + Pure Transformer decoder

There are limited papers involving Transformers only in Decoder as (f) in Fig 6, Li et al. (2021a) apply a pre-trained ResNet encoder to extract features and propose a squeeze-and-excitation Transformer in the decoder to learn the attention matrix. And Gong et al. (2022) propose a PVT-style decoder to interconnect multi-resolution CNN encoder seamlessly and evaluate methods on various datasets, including ISIC2018 (Tschandl et al. (2018)) and CVC-ClinicDB (Bernal et al. (2015)). Li et al. (2022d) proposes a window attention-up-sampling with Transformer to connect different resolution features through self-attention.

4.2.4. Transformer encoder + Transformer decoder

Recent works try to introduce Unet-like Pure Transformer, in which Transformer blocks are the basic blocks in both encoder and decoder parts (in Fig 6 (g)). Sagar (2021) adopt the UNet-like Transformer encoder-decoder architecture, and the Transformer input is a multi-scale feature encoded by a three-branch convolutional block with different kernel sizes. Cao et al. (2021) and Wu et al. (2021b) use Swin Transformer as the basic block of U-Net-like networks. The latter proposes the 3D Swin Transformer and introduces inductive bias, namely region prior. Inspired by the shifted window design in the Swin Transformer, Peiris et al. (2021) propose a hierarchical Transformer encoder-decoder network. Especially they design a decoder block that enables parallel window-based self- and cross-attention to capture details for boundary refinement. A convex combination approach and Fourier position encoding are also added to inject complementary information. Wu et al. (2022b) design a dilated Transformer as the basic block, which conducts self-attention for pair-wise patch relations captured alternatively in both local and global scopes. Li et al. (2021b) utilize the group Transformer (*i.e.*, a grouping architecture) to reduce computational costs and design a shape-sensitive Fourier Descriptor loss function for tooth root segmentation. Zhou et al. (2021a) propose nnFormer, which jointly uses local- and global-volume-based attention at different layers to construct feature pyramids. Skip attention replaces the traditional concatenation operation in the skip connection to improve the results.

Few works also consider merging multi-scale information in medical image segmentation. Lin et al. (2021) improve the Swin Transformer-based U-Net architecture by adapting dual encoder subnets under different input sizes, allowing for the extraction of coarse and fine-grained representations separately. They also incorporate a Transformer Interactive Fusion module to aggregate cues between the encoder subnets for fusing multi-scale information. Huang et al. (2021) propose an Enhanced Transformer-based U-Net, incorporating an Enhanced Transformer Context Bridge to combine all-level features instead of the traditional skip connection. They also introduce an efficient self-attention mechanism for spatial reduction to reduce computational costs.

We observe that these works with pure Transformer as the encoder/decoder tries to introduce the benefits of UNet-like architecture into the Transformer and the convolutional operations are inevitable in the up-/down-sampling or the patch embedding process. However, two main challenges exist. Firstly, the computational complexity of the Transformer is high, and the stacked pyramid framework exacerbates this issue. Therefore, various efficient self-attention computation methods(*e.g.*, Cao et al. (2021); Wu et al. (2021b); Peiris et al. (2021); Zhou et al. (2021a); Huang et al. (2021); Li et al. (2021b)) have been intro-

duced. Secondly, Transformer requires a large amount of data due to its lack of inductive bias(Dosovitskiy et al. (2020)), and labeled medical images are scarce compared to labeled natural images. Most methods rely on transferring learning via ImageNet pretraining. Wu et al. (2021b) has introduced inductive bias, and Tang et al. (2022); Xie et al. (2022) have explored self-supervised pretraining strategies in the medical image domain.

4.3. Where to use (Applications)

The Transformer-based medical applications are commonly used in multi-organ segmentation, cardiac diagnosis, polyp detection, brain tumor segmentation, and retinal segmentation tasks. Among these tasks, the universal Transformer-based medical segmentation models are often evaluated on the multi-organ segmentation task using the BCV dataset (Landman et al. (2015)), as shown in Table 13. Our observation shows that the hybrid encoder-CNN decoder framework performs well on 2D datasets, while the methods with Transformer encoder-Transformer decoder perform better on 3D datasets. Moreover, part of the universal approaches is also evaluated on the cardiac diagnosis task using the ACDC dataset (Bernard et al. (2018)) in Table 14.

Brain Tumor. Automated and accurate segmentation of brain tumors plays an essential role in the timely diagnosis of neurological diseases, and Transformer-based methods have been proposed for these tasks effectively. Jun et al. (2021) firstly introduce Transformer into the brain tumor segmentation task with a ViT-style pure Transformer encoder. Later, TransBTS (Wang et al. (2021f)) is proposed to take advantage of both CNN and Transformer in the local and global feature extraction, which inserts a Transformer module in the bottleneck of U-Net framework and inspires some following works on improving it(*e.g.*, Jia and Shu (2021); Dobko et al. (2021); Li et al. (2022a); Hatamizadeh et al. (2022a)). To further apply Transformer at each stage, Liang et al. (2022b) combine the convolution and Transformer blocks in parallel and integrate the two feature maps by a cross-attention fusion as an encoder basic block. Moreover, Swin Transformer-based encoder architectures (Liang et al. (2022a)) and Swin-Transformer based encoder-decoder architectures (Jiang et al. (2022b); Liang et al. (2022c); Peiris et al. (2022)) are designed for brain tumor segmentation to reduce the computational costs. It is noticed that multi-modality fusion is essential for precise brain tumor segmentation from MRI, thus Li et al. (2022c); Zhang et al. (2022b) and Xing et al. (2022) adopt multi-branch encoder for different modalities separately and the Transformer module act as the bottleneck between the encoder and the decoder. The proposed Transformer module consists of a self-attention module to enhance long-term dependencies within individual modalities and a cross-attention block to catch cross-modality contextual information. These methods are all experimented on the BraTS dataset in Table 15.

Polyp. Accurate polyp segmentation is a challenge due to the variable size and shape of polyps, as well as the indistinct boundaries between polyps and mucosa. Hence, local information is essential and the Transformer block is always combined with the convolutional block in the polyp segmentation task (as shown in Table 16). Tomar et al. (2022) combine a Swin Transformer block and a dilated convolutional block as a basic module at the bottleneck of U-Net. Dong et al. (2021); Wang et al. (2022c); Sanderson and Matuszewski (2022); Park and Lee (2022), and Duc et al. (2022) introduce PVT-like network to the polyp segmentation task. It is noted that a progressive decoder is also proposed for improved local emphasis and stepwise feature aggregation in works(*e.g.,* Wang et al. (2022c); Sanderson and Matuszewski (2022)).

Retinal. The applications of transformer in eye image include fovea localization (Song et al. (2022)), iris segmentation (Wei et al. (2021b)), retinal vessel segmentation (Chen et al. (2022); Jiang et al. (2022a)), and retinopathy lesion segmentation (Huang et al. (2022)). The network architecture is various and each of them makes some special designs. Song et al. (2022) propose a bi-branch CNN encoder, which consists of a main branch for retinal images and a vessel branch for pre-segmentation of vessel images. A Transformer block is used as the bottleneck in the main branch, similar to Tran-

Table 13: An overview of Transformer methods for multi-organ segmentation on the BCV dataset.

Method	Arch.	Type	Data split (Train:Val:Test)	Metric	
				Dice	HD
TransUNet (Chen et al. (2021b))	Hybrid encoder + CNN decoder	2D	18:NaN:12 (8 Organs Avg.)	0.7748	31.69
Swin-UNet (Cao et al. (2021))	Transformer encoder + Transformer decoder	2D	18:NaN:12 (8 Organs Avg.)	0.7913	21.55
LeViT-UNet (Xu et al. (2021b))	Hybrid encoder + CNN decoder	2D	18:NaN:12 (8 Organs Avg.)	0.7853	16.84
MISSFormer (Huang et al. (2021))	Transformer encoder + Transformer decoder	2D	18:NaN:12 (8 Organs Avg.)	0.8196	18.2
TransClaw U-Net (Chang et al. (2021))	Hybrid encoder + CNN decoder	2D	18:NaN:12 (8 Organs Avg.)	0.7809	26.38
LiteTrans (Xu and Quan (2021))	Transformer encoder + Transformer decoder	2D	18:NaN:12 (8 Organs Avg.)	0.7791	29.01
ViTBIS (Sagar (2021))	Transformer encoder + Transformer decoder	2D	20:NaN:10 (8 Organs Avg.)	0.8045	21.24
ViTBIS (Wang et al. (2022a))	Hybrid encoder + CNN decoder	2D	18:NaN:12 (8 Organs Avg.)	0.7859	26.59
CA-GANformer (You et al. (2022))	Hybrid encoder + CNN decoder	2D	18:NaN:12 (8 Organs Avg.)	0.8255	22.73
DSTUNet (Cai et al. (2022))	Hybrid encoder + CNN decoder	2D	18:NaN:12 (8 Organs Avg.)	0.8244	17.83
CS-Unet (Liu et al. (2022b))	Transformer encoder + Transformer decoder	2D	18:NaN:12 (8 Organs Avg.)	0.8221	27.02
DAE-Former (Azad et al. (2022))	Transformer encoder + Transformer decoder	2D	18:NaN:12 (8 Organs Avg.)	0.8243	17.46
TransCeption (Azad et al. (2023a))	Transformer encoder + Transformer decoder	2D	18:NaN:12 (8 Organs Avg.)	0.8224	20.89
FCT (Tragakis et al. (2023))	Transformer encoder + Transformer decoder	2D	18:NaN:12 (8 Organs Avg.)	0.8353	NaN
Cascaded MERIT (Rahman and Marculescu (2023))	Hybrid encoder + CNN decoder	2D	18:NaN:12 (8 Organs Avg.)	0.8490	13.22
CoTr (Xie et al. (2021c))	Hybrid encoder + CNN decoder	3D	21:NaN:9 (13 Organs Avg.)	0.8500	4.01
UNETR (Hatamizadeh et al. (2022b))	Pure Transformer encoder + CNN decoder	3D	30 (Standard) / 80 (Free): NaN:BCV test set	0.8560 (Standard), 0.8910 (Free)	NaN
AFTer-UNet (Yan et al. (2022))	Hybrid encoder + CNN decoder	3D	18:NaN:12 (8 Organs Avg.)	0.8102	NaN
UTNetV2 (Gao et al. (2022))	Transformer encoder + Transformer Decoder	3D	80%:20%:NaN (13 Organs Avg.)	0.8514	15.78
D-Former (Wu et al. (2022b))	Transformer encoder + Transformer decoder	3D	18:NaN:12 (8 Organs Avg.)	0.8883	NaN
nnFormer (Zhou et al. (2021a))	Transformer encoder + Transformer decoder	3D	18:NaN:12 (8 Organs Avg.)	0.8657	10.63
FINE (Themyr et al. (2022))	Transformer encoder + Transformer decoder	3D	18:NaN:12 (7 Organs Avg.)	0.8710	9.2
UNETR++ (Shaker et al. (2022))	Transformer encoder + Transformer decoder	3D	18:NaN:12 (8 Organs Avg.)	0.8722	7.53

Table 14: An overview of Transformer methods on cardiac segmentation on the ACDC dataset, in which RV is the right ventricle, LV means the left ventricle and the Myo is the myocardium. Ave. is the average dice.

Method	Arch.	Type	Data split(Train: Val: Test)	Metric			
				RV	Myo	LV	Ave.
TransUNet (Chen et al. (2021b))	Hybrid encoder + CNN decoder	2D	70: 10: 20	0.8886	0.8453	0.9573	0.8971
Swin-UNet (Cao et al. (2021))	Transformer encoder + Transformer decoder	2D	70: 10: 20	0.8855	0.8562	0.9583	0.9000
LeViT-UNet (Xu et al. (2021b))	Hybrid encoder + CNN decoder	2D	80: NaN: 20	0.8955	0.8764	0.9376	0.9032
MISSFormer (Huang et al. (2021))	Transformer encoder + Transformer decoder	2D	70: 10: 20	0.8955	0.8804	0.9499	0.9086
LiteTrans (Xu and Quan (2021))	Transformer encoder + Transformer decoder	2D	80: 20: NaN	0.8966	0.8797	0.8533	0.8966
DSTUNet (Cai et al. (2022))	Hybrid encoder + CNN decoder	2D	70: 10: 20	0.8036	0.8177	0.8834	0.8350
CS-Unet (Liu et al. (2022b))	Transformer encoder + Transformer decoder	2D	70: 10: 20	0.8920	0.8947	0.9542	0.9137
FCT (Tragakis et al. (2023))	Transformer encoder + Transformer decoder	2D	70: 10: 20	0.9264	0.9051	0.9550	0.9302
Cascaded MERIT (Rahman and Marculescu (2023))	Hybrid encoder + CNN decoder	2D	70: 10: 20	0.9023	0.8953	0.9580	0.9185
D-Former (Wu et al. (2022b))	Transformer encoder + Transformer decoder	3D	70: 10: 20	0.9133	0.8960	0.9593	0.9229
nnFormer (Zhou et al. (2021a))	Transformer encoder + Transformer decoder	3D	70: 10: 20	0.9094	0.8958	0.9565	0.9206
UNETR++ (Shaker et al. (2022))	Transformer encoder + Transformer decoder	3D	70: 10: 20	0.9189	0.9061	0.9600	0.9283

sUnet (Chen et al. (2021b)), to remove vessel interference and improve fovea localization. Wei et al. (2021b) design a bilateral self-attention module to apply spatial and visual branches to learn contextual clues for two characteristics in the iris segmentation and build a Transformer encoder-decoder architecture. While Huang et al. (2022) also propose a relation Transformer block to catch relationships between the lesions and vessel features. Moreover, Chen et al. (2022) propose a patch convolution attention Transformer in an encoder-decoder framework and Jiang et al. (2022a) design modified multi-head attention in the Transformer block and combine the CNN and Transformer block at each layer in the encoder for vessel retinal segmentation, both of them attempt to take the advantages of CNN and Transformers.

Conclusion. The Transformer-based medical segmentation methods can be categorized into different categories based on the encoder-decoder network components, *i.e.*, hybrid encoder + CNN decoder, Pure Transformer encoder + CNN decoder,

Table 15: An overview of Transformer methods on brain tumor segmentation application. The results are all tested on the BraTs validation dataset (Menze et al. (2014)), in which WT is the whole tumor, TC means the tumor core and ET is the enhanced tumor.

Method	Arch.	Dataset	Metric			Data split(Train:Val:Test)	Modality(Type)
			Dice WT	Dice TC	Dice ET		
Zhang et al. (2022b)	Hybrid encoder + CNN decoder	BraTs 2018	0.8964	0.8578	0.7761	285:66:NaN	MRI (3D)
Liang et al. (2022c)	Transformer encoder + Transformer decoder	BraTs 2018	0.9174	0.8553	0.8193	285:66:NaN	MRI (3D)
		BraTs 2019	0.9028	0.8173	0.7838	335:125:NaN	MRI (3D)
Liang et al. (2022b)	Hybrid encoder + CNN decoder	BraTs2018	0.9157	0.8568	0.8173	285:66:NaN	MRI (3D)
		BraTs 2019	0.9019	0.8257	0.7840	335:125:NaN	MRI (3D)
Wang et al. (2021f)	Hybrid encoder + CNN decoder	BraTs 2019	0.9000	0.8194	0.7893	335:125:NaN	MRI (3D)
		BraTs 2020	0.9009	0.8173	0.7873	369:125:NaN	MRI (3D)
Jun et al. (2021)	Hybrid encoder + CNN decoder	BraTs 2020	0.8695	0.6363	0.5063	369:125:NaN	MRI (3D)
Xing et al. (2022)	Pure Transformer encoder + CNN decoder	BraTs 2020	0.9200	0.8640	0.8000	315:17:37	MRI (3D)
Jiang et al. (2022b)	Hybrid encoder + CNN decoder	BraTs 2020	0.8906	0.8030	0.7736	369:125:NaN	MRI (3D)
		BraTs 2021	0.9183	0.8475	0.8321	1251:219:NaN	MRI (3D)
Liang et al. (2022a)	Pure Transformer encoder + CNN decoder	BraTs 2020	0.9076	0.8420	0.7948	371:127:NaN	MRI (3D)
		BraTs 2021	0.9183	0.8475	0.8321	1251:219:NaN	MRI (3D)
Jia and Shu (2021)	Hybrid encoder + CNN decoder	BraTs 2021	0.9097	0.8434	0.8187	1251:219:NaN	MRI (3D)

Table 16: An overview of Transformer methods on polyp segmentation application

Method	Arch.	Dataset	Metric	Data split(Train:Test)	Modality(Type)
Dong et al. (2021)	Pure Transformer encoder + CNN decoder	1.Kvasir-SEG (Jha et al. (2020))	Dice: 0.9170 mIoU: 0.8640	900:100	Endoscopy (2D)
		2.ClinicDB (Bernal et al. (2015))	Dice: 0.9370 mIoU: 0.8890	548:64	Endoscopy (2D)
		3.ColonDB (Tajbakhsh et al. (2015))	Dice: 0.8080 mIoU: 0.7270	NaN:380	Endoscopy (2D)
		4.Endoscene (Vázquez et al. (2017))	Dice: 0.9000 mIoU: 0.8330	NaN:60	Endoscopy (2D)
		5.ETIS (Silva et al. (2014))	Dice: 0.7870 mIoU: 0.7060	NaN;196	Endoscopy (2D)
Wang et al. (2022c)	Pure Transformer encoder + CNN decoder	1.ClinicDB (Bernal et al. (2015))	Dice: 0.9447 mIoU: 0.8995	548:64	Endoscopy (2D)
		2.Kvasir-SEG (Jha et al. (2020))	Dice: 0.9357 mIoU: 0.8905	900:100	Endoscopy (2D)
Sanderson and Matuszewski (2022)	Hybrid encoder + CNN decoder	1.Kvasir-SEG (Jha et al. (2020))	Dice: 0.9385 mIoU: 0.8903	900:100	Endoscopy (2D)
		2.ClinicDB (Bernal et al. (2015))	Dice: 0.9469 mIoU: 0.9020	548:64	Endoscopy (2D)
Park and Lee (2022)	Hybrid encoder + CNN decoder	1.Kvasir-SEG (Jha et al. (2020))	Dice: 0.9200 mIoU: 0.8700	900:100	Endoscopy (2D)
		2.ClinicDB (Bernal et al. (2015))	Dice: 0.9380 mIoU: 0.8920	550:62	Endoscopy (2D)
		3.ColonDB (Tajbakhsh et al. (2015))	Dice: 0.8040 mIoU: 0.7250	NaN:380	Endoscopy (2D)
		4.Endoscene (Vázquez et al. (2017))	Dice: 0.7580 mIoU: 0.6870	NaN:60	Endoscopy (2D)
		5.ETIS (Silva et al. (2014))	Dice: 0.9060 mIoU: 0.8420	NaN;196	Endoscopy (2D)
Duc et al. (2022)	Pure Transformer encoder + CNN decoder	1.Kvasir-SEG (Jha et al. (2020))	Dice: 0.9240 mIoU: 0.8760	900:100	Endoscopy (2D)
		2.ClinicDB (Bernal et al. (2015))	Dice: 0.9320 mIoU: 0.8840	548:64	Endoscopy (2D)
		3.ColonDB (Tajbakhsh et al. (2015))	Dice: 0.8110 mIoU: 0.7330	NaN:380	Endoscopy (2D)
		4.Endoscene (Vázquez et al. (2017))	Dice: 0.9060 mIoU: 0.8420	NaN:60	Endoscopy (2D)
		5.ETIS (Silva et al. (2014))	Dice: 0.8010 mIoU: 0.7220	NaN;196	Endoscopy (2D)
Tomar et al. (2022)	Hybrid encoder + CNN decoder	1.Kvasir-SEG (Jha et al. (2020))	Dice: 0.8884 mIoU: 0.8214	880:120	Endoscopy (2D)
		2.BKAI-IGH (Ngoc Lan et al. (2021))	Dice: 0.9154 mIoU: 0.8568	80:10:10	Endoscopy (2D)

CNN encoder + Pure Transformer decoder, Transformer encoder + Transformer decoder, and further subdivide them on the basis of inserted locations. The Transformer can be inserted into the CNN network as a plug-in, either as a bottleneck or multiple layers, to model long-term dependencies in various scales while retaining accurate spatial information and inductive bias from CNN. However, the interleaved CNN-Transformer structure can destroy consistency and delivery between features, which may be addressed by adopting a UNet-like Transformer architecture. Besides, we also discuss the performance of these methods for several important applications on popular benchmarks. However, it is important to note that the performance of these Transformer-based methods varies depending on the specific medical application and the benchmark dataset used for evaluation. For example, some methods may perform better on multi-organ segmentation tasks while others

may excel at cardiac diagnosis. Therefore, a thorough and impartial benchmark and a uniform public dataset are necessary to evaluate and compare the complexity and performance of these models across different medical applications. This would provide a clearer understanding of the strengths and weaknesses of each method and allow for more informed choices in selecting the most appropriate approach for a specific medical segmentation task.

5. Discussion

5.1. Summary

We have reviewed the diverse applications of the attention mechanism in medical image segmentation. The attention mechanism has been widely used in various tasks such as brain segmentation, breast segmentation, cardiac segmentation, lung segmentation, kidney segmentation, liver segmentation, pancreas segmentation, polyp segmentation, prostate segmentation, retinal segmentation, skin segmentation, and other miscellaneous tasks, as shown in Fig 7 (a). Most attention-based methods are designed to be universal and can be applied to different tasks. Various attention mechanisms have been introduced into medical image segmentation, as depicted in Fig 7 (b). This has contributed to the steady and sustained growth of attention-based methods in the field.

The traditional attention mechanism, which mimics the human visual system, focuses on the most important regions of an image and discards irrelevant parts, shifting the machine learning paradigm from large-scale vector transformation to more conscious processes. This approach is computationally efficient, low-cost, and interpretable, which is critical for medical images due to their large image resolution and small data scale. In the segmentation domain, channel and spatial attention are the most common non-Transformer attention types. Channel attention adaptively reweights each channel, acting as a feature selection process focusing on the target object (tissues, lesions, or organs). Spatial attention, conversely, can be seen as a spatial region selection mechanism that focuses on the area of interest or abnormality. We observe that spatial attention is more popular than channel attention, likely because edge information is significant for blurring boundaries in tasks like polyp segmentation, prostate segmentation, and skin lesion segmentation, among others. The non-Transformer attention mechanism is usually combined with convolutional layers in the encoder, decoder, or skip connection.

The limited receptive field and stationary weights of the traditional attention mechanisms have prompted the rapid and widespread adoption of the Transformer in numerous vision tasks, due to its ability to model long-range dependencies. As the Transformer was successfully applied in the field of vision, it has also gradually been applied in a very simple way to the field of medical image segmentation, such as being incorporated at the bottom of UNet or completely replacing standard convolutions. However, convolutional operations can capture local information, which is a weakness of Transformers. Consequently, researchers have improved Transformer blocks by combining them with convolutional blocks in multi-layers in a serial or parallel fashion to take advantage of multi-resolution features for more comprehensive modeling. Moreover, to address time and computational cost issues, DETR (Zhu et al. (2020)), Swin Transformer (Liu et al. (2021c)), and LeViT (Graham et al. (2021)) have replaced the vanilla Transformer block with more efficient ones. Additionally, researchers have continued to develop the integration of the Transformer into U-Net style architectures (e.g., pure Transformer encoder with CNN decoder) to leverage the U-like architecture's benefits.

5.2. Future Challenge

Despite the numerous successful applications of attention mechanisms in medical image segmentation, several challenges still need to be addressed.

5.2.1. Task-specific Attention

Different tasks may require distinct levels of attention to specific features or areas of interest. For instance, in lung nodule segmentation, a multi-scale attention mechanism is crucial for handling lesions of varying sizes. In brain tumor segmentation, a unique attention mechanism is essential for differentiat-

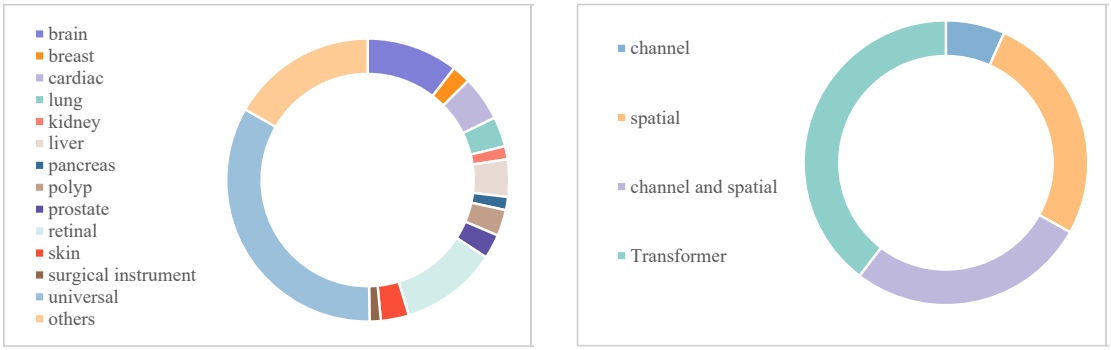

Fig. 7: The charts present statistics on attention-based medical image segmentation methods according to their applications and attention types. (a) displays the distribution of task-specific applications. (b) shows the ratio of different attention types.

ing between tumor and regular tissue. The required attention mechanism depends on the task at hand. Unfortunately, after reviewing the literature in this field, it has been found that most existing methods are not specifically designed for specialized medical image segmentation tasks. As a result, exploring this area further would be incredibly valuable.

5.2.2. *Robustness*

Although attention mechanisms can provide insight into the information that the model is focusing on, there may be instances where the attention results are partially or completely incorrect, leading to inaccurate model predictions. Currently, there is limited research focusing on the failure cases of attention models in medical image segmentation. It is crucial to study these cases in depth to understand the limitations and potential drawbacks of attention mechanisms, as well as to improve their robustness and generalizability in practical applications. By analyzing and addressing these failure cases, researchers can develop more reliable and accurate attention-based models, which in turn will enhance their adoption in clinical settings and their impact on patient care.

5.2.3. *Standard Evaluation*

The varying datasets, image processing modes, and data partitioning used for evaluation in the literature make it challenging to compare the accuracy and validity of the surveyed methods. It is essential to establish standard and diverse datasets that can fully represent the diversity of medical images, as well as a standardized training and validation process to confirm the effectiveness of proposed models. Additionally, exploring the performance impact of different types of attention mechanisms in various network locations or architectural designs can provide insights into better ways to apply attention mechanisms. This will help researchers identify optimal strategies for incorporating attention mechanisms into medical image segmentation tasks, ultimately leading to more accurate and efficient models.

5.2.4. *Multi-modality & Multi-task*

Multi-task learning helps improve a model's generalizability and performance by leveraging the relevance between different tasks. Thus, it is meaningful to build models capable of handling multiple tasks. Park et al. (2021) propose a federated split vision Transformer for COVID-19 diagnosis by co-training the segmentation, classification, and detection tasks. Beyond that, different medical imaging modalities provide complementary properties for diagnosis. Song et al. (2021) utilize the attention mechanism to model the pairwise relation between Optimal Coherence Tomography features and Visual Field features, in which the complementary information is passed from one modality to another by utilizing the Transformer model. While Xie et al. (2022) propose a Pyramid Transformer U-Net network to learn representations from diverse dimension data and transfer features to various downstream tasks through the switchable patch embedding layers, which outperforms the advanced SSL counterparts substantially. Thus, it is expected to see more work combing multi-task and multi-modality as the Transformer is inherently suitable for various sequence-based

inputs.

5.2.5. Complexity

The application of Transformers in medical image segmentation is often hindered by the high computational and memory costs. As a result, researchers have introduced efficient Transformer blocks into models. However, the multi-level feature fusion module, which can integrate features both locally and globally to enhance segmentation performance and address data sparsity issues, remains a high-cost component. CoTr (Xie et al. (2021c)) introduces the deformable self-attention mechanism to tackle this challenge. We anticipate further research and solutions in this specific direction, as it is essential for making Transformer-based architectures more accessible and efficient in medical image segmentation tasks.

6. Conclusion

In this paper, we have reviewed over 300 articles related to attention-based medical image segmentation applications, systematically surveying and summarizing the literature grouped into Non-Transformer and Transformer categories. We have provided an in-depth view of recent trends and future challenges in this field. Our aim is to give researchers a deeper understanding of the attention mechanisms applied in medical image segmentation and to serve as a springboard for future research. By examining the current state of the art and identifying potential areas for improvement, we hope to inspire the development of more advanced and effective attention-based techniques for medical image segmentation, ultimately contributing to better diagnosis and treatment in healthcare.

References

Abbasi-Sureshjani, S., Smit-Ockeloen, I., Zhang, J., Ter Haar Romeny, B., 2015. Biologically-inspired supervised vasculature segmentation in slo retinal fundus images, in: International Conference Image Analysis and Recognition, Springer. pp. 325–334.

Ahn, S.S., Ta, K., Thorn, S., Langdon, J., Sinusas, A.J., Duncan, J.S., 2021. Multi-frame attention network for left ventricle segmentation in 3d echocardiography, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 348–357.

Akil, M., Saouli, R., Kachouri, R., et al., 2020. Fully automatic brain tumor segmentation with deep learning-based selective attention using overlapping patches and multi-class weighted cross-entropy. Medical image analysis 63, 101692.

Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A., 2020. Dataset of breast ultrasound images. Data in brief 28, 104863.

Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., et al., 2019. 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 .

An, F.P., Liu, J.e., 2021. Medical image segmentation algorithm based on multilayer boundary perception-self attention deep learning model. Multimedia Tools and Applications 80, 15017–15039.

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al., 2022. The medical segmentation decathlon. Nature communications 13, 1–13.

Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al., 2011. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics 38, 915–931.

Arora, R., Raman, B., Nayyar, K., Awasthi, R., 2021. Automated skin lesion segmentation using attention-based deep convolutional neural network. Biomedical Signal Processing and Control 65, 102358.

Azad, R., Arimond, R., Aghdam, E.K., Kazerouni, A., Merhof, D., 2022. Daeformer: Dual attention-guided efficient transformer for medical image segmentation. arXiv preprint arXiv:2212.13504 .

Azad, R., Jia, Y., Aghdam, E.K., Cohen-Adad, J., Merhof, D., 2023a. Enhancing medical image segmentation with transception: A multi-scale feature fusion approach. arXiv preprint arXiv:2301.10847 .

Azad, R., Kazerouni, A., Heidari, M., Aghdam, E.K., Molaei, A., Jia, Y., Jose, A., Roy, R., Merhof, D., 2023b. Advances in medical image analysis with vision transformers: A comprehensive review. arXiv preprint arXiv:2301.03505 .

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C., 2017. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data 4, 1–13.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al., 2018. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 .

Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F., 2015. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized medical imaging and graphics 43, 99–111.

Bernal, J., Sánchez, J., Vilarino, F., 2012. Towards automatic polyp detection with a polyp appearance model. Pattern Recognition 45, 3166–3182.

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al., 2018. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging 37, 2514–2525.

Bhatkalkar, B.J., Reddy, D.R., Prabhu, S., Bhandary, S.V., 2020. Improving the performance of convolutional neural network for the segmentation of optic disc in fundus images using attention gates and conditional random fields. IEEE Access 8, 29299–29310.

Bi, R., Ji, C., Yang, Z., Qiao, M., Lv, P., Wang, H., 2022. Residual based attention-unet combing dac and rmp modules for automatic liver tumor segmentation in ct. Mathematical Biosciences and Engineering 19, 4703–4718.

Bilic, P., Christ, P.F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., Fu, C.W., Han, X., Heng, P.A., Hesser, J., et al., 2019. The liver tumor segmentation benchmark (lits). arXiv preprint arXiv:1901.04056 .

Buda, M., Saha, A., Mazurowski, M.A., 2019. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. Computers in biology and medicine 109, 218–225.

Budai, A., Bock, R., Maier, A., Hornegger, J., Michelson, G., 2013. Robust vessel segmentation in fundus images. International journal of biomedical imaging 2013.

Cai, Z., Xin, J., Shi, P., Wu, J., Zheng, N., 2022. Dstunet: Unet with efficient dense swin transformer pathway for medical image segmentation, in: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1–5.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2021. Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 .

Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H., 2019. Gcnet: Non-local networks meet squeeze-excitation networks and beyond, in: Proceedings of the IEEE/CVF international conference on computer vision workshops, pp. 0–0.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers, in: European conference on computer vision, Springer. pp. 213–229.

Chang, Y., Menghan, H., Guangtao, Z., Xiao-Ping, Z., 2021. Transclaw u-net: Claw u-net with transformers for medical image segmentation. arXiv preprint arXiv:2107.05188 .

Chen, B., Liu, Y., Zhang, Z., Lu, G., Zhang, D., 2021a. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. arXiv preprint arXiv:2107.05274 .

Chen, D., Yang, W., Wang, L., Tan, S., Lin, J., Bu, W., 2022. Pcat-unet: Unet-like network fused convolution and transformer for retinal vessel segmentation. PloS one 17, e0262689.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021b. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 .

Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S., 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5659–5667.

Chen, S., Bortsova, G., García-Uceda Juárez, A., Tulder, G.v., Bruijne, M.d., 2019a. Multi-task attention-based semi-supervised learning for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 457–465.

Chen, X., Yao, L., Zhang, Y., 2020. Residual attention u-net for automated multi-class segmentation of covid-19 chest ct images. arXiv preprint arXiv:2004.05645 .

Chen, X., Zhang, R., Yan, P., 2019b. Feature fusion encoder decoder network for automatic liver lesion segmentation, in: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019), IEEE. pp. 430–433.

Cheng, J., Liu, J., Kuang, H., Wang, J., 2022a. A fully automated multimodal mri-based multi-task learning for glioma segmentation and idh genotyping. IEEE Transactions on Medical Imaging .

Cheng, J., Tian, S., Yu, L., Lu, H., Lv, X., 2020. Fully convolutional attention network for biomedical image segmentation. Artificial Intelligence in Medicine 107, 101899.

Cheng, Z., Qu, A., He, X., 2022b. Contour-aware semantic segmentation network with spatial attention mechanism for medical image. The Visual Computer 38, 749–762.

Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al., 2018. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic), in: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), IEEE. pp. 168–172.

Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J., 2016. Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association 23, 304–310.

Ding, F., Yang, G., Liu, J., Wu, J., Ding, D., Xv, J., Cheng, G., Li, X., 2019. Hierarchical attention networks for medical image segmentation. arXiv preprint arXiv:1911.08777 .

Ding, F., Yang, G., Wu, J., Ding, D., Xv, J., Cheng, G., Li, X., 2020a. High-order attention networks for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 253–262.

Ding, X., Peng, Y., Shen, C., Zeng, T., 2020b. Cab u-net: An end-to-end category attention boosting algorithm for segmentation. Computerized Medical Imaging and Graphics 84, 101764.

Dobko, M., Kolinko, D.I., Viniavskyi, O., Yelisieiev, Y., 2021. Combining cnns with transformer for multimodal 3d mri brain tumor segmentation with self-supervised pretraining. arXiv preprint arXiv:2110.07919 .

Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L., 2021. Polyp-pvt: Polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:2108.06932 .

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .

Duc, N.T., Oanh, N.T., Thuy, N.T., Triet, T.M., Dinh, V.S., 2022. Colonformer: an efficient transformer based method for colon polyp segmentation. IEEE Access 10, 80575–80586.

Duran, A., Dussert, G., Rouviére, O., Jaouen, T., Jodoin, P.M., Lartizien, C., 2022. Prostattention-net: a deep attention model for prostate cancer segmentation by aggressiveness in mri scans. Medical Image Analysis , 102347.

Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L., 2020a. Pranet: Parallel reverse attention network for polyp segmentation, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 263–273.

Fan, T., Wang, G., Li, Y., Wang, H., 2020b. Ma-net: A multi-scale attention network for liver and tumor segmentation. IEEE Access 8, 179656–179665.

Fang, W., Han, X.h., 2020. Spatial and channel attention modulated network for medical image segmentation, in: Proceedings of the Asian Conference on Computer Vision.

Fang, Y., Huang, H., Yang, W., Xu, X., Jiang, W., Lai, X., 2022. Nonlocal convolutional block attention module vnet for gliomas automatic segmentation. International Journal of Imaging Systems and Technology 32, 528–543.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3146–3154.

Gao, C., Ye, H., Cao, F., Wen, C., Zhang, Q., Zhang, F., 2021a. Multiscale fused network with additive channel–spatial attention for image segmentation. Knowledge-Based Systems 214, 106754.

Gao, Y., Zhou, M., Liu, D., Metaxas, D., 2022. A multi-scale transformer for medical image segmentation: Architectures, model efficiency, and benchmarks. arXiv preprint arXiv:2203.00131 .

Gao, Y., Zhou, M., Metaxas, D.N., 2021b. Utnet: a hybrid transformer architecture for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 61–71.

Ge, R., Yang, G., Chen, Y., Luo, L., Feng, C., Ma, H., Ren, J., Li, S., 2019. K-net: Integrate left ventricle segmentation and direct quantification of paired echo sequence. IEEE transactions on medical imaging 39, 1690–1702.

Gonçalves, T., Rio-Torto, I., Teixeira, L.F., Cardoso, J.S., 2022. A survey on attention mechanisms for medical applications: are we moving towards better algorithms? .

Gong, Z., French, A.P., Qiu, G., Chen, X., 2022. Convtransseg: A multi-resolution convolution-transformer network for medical image segmentation. arXiv preprint arXiv:2210.07072 .

Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., Douze, M., 2021. Levit: a vision transformer in convnet's clothing for faster inference, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12259–12269.

Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S., 2020. Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE transactions on medical imaging 40, 699–711.

Guo, C., Szemenyei, M., Hu, Y., Wang, W., Zhou, W., Yi, Y., 2021a. Channel attention residual u-net for retinal vessel segmentation, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1185–1189.

Guo, C., Szemenyei, M., Yi, Y., Wang, W., Chen, B., Fan, C., 2021b. Sa-unet: Spatial attention u-net for retinal vessel segmentation, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE. pp. 1236–1242.

Guo, C., Szemenyei, M., Yi, Y., Zhou, W., Bian, H., 2020. Residual spatial attention network for retinal vessel segmentation, in: International Conference on Neural Information Processing, Springer. pp. 509–519.

Guo, D., Terzopoulos, D., 2021. A transformer-based network for anisotropic 3d medical image segmentation, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE. pp. 8857–8861.

Guo, L., Lei, B., Chen, W., Du, J., Frangi, A.F., Qin, J., Zhao, C., Shi, P., Xia, B., Wang, T., 2021c. Dual attention enhancement feature fusion network for segmentation and quantitative analysis of paediatric echocardiography. Medical Image Analysis 71, 102042.

Guo, M.H., Xu, T.X., Liu, J.J., Liu, Z.N., Jiang, P.T., Mu, T.J., Zhang, S.H., Martin, R.R., Cheng, M.M., Hu, S.M., 2022. Attention mechanisms in computer vision: A survey. Computational Visual Media , 1–38.

Gutman, D., Codella, N.C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A., 2016. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). arXiv preprint

arXiv:1605.01397 .
Hamilton, B., 2018. Kaggle data science bowl: Find the nuclei in divergent images to advance medical discovery.
Haseljić, H., Chatterjee, S., Frysch, R., Kulvait, V., Semshchikov, V., Hensen, B., Wacker, F., Brüsch, I., Werncke, T., Speck, O., et al., 2023. Liver segmentation using turbolift learning for ct and cone-beam c-arm perfusion imaging. Computers in Biology and Medicine , 106539.
Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., Xu, D., 2022a. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. arXiv preprint arXiv:2201.01266 .
Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D., 2022b. Unetr: Transformers for 3d medical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574–584.
He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., Gao, Y., Wang, Q., Zhang, J., Shen, D., 2022. Transformers in medical image analysis: A review. arXiv preprint arXiv:2202.12165 .
Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al., 2019. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445 .
van den Heuvel, T.L., de Bruijn, D., de Korte, C.L., Ginneken, B.v., 2018. Automated measurement of fetal head circumference using 2d ultrasound images. PloS one 13, e0200412.
Hoover, A., Kouznetsova, V., Goldbaum, M., 2000. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Transactions on Medical imaging 19, 203–210.
Hu, H., Shen, L., Guan, Q., Li, X., Zhou, Q., Ruan, S., 2022. Deep co-supervision and attention fusion strategy for automatic covid-19 lung infection segmentation on ct images. Pattern Recognition 124, 108452.
Hu, J., Shen, L., Albanie, S., Sun, G., Vedaldi, A., 2018a. Gather-excite: Exploiting feature context in convolutional neural networks. Advances in neural information processing systems 31.
Hu, J., Shen, L., Sun, G., 2018b. Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141.
Hu, J., Wang, H., Gao, S., Bao, M., Liu, T., Wang, Y., Zhang, J., 2019a. S-unet: A bridge-style u-net framework with a saliency mechanism for retinal vessel segmentation. IEEE Access 7, 174167–174177.
Hu, J., Wang, H., Wang, J., Wang, Y., He, F., Zhang, J., 2021. Sa-net: A scale-attention network for medical image segmentation. PloS one 16, e0247388.
Hu, Y., Deng, H., Zhou, Y., Chen, Y., Hao, Z., Yang, W., 2019b. Automatic kidney and tumor segmentation with attention-based v-net .
Huang, S., Li, J., Xiao, Y., Shen, N., Xu, T., 2022. Rtnet: relation transformer network for diabetic retinopathy multi-lesion segmentation. IEEE Transactions on Medical Imaging 41, 1596–1607.
Huang, X., Deng, Z., Li, D., Yuan, X., 2021. Missformer: An effective medical image segmentation transformer. arXiv preprint arXiv:2109.07162 .
Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W., 2019. Ccnet: Criss-cross attention for semantic segmentation, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 603–612.
Intelligent Medical Imaging Research Group, C., . Corneal nerve fiber dataset. [EB/OL]. https://imed.nimte.ac.cn/CORN.html.
Islam, M., Vibashan, V., Jose, V., Wijethilake, N., Utkarsh, U., Ren, H., 2019. Brain tumor segmentation and survival prediction using 3d attention unet, in: International MICCAI Brainlesion Workshop, Springer. pp. 262–272.
Jaeger, S., Candemir, S., Antani, S., Wáng, Y.X.J., Lu, P.X., Thoma, G., 2014. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. Quantitative imaging in medicine and surgery 4, 475.
Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., Lange, T.d., Johansen, D., Johansen, H.D., 2020. Kvasir-seg: A segmented polyp dataset, in: International Conference on Multimedia Modeling, Springer. pp. 451–462.
Ji, Y., Zhang, R., Wang, H., Li, Z., Wu, L., Zhang, S., Luo, P., 2021. Multi-compound transformer for accurate biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 326–336.
Jia, H., Song, Y., Huang, H., Cai, W., Xia, Y., 2019. Hd-net: Hybrid discriminative network for prostate segmentation in mr images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 110–118.
Jia, Q., Shu, H., 2021. Bitr-unet: a cnn-transformer combined network for mri brain tumor segmentation. arXiv preprint arXiv:2109.12271 .
Jia, Z., Wang, C., Sun, Z., Geng, H., Fu, H., 2022. A new segmentation network of pediatric renography based on attention mechanism, in: Proceedings of 2021 Chinese Intelligent Systems Conference, Springer. pp. 161–172.
Jiang, H., Shi, T., Bai, Z., Huang, L., 2019a. Ahcnet: An application of attention mechanism and hybrid connection for liver tumor segmentation in ct volumes. IEEE Access 7, 24898–24909.
Jiang, Y., Duan, L., Cheng, J., Gu, Z., Xia, H., Fu, H., Li, C., Liu, J., 2019b. Jointrcnn: a region-based convolutional neural network for optic disc and cup segmentation. IEEE Transactions on Biomedical Engineering 67, 335–343.
Jiang, Y., Liang, J., Cheng, T., Lin, X., Zhang, Y., Dong, J., 2022a. Mtpa_unet: Multi-scale transformer-position attention retinal vessel segmentation network joint transformer and cnn. Sensors 22, 4592.
Jiang, Y., Yao, H., Ma, Z., Zhang, J., 2021. Bi-sanet—bilateral network with scale attention for retinal vessel segmentation. Symmetry 13, 1820.
Jiang, Y., Zhang, Y., Lin, X., Dong, J., Cheng, T., Liang, J., 2022b. Swinbts: A method for 3d multimodal brain tumor segmentation using swin transformer. Brain Sciences 12, 797.
Jin, Q., Meng, Z., Sun, C., Cui, H., Su, R., 2020. Ra-unet: A hybrid deep attention-aware network to extract liver and tumor in ct scans. Frontiers in Bioengineering and Biotechnology , 1471.
Jin, Y., Cheng, K., Dou, Q., Heng, P.A., 2019. Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 440–448.
Jun, E., Jeong, S., Heo, D.W., Suk, H.I., 2021. Medical transformer: Universal brain encoder for 3d mri analysis. arXiv preprint arXiv:2104.13633 .
Karimi, D., Vasylechko, S.D., Gholipour, A., 2021. Convolution-free medical image segmentation using transformers, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 78–88.
Karthik, R., Menaka, R., Hariharan, M., Won, D., 2022. Contour-enhanced attention cnn for ct-based covid-19 segmentation. Pattern Recognition 125, 108538.
Kaul, C., Manandhar, S., Pears, N., 2019. Focusnet: An attention-based fully convolutional network for medical image segmentation, in: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019), IEEE. pp. 455–458.
Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., et al., 2021a. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. Medical Image Analysis 69, 101950.
Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., Baydar, B., Lachinov, D., Han, S., Pauli, J., Isensee, F., Perkonigg, M., Sathish, R., Rajan, R., Sheet, D., Dovletov, G., Speck, O., Nürnberger, A., Maier-Hein, K.H., Bozdağı Akar, G., Ünal, G., Dicle, O., Selver, M.A., 2021b. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. Medical Image Analysis 69, 101950. URL: http://www.sciencedirect.com/science/article/pii/S1361841520303145, doi:https://doi.org/10.1016/j.media.2020.101950.
Kearney, V., Chan, J.W., Wang, T., Perry, A., Yom, S.S., Solberg, T.D., 2019. Attention-enabled 3d boosted convolutional neural networks for semantic ct segmentation using deep supervision. Physics in Medicine & Biology 64, 135001.
Khanh, T.L.B., Dao, D.P., Ho, N.H., Yang, H.J., Baek, E.T., Lee, G., Kim, S.H., Yoo, S.B., 2020. Enhancing u-net with spatial-channel attention gate for abnormal tissue segmentation in medical imaging. Applied Sciences 10, 5729.
Kim, T., Lee, H., Kim, D., 2021. Uacanet: Uncertainty augmented context attention for polyp segmentation, in: Proceedings of the 29th ACM International Conference on Multimedia, pp. 2167–2175.
Kistler, M., Bonaretti, S., Pfahrer, M., Niklaus, R., Büchler, P., et al., 2013. The virtual skeleton database: an open access repository for biomedical research and collaboration. Journal of medical Internet research 15, e2930.
Kuang, H., Yang, D., Wang, S., Wang, X., Zhang, L., 2023. Towards simultaneous segmentation of liver tumors and intrahepatic vessels via cross-attention mechanism. arXiv preprint arXiv:2302.09785 .
Kuijf, H.J., Biesbroek, J.M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., Casamitjana, A., et al., 2019. Standardized assessment of automatic segmentation of white matter hyper-

intensities and results of the wmh segmentation challenge. IEEE transactions on medical imaging 38, 2556–2568.

Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A., 2015. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge, in: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge, p. 12.

Landman, B.A., Warfield, S., 2012. Miccai 2012: grand challenge and workshop on multi-atlas labeling, in: Proc. international conference on medical image computing and computer assisted intervention, MICCAI.

Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., et al., 2019. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE transactions on medical imaging 38, 2198–2210.

Lee, H., Kim, H.E., Nam, H., 2019. Srm: A style-based recalibration module for convolutional neural networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1854–1862.

Lee, H., Park, J., Hwang, J.Y., 2020. Channel attention module with multiscale grid average pooling for breast cancer segmentation in an ultrasound image. IEEE transactions on ultrasonics, ferroelectrics, and frequency control 67, 1344–1353.

Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., Rubin, D.L., 2017. A curated mammography data set for use in computer-aided detection and diagnosis research. Scientific data 4, 1–9.

Lei, B., Huang, S., Li, H., Li, R., Bian, C., Chou, Y.H., Qin, J., Zhou, P., Gong, X., Cheng, J.Z., 2020a. Self-co-attention neural network for anatomy segmentation in whole breast ultrasound. Medical image analysis 64, 101753.

Lei, Y., Dong, X., Tian, Z., Liu, Y., Tian, S., Wang, T., Jiang, X., Patel, P., Jani, A.B., Mao, H., et al., 2020b. Ct prostate segmentation based on synthetic mri-aided deep attention fully convolution network. Medical physics 47, 530–540.

Li, C., Tan, Y., Chen, W., Luo, X., Gao, Y., Jia, X., Wang, Z., 2020a. Attention unet++: A nested attention-aware u-net for liver ct image segmentation, in: 2020 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 345–349.

Li, C., Tan, Y., Chen, W., Luo, X., He, Y., Gao, Y., Li, F., 2020b. Anu-net: Attention-based nested u-net to exploit full resolution features for medical image segmentation. Computers & Graphics 90, 11–20.

Li, D., Rahardja, S., 2021. Bseresu-net: An attention-based before-activation residual u-net for retinal vessel segmentation. Computer Methods and Programs in Biomedicine 205, 106070.

Li, H., Xiong, P., An, J., Wang, L., 2018. Pyramid attention network for semantic segmentation. arXiv preprint arXiv:1805.10180 .

Li, J., Chen, J., Tang, Y., Wang, C., Landman, B.A., Zhou, S.K., 2023. Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. Medical image analysis , 102762.

Li, J., Wang, J., Tian, Q., Gao, W., Zhang, S., 2019a. Global-local temporal representations for video person re-identification, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 3958–3967.

Li, J., Wang, W., Chen, C., Zhang, T., Zha, S., Yu, H., Wang, J., 2022a. Transbtsv2: Wider instead of deeper transformer for medical image segmentation. arXiv preprint arXiv:2201.12785 .

Li, K., Qi, X., Luo, Y., Yao, Z., Zhou, X., Sun, M., 2020c. Accurate retinal vessel segmentation in color fundus images via fully attention-based networks. IEEE Journal of Biomedical and Health Informatics 25, 2071–2081.

Li, L., Weng, X., Schnabel, J.A., Zhuang, X., 2020d. Joint left atrial segmentation and scar quantification based on a dnn with spatial encoding and shape attention, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 118–127.

Li, L., Wu, F., Wang, S., Luo, X., Martin-Isla, C., Zhai, S., Zhang, J., Liu, Y., Zhang, Z., Ankenbrand, M.J., et al., 2022b. Myops: A benchmark of myocardial pathology segmentation combining three-sequence cardiac magnetic resonance images. arXiv preprint arXiv:2201.03186 .

Li, L., Xu, M., Wang, X., Jiang, L., Liu, H., 2019b. Attention based glaucoma detection: a large-scale database and cnn model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10571–10580.

Li, R., Li, M., Li, J., Zhou, Y., 2019c. Connection sensitive attention u-net for accurate retinal vessel segmentation. arXiv preprint arXiv:1903.05558 .

Li, S., Sui, X., Luo, X., Xu, X., Liu, Y., Goh, R., 2021a. Medical image segmentation using squeeze-and-expansion transformers. arXiv preprint arXiv:2105.09511 .

Li, X., Jiang, Y., Li, M., Yin, S., 2020e. Lightweight attention convolutional neural network for retinal vessel image segmentation. IEEE Transactions on Industrial Informatics 17, 1958–1967.

Li, X., Ma, S., Tang, J., Guo, F., 2022c. Transiam: Fusing multimodal visual features using transformer for medical image segmentation. arXiv preprint arXiv:2204.12185 .

Li, Y., Cai, W., Gao, Y., Li, C., Hu, X., 2022d. More than encoder: Introducing transformer decoder to upsample, in: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE. pp. 1597–1602.

Li, Y., Wang, S., Wang, J., Zeng, G., Liu, W., Zhang, Q., Jin, Q., Wang, Y., 2021b. Gt u-net: A u-net like group transformer network for tooth root segmentation, in: International Workshop on Machine Learning in Medical Imaging, Springer. pp. 386–395.

Li, Y., Yang, J., Ni, J., Elazab, A., Wu, J., 2021c. Ta-net: Triple attention network for medical image segmentation. Computers in Biology and Medicine 137, 104836.

Lian, S., Luo, Z., Zhong, Z., Lin, X., Su, S., Li, S., 2018. Attention guided u-net for accurate iris segmentation. Journal of Visual Communication and Image Representation 56, 296–304.

Liang, J., Yang, C., Zeng, L., 2022a. 3d pswinbts: An efficient transformer-based unet using 3d parallel shifted windows for brain tumor segmentation. Digital Signal Processing 131, 103784.

Liang, J., Yang, C., Zeng, M., Wang, X., 2022b. Transconver: transformer and convolution parallel network for developing automatic brain tumor segmentation in mri images. Quantitative Imaging in Medicine and Surgery 12, 2397.

Liang, J., Yang, C., Zhong, J., Ye, X., 2022c. Btswin-unet: 3d u-shaped symmetrical swin transformer-based network for brain tumor segmentation with self-supervised pre-training. Neural Processing Letters , 1–19.

Liao, F., Liang, M., Li, Z., Hu, X., Song, S., 2019. Evaluate the malignancy of pulmonary nodules using the 3-d deep leaky noisy-or network. IEEE transactions on neural networks and learning systems 30, 3484–3495.

Lin, A., Chen, B., Xu, J., Zhang, Z., Lu, G., 2021. Ds-transunet: Dual swin transformer u-net for medical image segmentation. arXiv preprint arXiv:2106.06716 .

Lin, D., Li, Y., Nwe, T.L., Dong, S., Oo, Z.M., 2020. Refineu-net: Improved u-net with progressive global feedbacks and residual attention guided local refinement for medical image segmentation. Pattern Recognition Letters 138, 267–275.

Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., Huisman, H., 2014a. Computer-aided detection of prostate cancer in mri. IEEE transactions on medical imaging 33, 1083–1092.

Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., et al., 2014b. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. Medical image analysis 18, 359–373.

Liu, C., Gu, P., Xiao, Z., 2022a. Multiscale u-net with spatial positional attention for retinal vessel segmentation. Journal of Healthcare Engineering 2022.

Liu, F., Wang, K., Liu, D., Yang, X., Tian, J., 2021a. Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography. Medical Image Analysis 67, 101873.

Liu, J., Liu, H., Gong, S., Tang, Z., Xie, Y., Yin, H., Niyoyita, J.P., 2021b. Automated cardiac segmentation of cross-modal medical images using unsupervised multi-domain adaptation and spatial neural attention structure. Medical Image Analysis 72, 102135.

Liu, Q., Kaul, C., Anagnostopoulos, C., Murray-Smith, R., Deligianni, F., 2022b. Optimizing vision transformers for medical image segmentation and few-shot domain adaptation. arXiv preprint arXiv:2210.08066 .

Liu, Y., Yang, G., Mirak, S.A., Hosseiny, M., Azadikhah, A., Zhong, X., Reiter, R.E., Lee, Y., Raman, S.S., Sung, K., 2019. Automatic prostate zonal segmentation using fully convolutional network with feature pyramid attention. IEEE access 7, 163626–163632.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021c. Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022.

Liu, Z., Wang, L., Wu, W., Qian, C., Lu, T., 2021d. Tam: Temporal adaptive module for video recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13708–13718.

Lo, P., Van Ginneken, B., Reinhardt, J.M., Yavarna, T., De Jong, P.A., Irving, B., Fetita, C., Ortner, M., Pinho, R., Sijbers, J., et al., 2012. Extraction

of airways from ct (exact'09). IEEE Transactions on Medical Imaging 31, 2093–2107.

Lou, A., Guan, S., Loew, M., 2021. Caranet: Context axial reverse attention network for segmentation of small medical objects. arXiv preprint arXiv:2108.07368 .

Lu, C., Guo, Z., Yuan, J., Xia, K., Yu, H., 2022. Fine-grained calibrated double-attention convolutional network for left ventricular segmentation. Physics in Medicine & Biology 67, 055013.

Lu, Z., Carneiro, G., Bradley, A.P., 2015. An improved joint optimization of multiple level set functions for the segmentation of overlapping cervical cells. IEEE Transactions on Image Processing 24, 1261–1272.

Lu, Z., Carneiro, G., Bradley, A.P., Ushizima, D., Nosrati, M.S., Bianchi, A.G., Carneiro, C.M., Hamarneh, G., 2016. Evaluation of three algorithms for the segmentation of overlapping cervical cells. IEEE journal of biomedical and health informatics 21, 441–450.

Luo, Z., Zhang, Y., Zhou, L., Zhang, B., Luo, J., Wu, H., 2019. Micro-vessel image segmentation based on the ad-unet model. IEEE Access 7, 143402–143411.

Lv, Y., Ma, H., Li, J., Liu, S., 2020. Attention guided u-net with atrous convolution for accurate retinal vessels segmentation. IEEE Access 8, 32826–32839.

Lyu, C., Hu, G., Wang, D., 2020. Attention to fine-grained information: hierarchical multi-scale network for retinal vessel segmentation. The Visual Computer , 1–11.

Ma, J., Wang, Y., An, X., Ge, C., Yu, Z., Chen, J., Zhu, Q., Dong, G., He, J., He, Z., Cao, T., Zhu, Y., Nie, Z., Yang, X., 2021. Towards data-efficient learning: A benchmark for covid-19 ct lung and infection segmentation. Medical Physics 48, 1197–1210. doi:https://doi.org/10.1002/mp.14676.

Maier-Hein, L., Wagner, M., Ross, T., Reinke, A., Bodenstedt, S., Full, P.M., Hempe, H., Mindroc-Filimon, D., Scholz, P., Tran, T.N., et al., 2021. Heidelberg colorectal data set for surgical data science in the sensor operating room. Scientific data 8, 101.

Maji, D., Sigedar, P., Singh, M., 2022. Attention res-unet with guided decoder for semantic segmentation of brain tumors. Biomedical Signal Processing and Control 71, 103077.

of Massachusetts General Hospital, M.M.C., . Internet brain segmentation repository. [EB/OL]. https://www.nitrc.org/projects/ibsr.

of Medical, I.S., Radiology, I., . Covid-19 ct segmentation dataset. [EB/OL]. http://medicalsegmentation.com/covid19/.

Mendonça, T., Ferreira, P.M., Marques, J.S., Marcal, A.R., Rozeira, J., 2013. Ph 2-a dermoscopic image database for research and benchmarking, in: 2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC), IEEE. pp. 5437–5440.

Mendrik, A.M., Vincken, K.L., Kuijf, H.J., Breeuwer, M., Bouvy, W.H., De Bresser, J., Alansary, A., De Bruijne, M., Carass, A., El-Baz, A., et al., 2015. Mrbrains challenge: online evaluation framework for brain image segmentation in 3t mri scans. Computational intelligence and neuroscience 2015.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al., 2014. The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging 34, 1993–2024.

Min, S., Chen, X., Zha, Z.J., Wu, F., Zhang, Y., 2019. A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4578–4585.

Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S., 2012. Inbreast: toward a full-field digital mammographic database. Academic radiology 19, 236–248.

Mostayed, A., Wee, W.G., Zhou, X., 2019. Content-adaptive u-net architecture for medical image segmentation, in: 2019 International Conference on Computational Science and Computational Intelligence (CSCI), IEEE. pp. 698–702.

Mou, L., Zhao, Y., Chen, L., Cheng, J., Gu, Z., Hao, H., Qi, H., Zheng, Y., Frangi, A., Liu, J., 2019. Cs-net: channel and spatial attention network for curvilinear structure segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 721–730.

Myronenko, A., Hatamizadeh, A., 2019. 3d kidneys and kidney tumor semantic segmentation using boundary-aware networks. arXiv preprint arXiv:1909.06684 .

Ngoc Lan, P., An, N.S., Hang, D.V., Long, D.V., Trung, T.Q., Thuy, N.T., Sang, D.V., 2021. Neounet: Towards accurate colon polyp segmentation and neoplasm detection, in: Advances in Visual Computing: 16th International Symposium, ISVC 2021, Virtual Event, October 4-6, 2021, Proceedings, Part II, Springer. pp. 15–28.

Ni, Z.L., Bian, G.B., Xie, X.L., Hou, Z.G., Zhou, X.H., Zhou, Y.J., 2019a. Rasnet: segmentation for tracking surgical instruments in surgical videos using refined attention segmentation network, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE. pp. 5735–5738.

Ni, Z.L., Bian, G.B., Zhou, X.H., Hou, Z.G., Xie, X.L., Wang, C., Zhou, Y.J., Li, R.Q., Li, Z., 2019b. Raunet: Residual attention u-net for semantic segmentation of cataract surgical instruments, in: International Conference on Neural Information Processing, Springer. pp. 139–149.

Ni, Z.L., Zhou, X.H., Wang, G.A., Yue, W.Q., Li, Z., Bian, G.B., Hou, Z.G., 2022. Surginet: Pyramid attention aggregation and class-wise self-distillation for surgical instrument segmentation. Medical Image Analysis 76, 102310.

Noori, M., Bahri, A., Mohammadi, K., 2019. Attention-guided version of 2d unet for automatic brain tumor segmentation, in: 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE), IEEE. pp. 269–275.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 .

Orlando, J.I., Fu, H., Breda, J.B., van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.A., Kim, J., Lee, J., et al., 2020. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical image analysis 59, 101570.

Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., Heidenreich, P.A., Harrington, R.A., Liang, D.H., Ashley, E.A., et al., 2020. Video-based ai for beat-to-beat assessment of cardiac function. Nature 580, 252–256.

Owen, C.G., Rudnicka, A.R., Mullen, R., Barman, S.A., Monekosso, D., Whincup, P.H., Ng, J., Paterson, C., 2009. Measuring retinal vessel tortuosity in 10-year-old children: validation of the computer-assisted image analysis of the retina (caiar) program. Investigative ophthalmology & visual science 50, 2004–2010.

Pace, D.F., Dalca, A.V., Geva, T., Powell, A.J., Moghari, M.H., Golland, P., 2015. Interactive whole-heart segmentation in congenital heart disease, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 80–88.

Park, K.B., Lee, J.Y., 2022. Swine-net: hybrid deep learning approach to novel polyp segmentation using convolutional neural network and swin transformer. Journal of Computational Design and Engineering 9, 616–632.

Park, S., Kim, G., Kim, J., Kim, B., Ye, J.C., 2021. Federated split vision transformer for covid-19cxr diagnosis using task-agnostic training. arXiv preprint arXiv:2111.01338 .

Peiris, H., Hayat, M., Chen, Z., Egan, G., Harandi, M., 2021. A volumetric transformer for accurate 3d tumor segmentation. arXiv preprint arXiv:2111.13300 .

Peiris, H., Hayat, M., Chen, Z., Egan, G., Harandi, M., 2022. Hybrid window attention based transformer architecture for brain tumor segmentation. arXiv preprint arXiv:2209.07704 .

Petit, O., Thome, N., Rambour, C., Themyr, L., Collins, T., Soler, L., 2021. U-net transformer: Self and cross attention for medical image segmentation, in: International Workshop on Machine Learning in Medical Imaging, Springer. pp. 267–276.

PUB, M.H., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, P., . The digital database for screening mammography, in: Proceedings of the Fifth International Workshop on Digital Mammography, pp. 212–218.

Punn, N.S., Agarwal, S., 2020. Chs-net: a deep learning approach for hierarchical segmentation of covid-19 infected ct images. arXiv preprint arXiv:2012.07079 .

Punn, N.S., Agarwal, S., 2022a. Chs-net: A deep learning approach for hierarchical segmentation of covid-19 via ct images. Neural Processing Letters , 1–22.

Punn, N.S., Agarwal, S., 2022b. Rca-iunet: a residual cross-spatial attention-guided inception u-net model for tumor segmentation in breast ultrasound imaging. Machine Vision and Applications 33, 1–10.

Qin, Y., Zheng, H., Gu, Y., Huang, X., Yang, J., Wang, L., Zhu, Y.M., 2020. Learning bronchiole-sensitive airway segmentation cnns by feature recali-

bration and attention distillation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 221–231.

Qin, Z., Zhang, P., Wu, F., Li, X., 2021. Fcanet: Frequency channel attention networks, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 783–792.

Qu, T., Wang, X., Fang, C., Mao, L., Li, J., Li, P., Qu, J., Li, X., Xue, H., Yu, Y., et al., 2022. M3net: A multi-scale multi-view framework for multi-phase pancreas segmentation based on cross-phase non-local attention. Medical image analysis 75, 102232.

Rahman, M.M., Marculescu, R., 2023. Multi-scale hierarchical vision transformer with cascaded attention decoding for medical image segmentation. arXiv preprint arXiv:2303.16892 .

Ren, Y., Yu, L., Tian, S., Cheng, J., Guo, Z., Zhang, Y., 2022. Serial attention network for skin lesion segmentation. Journal of Ambient Intelligence and Humanized Computing 13, 799–810.

Reza, A., Moein, H., Yuli, W., Dorit, M., 2022. Contextual attention network: Transformer meets u-net. arXiv preprint arXiv:2203.01932 .

Rikitake, R., Tsukada, Y., Ando, M., Yoshida, M., Iwamoto, M., Yamasoba, T., Higashi, T., 2019. Use of intensity-modulated radiation therapy for nasopharyngeal cancer in japan: analysis using a national database. Japanese journal of clinical oncology 49, 639–645.

Roy, A.G., Navab, N., Wachinger, C., 2018. Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 421–429.

Sabarinathan, D., Parisa Beham, M., Mansoor Roomi, S., 2019. Hyper vision net: kidney tumor segmentation using coordinate convolutional layer and attention unit, in: National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics, Springer. pp. 609–618.

Sagar, A., 2021. Vitbis: Vision transformer for biomedical image segmentation, in: Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-Preserving Machine Learning. Springer, pp. 34–45.

Sanderson, E., Matuszewski, B.J., 2022. Fcn-transformer feature fusion for polyp segmentation, in: Annual Conference on Medical Image Understanding and Analysis, Springer. pp. 892–907.

Scherer, D., Müller, A., Behnke, S., 2010. Evaluation of pooling operations in convolutional architectures for object recognition, in: Artificial Neural Networks–ICANN 2010: 20th International Conference, Thessaloniki, Greece, September 15-18, 2010, Proceedings, Part III 20, Springer. pp. 92–101.

Sha, Y., Zhang, Y., Ji, X., Hu, L., 2021. Transformer-unet: Raw image processing with unet. arXiv preprint arXiv:2109.08417 .

Shaker, A., Maaz, M., Rasheed, H., Khan, S., Yang, M.H., Khan, F.S., 2022. Unetr++: Delving into efficient and accurate 3d medical image segmentation. arXiv preprint arXiv:2212.04497 .

Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H., 2023. Transformers in medical imaging: A survey. Medical Image Analysis , 102802.

Shapey, J., Kujawa, A., Dorent, R., Wang, G., Dimitriadis, A., Grishchuk, D., Paddick, I., Kitchen, N., Bradford, R., Saeed, S.R., et al., 2021. Segmentation of vestibular schwannoma from mri, an open annotated dataset and baseline algorithm. Scientific Data 8, 1–6.

Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.i., Matsui, M., Fujita, H., Kodera, Y., Doi, K., 2000. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. American Journal of Roentgenology 174, 71–74.

Silva, J., Histace, A., Romain, O., Dray, X., Granado, B., 2014. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. International journal of computer assisted radiology and surgery 9, 283–293.

Simoncelli, E.P., Olshausen, B.A., 2001. Natural image statistics and neural representation. Annual review of neuroscience 24, 1193–1216.

Singh, V.K., Abdel-Nasser, M., Rashwan, H.A., Akram, F., Pandey, N., Lalande, A., Presles, B., Romani, S., Puig, D., 2019. Fca-net: Adversarial learning for skin lesion segmentation based on multi-scale features and factorized channel attention. IEEE Access 7, 130552–130565.

Sinha, A., Dolz, J., 2020. Multi-scale self-guided attention for medical image segmentation. IEEE journal of biomedical and health informatics 25, 121–130.

Sivaswamy, J., Krishnadas, S., Joshi, G.D., Jain, M., Tabish, A.U.S., 2014. Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation, in: 2014 IEEE 11th international symposium on biomedical imaging (ISBI), IEEE. pp. 53–56.

Soler, L., Hostettler, A., Agnus, V., Charnoz, A., Fasquel, J., Moreau, J., Osswald, A., Bouhadjar, M., Marescaux, J., 2010. 3d image reconstruction for comparison of algorithm database: A patient specific anatomical and medical image database. IRCAD, Strasbourg, France, Tech. Rep 1.

Song, D., Fu, B., Li, F., Xiong, J., He, J., Zhang, X., Qiao, Y., 2021. Deep relation transformer for diagnosing glaucoma with optical coherence tomography and visual field function. IEEE Transactions on Medical Imaging 40, 2392–2402.

Song, S., Dang, K., Yu, Q., Wang, Z., Coenen, F., Su, J., Ding, X., 2022. Bilateral-vit for robust fovea localization, in: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1–5.

Staal, J., Abràmoff, M.D., Niemeijer, M., Viergever, M.A., Van Ginneken, B., 2004. Ridge-based vessel segmentation in color images of the retina. IEEE transactions on medical imaging 23, 501–509.

Sun, H., Li, C., Liu, B., Liu, Z., Wang, M., Zheng, H., Feng, D.D., Wang, S., 2020. Aunet: attention-guided dense-upsampling networks for breast mass segmentation in whole mammograms. Physics in Medicine & Biology 65, 055005.

Sun, L., Ma, W., Ding, X., Huang, Y., Liang, D., Paisley, J., 2019. A 3d spatially weighted network for segmentation of brain tissue from mri. IEEE transactions on medical imaging 39, 898–909.

Sun, Q., Fang, N., Liu, Z., Zhao, L., Wen, Y., Lin, H., 2021. Hybridctrm: Bridging cnn and transformer for multimodal brain image segmentation. Journal of Healthcare Engineering 2021.

Tajbakhsh, N., Gurudu, S.R., Liang, J., 2015. Automated polyp detection in colonoscopy videos using shape and context information. IEEE transactions on medical imaging 35, 630–644.

Tan, W., Liu, P., Li, X., Xu, S., Chen, Y., Yang, J., 2022. Segmentation of lung airways based on deep learning methods. IET Image Processing .

Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A., 2022. Self-supervised pre-training of swin transformers for 3d medical image analysis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20730–20740.

Tang, Y.B., Tang, Y.X., Xiao, J., Summers, R.M., 2019. Xlsor: A robust and accurate lung segmentor on chest x-rays using criss-cross attention and customized radiorealistic abnormalities generation, in: International Conference on Medical Imaging with Deep Learning, PMLR. pp. 457–467.

Themyr, L., Rambour, C., Thome, N., Collins, T., Hostettler, A., 2022. Memory transformers for full context and high-resolution 3d medical segmentation, in: Machine Learning in Medical Imaging: 13th International Workshop, MLMI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings, Springer. pp. 121–130.

Tomar, N.K., Jha, D., Ali, S., Johansen, H.D., Johansen, D., Riegler, M.A., Halvorsen, P., 2021. Ddanet: Dual decoder attention network for automatic polyp segmentation, in: International Conference on Pattern Recognition, Springer. pp. 307–314.

Tomar, N.K., Shergill, A., Rieders, B., Bagci, U., Jha, D., 2022. Transresu-net: Transformer based resu-net for real-time colonoscopy polyp segmentation. arXiv preprint arXiv:2206.08985 .

Tong, H., Fang, Z., Wei, Z., Cai, Q., Gao, Y., 2021a. Sat-net: a side attention network for retinal image segmentation. Applied Intelligence 51, 5146–5156.

Tong, Q., Li, C., Si, W., Liao, X., Tong, Y., Yuan, Z., Heng, P.A., 2019. Rianet: Recurrent interleaved attention network for cardiac mri segmentation. Computers in biology and medicine 109, 290–302.

Tong, X., Wei, J., Sun, B., Su, S., Zuo, Z., Wu, P., 2021b. Ascu-net: attention gate, spatial and channel attention u-net for skin lesion segmentation. Diagnostics 11, 501.

Tragakis, A., Kaul, C., Murray-Smith, R., Husmeier, D., 2023. The fully convolutional transformer for medical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3660–3669.

Tschandl, P., Rosendahl, C., Kittler, H., 2018. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data 5, 1–9.

Vakanski, A., Xian, M., Freer, P.E., 2020. Attention-enriched deep learning model for breast tumor segmentation in ultrasound images. Ultrasound in medicine & biology 46, 2819–2833.

Van Noord, N., Postma, E., 2017. Learning scale-variant and scale-invariant features for deep image classification. Pattern Recognition 61, 583–592.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, A., 2017. A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of healthcare engineering 2017.

Video, .I.S.P.S., Cup, I.P.V., . Vip-cup18. [EB/OL]. https://users.encs.concordia.ca/~i-sip/2018VIP-Cup/index.html.

Wang, B., Qiu, S., He, H., 2019a. Dual encoding u-net for retinal vessel segmentation, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 84–92.

Wang, D., Haytham, A., Pottenburgh, J., Saeedi, O., Tao, Y., 2020a. Hard attention net for automatic retinal vessel segmentation. IEEE Journal of Biomedical and Health Informatics 24, 3384–3396.

Wang, D., Li, M., Ben-Shlomo, N., Corrales, C.E., Cheng, Y., Zhang, T., Jayender, J., 2019b. Mixed-supervised dual-network for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 192–200.

Wang, H., Cao, P., Wang, J., Zaiane, O.R., 2021a. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer, in: 36th AAAI Conference on Artificial intelligence, Vancouver.

Wang, H., Xie, S., Lin, L., Iwamoto, Y., Han, X.H., Chen, Y.W., Tong, R., 2022a. Mixed transformer u-net for medical image segmentation, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 2390–2394.

Wang, H., Xu, G., Pan, X., Liu, Z., Tang, N., Lan, R., Luo, X., 2022b. Attention-inception-based u-net for retinal vessel segmentation with advanced residual. Computers & Electrical Engineering 98, 107670.

Wang, J., Huang, Q., Tang, F., Meng, J., Su, J., Song, S., 2022c. Stepwise feature fusion: Local guides global. arXiv preprint arXiv:2203.03635 .

Wang, J., Peng, Y., Guo, Y., Li, D., Sun, J., 2021b. Ccut-net: pixel-wise global context channel attention ut-net for head and neck tumor segmentation, in: 3D Head and Neck Tumor Segmentation in PET/CT Challenge, Springer. pp. 38–49.

Wang, J., Wei, L., Wang, L., Zhou, Q., Zhu, L., Qin, J., 2021c. Boundary-aware transformers for skin lesion segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 206–216.

Wang, J., Zhao, X., Ning, Q., Qian, D., 2020b. Aec-net: attention and edge constraint network for medical image segmentation, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE. pp. 1616–1619.

Wang, K.N., Yang, X., Miao, J., Li, L., Yao, J., Zhou, P., Xue, W., Zhou, G.Q., Zhuang, X., Ni, D., 2022d. Awsnet: An auto-weighted supervision attention network for myocardial scar and edema segmentation in multi-sequence cardiac magnetic resonance images. Medical Image Analysis , 102362.

Wang, L., Wang, X., Zhang, B., Huang, X., Bai, C., Xia, M., Sun, P., 2021d. Multi-scale hierarchical transformer structure for 3d medical image segmentation, in: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE. pp. 1542–1545.

Wang, M., Zhu, W., Shi, F., Su, J., Chen, H., Yu, K., Zhou, Y., Peng, Y., Chen, Z., Chen, X., 2021e. Mstganet: Automatic drusen segmentation from retinal oct images. IEEE Transactions on Medical Imaging 41, 394–406.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q., 2020c. Eca-net: Efficient channel attention for deep convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J., 2021f. Transbts: Multimodal brain tumor segmentation using transformer, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 109–119.

Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L., 2021g. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568–578.

Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7794–7803.

Wang, X., Han, S., Chen, Y., Gao, D., Vasconcelos, N., 2019c. Volumetric attention for 3d medical image segmentation and detection, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 175–184.

Wang, Y., Dou, H., Hu, X., Zhu, L., Yang, X., Xu, M., Qin, J., Heng, P.A., Wang, T., Ni, D., 2019d. Deep attentive features for prostate segmentation in 3d transrectal ultrasound. IEEE transactions on medical imaging 38, 2768–2778.

Wang, Y., Wang, S., 2022. Skin lesion segmentation with attention-based sc-conv u-net and feature map distortion. Signal, Image and Video Processing , 1–9.

Wang, Y., Zhou, Y., Shen, W., Park, S., Fishman, E.K., Yuille, A.L., 2019e. Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. Medical image analysis 55, 88–102.

Wang, Z., Zou, Y., Liu, P.X., 2021h. Hybrid dilation and attention residual u-net for medical image segmentation. Computers in Biology and Medicine 134, 104449.

Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S., 2021a. Shallow attention network for polyp segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 699–708.

Wei, J., Huang, H., Sun, M., Wang, Y., Ren, M., He, R., Sun, Z., 2021b. Toward accurate and reliable iris segmentation using uncertainty learning. arXiv preprint arXiv:2110.10334 .

Wei, Z., Song, H., Chen, L., Li, Q., Han, G., 2019. Attention-based denseunet network with adversarial training for skin lesion segmentation. IEEE Access 7, 136616–136629.

Woo, S., Park, J., Lee, J.Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), pp. 3–19.

Wu, C., Zou, Y., Zhan, J., 2019. Da-u-net: densely connected convolutional networks and decoder with attention gate for retinal vessel segmentation, in: IOP Conference Series: Materials Science and Engineering, IOP Publishing. p. 012053.

Wu, C.Z., Sun, J., Wang, J., Xu, L.F., Zhan, S., 2021a. Encoding-decoding network with pyramid self-attention module for retinal vessel segmentation. International Journal of Automation and Computing 18, 973–980.

Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., Wen, Z., 2022a. Fat-net: Feature adaptive transformers for automated skin lesion segmentation. Medical Image Analysis 76, 102327.

Wu, H., Pan, J., Li, Z., Wen, Z., Qin, J., 2020. Automated skin lesion segmentation via an adaptive dual attention module. IEEE Transactions on Medical Imaging 40, 357–370.

Wu, M., Qian, Y., Liao, X., Wang, Q., Heng, P.A., 2021b. Hepatic vessel segmentation based on 3dswin-transformer with inductive biased multi-head self-attention. arXiv preprint arXiv:2111.03368 .

Wu, Y., Liao, K., Chen, J., Chen, D.Z., Wang, J., Gao, H., Wu, J., 2022b. D-former: A u-shaped dilated transformer for 3d medical image segmentation. arXiv preprint arXiv:2201.00462 .

Xia, H., Ma, M., Li, H., Song, S., 2022. Mc-net: multi-scale context-attention network for medical ct image segmentation. Applied Intelligence 52, 1508–1519.

Xian, M., Zhang, Y., Cheng, H.D., Xu, F., Huang, K., Zhang, B., Ding, J., Ning, C., Wang, Y., 2018. A benchmark for breast ultrasound image segmentation (BUSIS). Infinite Study.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021a. Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems 34, 12077–12090.

Xie, F., Huang, Z., Shi, Z., Wang, T., Song, G., Wang, B., Liu, Z., 2021b. Dudanet: a double u-shaped dilated attention network for automatic infection area segmentation in covid-19 lung ct images. International Journal of Computer Assisted Radiology and Surgery 16, 1425–1434.

Xie, Y., Zhang, J., Shen, C., Xia, Y., 2021c. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 171–180.

Xie, Y., Zhang, J., Xia, Y., Wu, Q., 2022. Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI, Springer. pp. 558–575.

Xing, Z., Yu, L., Wan, L., Han, T., Zhu, L., 2022. Nestedformer: Nested modality-aware transformer for brain tumor segmentation, in: International

Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 140–150.

Xiong, Z., Xia, Q., Hu, Z., Huang, N., Bian, C., Zheng, Y., Vesal, S., Ravikumar, N., Maier, A., Yang, X., et al., 2021. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Medical Image Analysis 67, 101832.

Xu, F., Zhang, Y., Cheng, H.D., Zhang, B., Ding, J., Ning, C., Wang, Y., 2021a. Tumor saliency estimation for breast ultrasound images via breast anatomy modeling. Artificial Intelligence in Medicine 119, 102155.

Xu, G., Wu, X., Zhang, X., He, X., 2021b. Levit-unet: Make faster encoders with transformer for medical image segmentation. arXiv preprint arXiv:2107.08623 .

Xu, H., Xie, H., Liu, Y., Cheng, C., Niu, C., Zhang, Y., 2019. Deep cascaded attention network for multi-task brain tumor segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 420–428.

Xu, L., Gao, S., Shi, L., Wei, B., Liu, X., Zhang, J., He, Y., 2021c. Exploiting vector attention and context prior for ultrasound image segmentation. Neurocomputing 454, 461–473.

Xu, S., Quan, H., 2021. Litetrans: Reconstruct transformer with convolution for medical image segmentation, in: International Symposium on Bioinformatics Research and Applications, Springer. pp. 300–313.

Xu, X., Lian, C., Wang, S., Wang, A., Royce, T., Chen, R., Lian, J., Shen, D., 2020. Asymmetrical multi-task attention u-net for the segmentation of prostate bed in ct image, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 470–479.

Xu, Y., Cai, M., Lin, L., Zhang, Y., Hu, H., Peng, Z., Zhang, Q., Chen, Q., Mao, X., Iwamoto, Y., et al., 2021d. Pa-resseg: A phase attention residual network for liver tumor segmentation from multiphase ct images. Medical Physics 48, 3752–3766.

Xuan, P., Cui, H., Zhang, H., Zhang, T., Wang, L., Nakaguchi, T., Duh, H.B., 2022. Dynamic graph convolutional autoencoder with node-attribute-wise attention for kidney and tumor segmentation from ct volumes. Knowledge-Based Systems 236, 107360.

Yan, Q., Wang, B., Zhang, W., Luo, C., Xu, W., Xu, Z., Zhang, Y., Shi, Q., Zhang, L., You, Z., 2020. Attention-guided deep neural network with multi-scale feature fusion for liver vessel segmentation. IEEE Journal of Biomedical and Health Informatics 25, 2629–2642.

Yan, X., Tang, H., Sun, S., Ma, H., Kong, D., Xie, X., 2022. After-unet: Axial fusion transformer unet for medical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3971–3981.

Yang, J., Qiu, K., 2021. An improved segmentation algorithm of ct image based on u-net network and attention mechanism. Multimedia Tools and Applications , 1–24.

Yang, K., Chang, S., Tian, Z., Gao, C., Du, Y., Zhang, X., Liu, K., Meng, J., Xue, L., 2022a. Automatic polyp detection and segmentation using shuffle efficient channel attention network. Alexandria Engineering Journal 61, 917–926.

Yang, X., Li, Z., Guo, Y., Zhou, D., 2022b. Dcu-net: a deformable convolutional neural network based on cascade u-net for retinal vessel segmentation. Multimedia Tools and Applications 81, 15593–15607.

Yang, X., Tian, X., 2022. Transnunet: Using attention mechanism for whole heart segmentation, in: 2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA), IEEE. pp. 553–556.

Yao, K., Su, Z., Huang, K., Yang, X., Sun, J., Hussain, A., Coenen, F., 2022. A novel 3d unsupervised domain adaptation framework for cross-modality medical image segmentation. IEEE Journal of Biomedical and Health Informatics .

Yap, M.H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwiggelaar, R., Davison, A.K., Marti, R., 2017. Automated breast ultrasound lesions detection using convolutional neural networks. IEEE journal of biomedical and health informatics 22, 1218–1226.

Yeung, M., Sala, E., Schönlieb, C.B., Rundo, L., 2021. Focus u-net: A novel dual attention-gated cnn for polyp segmentation during colonoscopy. Computers in biology and medicine 137, 104815.

Yin, S., Deng, H., Xu, Z., Zhu, Q., Cheng, J., 2022. Sd-unet: A novel segmentation framework for ct images of lung infections. Electronics 11, 130.

You, C., Zhao, R., Liu, F., Chinchali, S., Topcu, U., Staib, L., Duncan, J.S., 2022. Class-aware generative adversarial transformers for medical image segmentation. arXiv preprint arXiv:2201.10737 .

Yu, H., Shim, J.h., Kwak, J., Song, J.W., Kang, S.J., 2022. Vision transformer-based retina vessel segmentation with deep adaptive gamma correction, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1456–1460.

Yuan, W., Wei, J., Wang, J., Ma, Q., Tasdizen, T., 2019. Unified attentional generative adversarial network for brain tumor segmentation from multimodal unpaired images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 229–237.

Yun, B., Wang, Y., Chen, J., Wang, H., Shen, W., Li, Q., 2021. Spectr: Spectral transformer for hyperspectral pathology image segmentation. arXiv preprint arXiv:2103.03604 .

Zhang, C., Lu, J., Hua, Q., Li, C., Wang, P., 2022a. Saa-net: U-shaped network with scale-axis-attention for liver tumor segmentation. Biomedical Signal Processing and Control 73, 103460.

Zhang, C., Lu, J., Yang, L., Li, C., 2021a. Caagp: Rethinking channel attention with adaptive global pooling for liver tumor segmentation. Computers in Biology and Medicine 138, 104875.

Zhang, H., Goodfellow, I., Metaxas, D., Odena, A., 2019a. Self-attention generative adversarial networks, in: International conference on machine learning, PMLR. pp. 7354–7363.

Zhang, J., Dashtbozorg, B., Bekkers, E., Pluim, J.P., Duits, R., ter Haar Romeny, B.M., 2016. Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores. IEEE transactions on medical imaging 35, 2631–2644.

Zhang, J., Jiang, Z., Dong, J., Hou, Y., Liu, B., 2020. Attention gate resu-net for automatic mri brain tumor segmentation. IEEE Access 8, 58533–58545.

Zhang, J., Yu, L., Chen, D., Pan, W., Shi, C., Niu, Y., Yao, X., Xu, X., Cheng, Y., 2021b. Dense gan and multi-layer attention based lesion segmentation method for covid-19 ct images. Biomedical Signal Processing and Control 69, 102901.

Zhang, L., Lu, L., Nogues, I., Summers, R.M., Liu, S., Yao, J., 2017. Deeppap: deep convolutional networks for cervical cell classification. IEEE journal of biomedical and health informatics 21, 1633–1643.

Zhang, S., Fu, H., Yan, Y., Zhang, Y., Wu, Q., Yang, M., Tan, M., Xu, Y., 2019b. Attention guided network for retinal image segmentation, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 797–805.

Zhang, Y., He, N., Yang, J., Li, Y., Wei, D., Huang, Y., Zhang, Y., He, Z., Zheng, Y., 2022b. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. arXiv preprint arXiv:2206.02425 .

Zhang, Y., Higashita, R., Fu, H., Xu, Y., Zhang, Y., Liu, H., Zhang, J., Liu, J., 2021c. A multi-branch hybrid transformer network for corneal endothelial cell segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 99–108.

Zhang, Y., Liu, H., Hu, Q., 2021d. Transfuse: Fusing transformers and cnns for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 14–24.

Zhang, Z., Fu, H., Dai, H., Shen, J., Pang, Y., Shao, L., 2019c. Et-net: A generic edge-attention guidance network for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 442–450.

Zhang, Z., Sun, B., Zhang, W., 2021e. Pyramid medical transformer for medical image segmentation. arXiv preprint arXiv:2104.14702 .

Zhang, Z., Yin, F.S., Liu, J., Wong, W.K., Tan, N.M., Lee, B.H., Cheng, J., Wong, T.Y., 2010. Origa-light: An online retinal fundus image database for glaucoma analysis and research, in: 2010 Annual international conference of the IEEE engineering in medicine and biology, IEEE. pp. 3065–3068.

Zhao, J., Dai, L., Zhang, M., Yu, F., Li, M., Li, H., Wang, W., Zhang, L., 2020. Pgu-net+: progressive growing of u-net+ for automated cervical nuclei segmentation, in: Multiscale Multimodal Medical Imaging: First International Workshop, MMMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 1, Springer. pp. 51–58.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al., 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6881–6890.

Zhou, C., Chen, S., Ding, C., Tao, D., 2018. Learning contextual and attentive information for brain tumor segmentation, in: International MICCAI brainlesion workshop, Springer. pp. 497–507.

Zhou, C., Ding, C., Wang, X., Lu, Z., Tao, D., 2020a. One-pass multi-task net-

works with cross-task guided attention for brain tumor segmentation. IEEE Transactions on Image Processing 29, 4516–4529.

Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y., 2021a. nnformer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201 .

Zhou, T., Canu, S., Ruan, S., 2021b. Automatic covid-19 ct segmentation using u-net integrated spatial and channel attention mechanism. International Journal of Imaging Systems and Technology 31, 16–27.

Zhou, T., Ruan, S., Guo, Y., Canu, S., 2020b. A multi-modality fusion network based on attention mechanism for brain tumor segmentation, in: 2020 IEEE 17th international symposium on biomedical imaging (ISBI), IEEE. pp. 377–380.

Zhu, X., Hu, H., Wang, H., Yao, J., Ou, D., Xu, D., et al., 2021. Region aware transformer for automatic breast ultrasound tumor segmentation, in: Medical Imaging with Deep Learning.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J., 2020. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 .

Zhuang, Z., Li, N., Joseph Raj, A.N., Mahesh, V.G., Qiu, S., 2019. An rdau-net model for lesion segmentation in breast ultrasound images. PloS one 14, e0221535.

Zuo, Q., Chen, S., Wang, Z., 2021. R2au-net: attention recurrent residual convolutional neural network for multimodal medical image segmentation. Security and Communication Networks 2021.