International Journal of Network Dynamics and Intelligence Scilight

Survey/review study

Deep Learning Attention Mechanism in Medical Image Analysis: Basics and Beyonds

Xiang Li [1], Minglei Li [1], Pengfei Yan [1], Guanyi Li [1], Yuchen Jiang [1], Hao Luo [1,*], and Shen Yin [2]

[1] Department of Control Science and Engineering, Harbin Institute of Technology, Harbin 150001, China
[2] Department of Mechanical and Industrial Engineering, Faculty of Engineering, Norwegian University of Science and Technology, Trondheim 7034, Norway
[*] Correspondence: hao.luo@hit.edu.cn

Received: 16 October 2022
Accepted: 25 November 2022
Published: 27 March 2023

Abstract: With the improvement of hardware computing power and the development of deep learning algorithms, a revolution of "artificial intelligence (AI) + medical image" is taking place. Benefiting from diversified modern medical measurement equipment, a large number of medical images will be produced in the clinical process. These images improve the diagnostic accuracy of doctors, but also increase the labor burden of doctors. Deep learning technology is expected to realize an auxiliary diagnosis and improve diagnostic efficiency. At present, the method of deep learning technology combined with attention mechanism is a research hotspot and has achieved state-of-the-art results in many medical image tasks. This paper reviews the deep learning attention methods in medical image analysis. A comprehensive literature survey is first conducted to analyze the keywords and literature. Then, we introduce the development and technical characteristics of the attention mechanism. For its application in medical image analysis, we summarize the related methods in medical image classification, segmentation, detection, and enhancement. The remaining challenges, potential solutions, and future research directions are also discussed.

Keywords: medical image; attention mechanism; deep learning

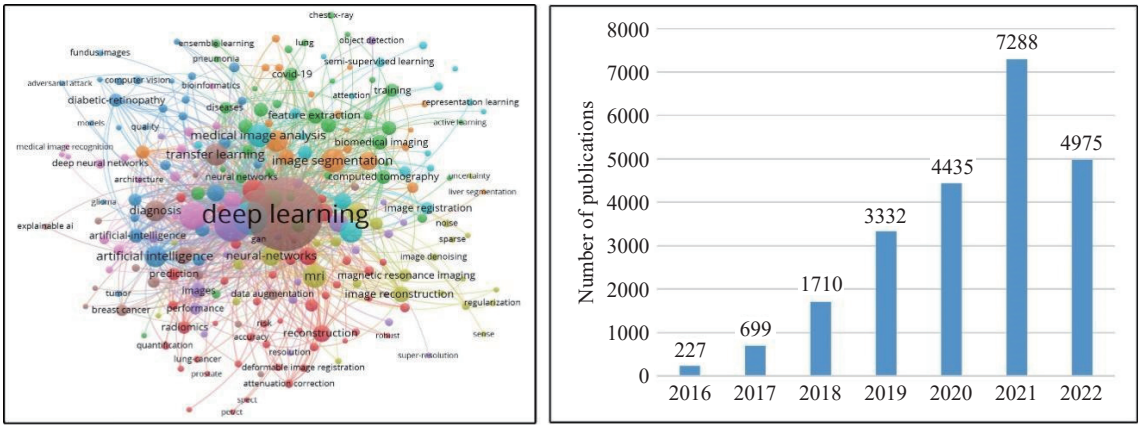
1. Introduction

Artificial intelligence (AI) augmented healthcare systems and services are changing the routine medical practice [1]. In modern medicine, the large amount of medical data provided by the advanced biological signal and medical imaging equipment has greatly promoted medical treatment, but has also greatly increased the cost of data analysis. AI technology is good at summarizing the patterns in large amounts of data to mimic human experts. Therefore, AI-augmented healthcare systems play a significant role in the entire medical process including preventive medicine, accurate diagnosis, and rehabilitation. The development of deep learning technology and the increased computing power of hardware devices are undoubtedly the key to launching the "AI + medical image" revolution [2]. Deep learning technology has good learning ability and can be applied to a variety of medical images, and its end-to-end overall structure can greatly improve the convenience of practical application by doctors. In the actual clinical process, medical images are qualitatively analyzed by professional physicians, but there exists an experience gap between different physicians, which may lead to the bias of qualitative analysis. The deep learning approach can assist clinical processes to reduce experience differences and reduce time and labor costs.

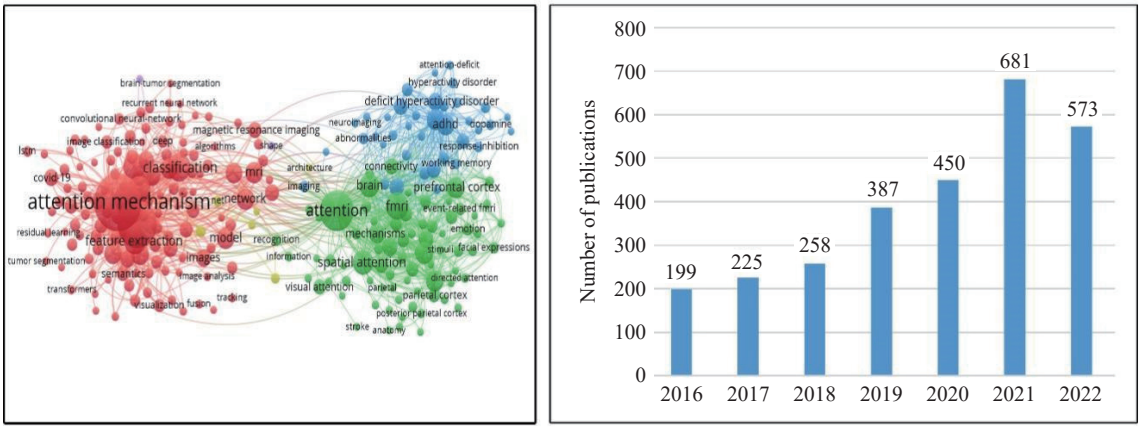
In recent years, the medical image analysis method based on attention mechanism combined with deep learning has attracted wide attention. These methods achieved the most advanced performance in multiple deep learning medical image tasks [3,4]. This paper summarizes the application of deep learning attention methods in medical image analysis. The attention mechanism is inspired by the human cognitive system and can be regarded as a dynamic selection process. Generally, the attention mechanism improves the performance by weighting important parts of data or building data correlations [5]. One of the important reasons that the two can be well combined and used in medical image analysis is that the attention mechanism has good embeddedness. The attention mechanism is usually used as a module of deep learning methods. The traditional convolutional neural network (CNN) is usually composed of basic layers such as the convolutional layer and pooling layer [6], and the receptive field range is very small (e.g., only 3 or 5 pixels). The visual attention module can be easily embedded between CNN layers to model global fea-

tures and analyze feature importance. Thus, the good embeddedness is not limited to CNN, and the attention mechanism can be applied to many methods in deep learning. For example, graph convolutional network (GCN) has good effect on graph data analysis in recent years, which is developed based on the graph neural network (GNN) and CNN. Many data can be summarized as graph data, which has specific nodes and edges. The input of GCN includes the node features and the adjacency matrix formed by the edges. GCN can learn and update the relationship between nodes according to node features, or integrate all node features to form an overall graph representation. Medical images can be constructed into graph data by feature extraction and analyzed by GCN in many cases. Li et al. [7] proposed a region aggregation GCN, which divided the hand bone X-ray image into multiple regions, and the attention mechanism integrated the features of the multiple regions for bone age prediction. In addition, another benefit of attention mechanism and deep learning in medical images is that it can improve the interpretability of deep learning [8]. The essence of the medical image analysis method is to serve the clinical process well, and the lesion attention information provided by the attention mechanism can give doctors intuitive clinical guidance.

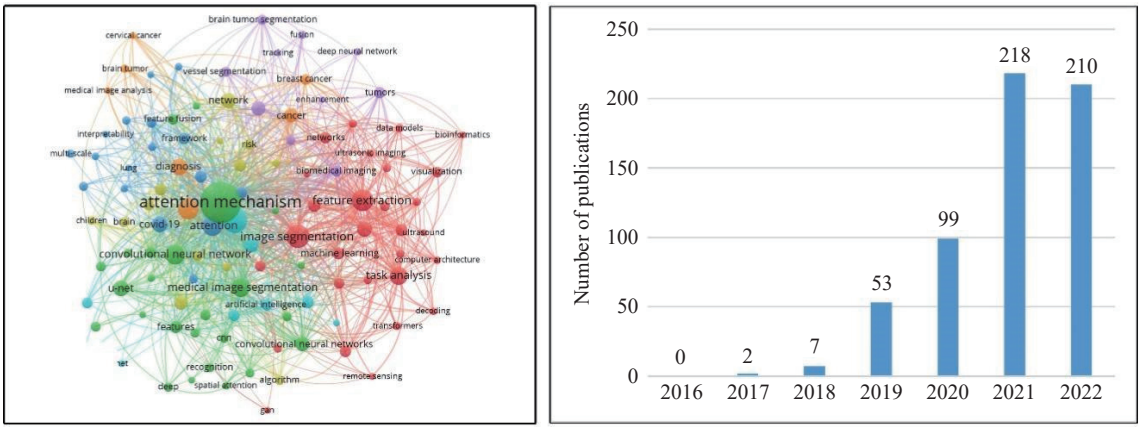
The literature survey of the medical image analysis based on attention mechanism and deep learning is shown in Figure 1. The literature survey is conducted on the Web of Science database. We first searched the keywords "Deep learning" and "Medical image", and the keyword network graph and the number of publications are shown in Figure 1 (**a**). As can be seen, the application of deep learning in the medical image is a hot research field, with the number of publications exceeding 4000 in each of the last three years. Then, we conducted a search using the keywords "attention mechanism" and "medical image", and the keyword network graph and the number of publications are shown in Figure 1 (**b**). It should be noted that the methods in Figure 1 (**b**) are not necessarily related to deep learning, as there are also many attention mechanisms used in traditional methods. Finally, we conducted the search using the keywords "deep learning", "attention mechanism" and "medical image", and the keyword network graph and the number of publications are shown in Figure 1 (**c**). The methods shown in Figure 1 (**c**) are a pure combination of deep learning and attention mechanism, which is the focus of this paper. As can be observed, the keyword of these approaches coveres many of the methodological subfields of deep learning. In addition, there are multiple categories of medical images appearing in the keyword network in Figure 1 (**c**), including "ultrasound", "MRI", etc., multiple major diseases, including "COVID-19", "brain tumor", etc. As the deep learning attention method is widely used in medical images, we conducted a comprehensive analysis of the existing methods.



(a) "Deep learning" and "Medical image"



(b) "Attention mechanism" and "Medical image"



(c) "Deep learning", "Attention mechanism" and "Medical image"

Figure 1. Keyword network graph of the researched literature and the number of publications.

Channel attention, hybrid attention, and transformer are the three main categories of attention mechanisms. Hybrid attention mechanism can integrate various attention mechanisms including channel attention and change the connection method of various attention mechanisms. Transformer is essentially a spatial attention mechanism, which decomposes the whole image into multiple patches, and then constructs the overall spatial relationship and spatial long-distance dependence. In most methods, the three attention mechanisms are independent of each other, and are combined by some algorithms in the way of hybrid attention mechanism.

In this paper, the application of channel attention, hybrid attention, and transformer in medical image analysis is mainly introduced. The main content of this paper is shown in Figure 2. The rest of this paper is given as follows. The second section introduces the concept, development, and attention mechanism technical characteristics. Section 3, Section 4, and Section 5 respectively introduce the application of channel attention mechanism, hybrid attention mechanism, and transformer in medical image tasks. In addition, we defined the medical image task as image classification task, image segmentation task, image detection task, and image enhancement task. Since the image classification task, image segmentation task and image detection task are three large fields in the medical image task, they are introduced separately. We combine reconstruction, denoising, registration, and super-resolution as the image enhancement task. Section 6 describes the remaining challenges and potential solutions. Section 7 is the conclusion of this paper.

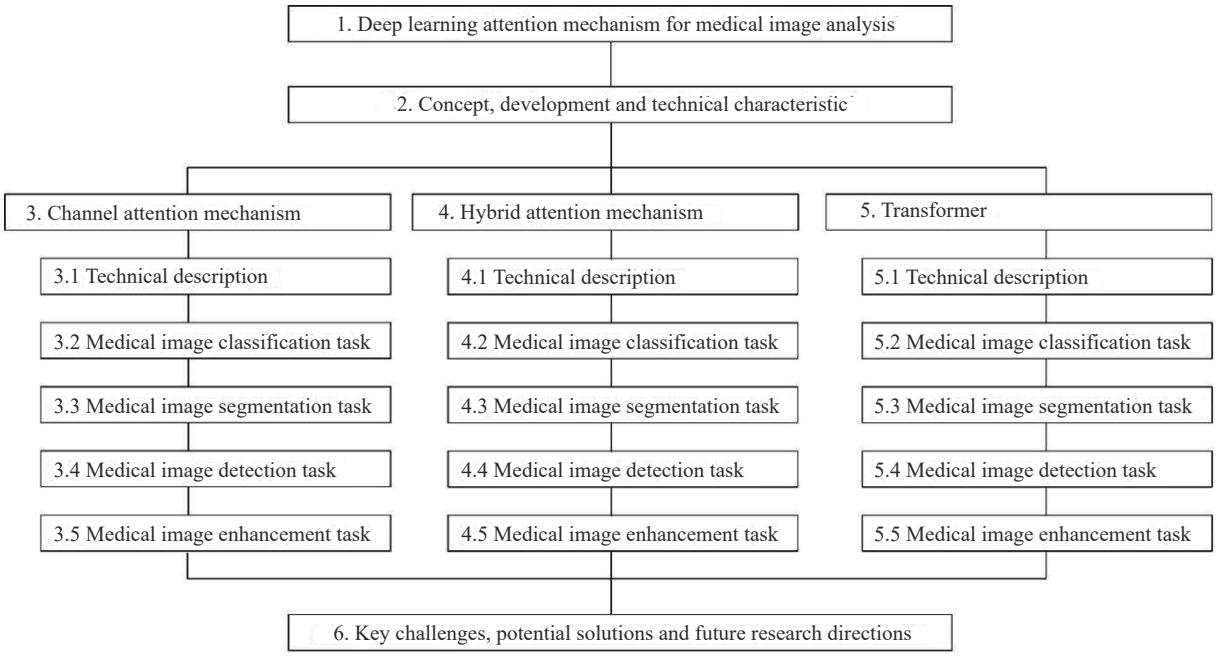


Figure 2. The main content of this paper.

2. Technical Concept of Attention Mechanism

2.1. Concept and Development

In recent years, attention mechanism has received widespread concern in AI research. The development of the attention mechanism is inspired by the human biological cognitive system and its information processing mechanism [9]. The human biological cognitive system usually does not equally treat all information when processing a large amount of external information. Based on previous life experience, the human cognitive system will first determines the importance of all information [10]. Then, the human brain will prioritizes important information and gives such information more thinking resources, while unimportant information is often selectively ignored. For example, the human eye is a part of the human cognitive system, and the high-resolution image only occupies a small part of the visual field of the eye, and all the peripheral images are low-resolution images [11]. The high-resolution part is called the fovea. The eye extracts feature from the fovea through continuous saccade movements [12]. Moreover, this process (information reception→importance discrimination→priority information processing) is learnable. As humans grow older and gain more life experience, the biological cognitive system will be enhanced through learning, which improves the efficiency and accuracy of information processing. The process of information processing in the human biological cognitive system is summarized as the attention mechanism.

In the 1980s, the concept of attention mechanisms was applied to engineering. Many researchers combined engineering, psychology, and biology to study AI technology in early studies. VISIT [13] is a visual attention model that uses psychophysical data and visual brain partitions biological data in the modeling process to improve the plausibility of attention mechanisms. They also explored the relationship between the model and the primary visual cortex, occipital, superior colliculus, and posterior parietal regions. Zhang et al. [14] combined computer vision object detection with human eye movement, and the model was consistent with the real human eye movement measurement data, including fixation times, cumulative probabilities and saccade paths, etc. Larochelle et al. [15] proposed an image classification system based on the fovea of the human eye, which includes a restricted boltzmann machine (RBM) and a gaze controller that are jointly trained to classify images. In the following research [16], a new attentional model for target tracking and recognition was developed. The model is closer to the human visual system, which consists of two interacting pathways. The first path models the object's appearance and category, and the second path simulates the location, orientation, scale, and speed of the object. It can be seen that early attention studies combined with biology mainly simulated the attention area of human eyes to the object, or simulated the thinking process of the human brain.

Before deep neural networks became the research hotspot, traditional machine learning was usually used as the carrier of attention mechanism. Fukushima et al. [17] designed a recognition system for hyphens in cursive. A composite graph was usually composed of multiple patterns. The proposed method adopted the attention mechanism to identify one pattern separately, and then focus attention on other patterns. Milanese et al. [18] studied the combination of data regions concerned by the attention model. They decomposed the image into features and saliency maps as cues, and then combined cues using a top-down approach. In addition, reinforcement learning was also applied to the attention model to learn object the order in object recognition [19]. The method of combining neural networks with attention mechanisms also appeared in early research. It should be noted that neural networks are mostly shallow networks rather than deep networks. Postma et al. [20] presented a signal channel attention network, which is a sparsely connected neural network and performs spatial selection through covert attention. In addition, reinforcement learning (RL) can also guide the learning process of attention. DasNet [21] used RL to select attention mechanisms and extract features from images. It also had feedback connections to optimize network parameters. Salah et al. [22] developed a neural network with selective attention, which used the Markov model to improve visual pattern recognition and was applied to handwritten digit recognition and face recognition.

After 2014, the attention mechanism and deep learning network have been well combined, and significant performance improvement has been achieved in multiple learning tasks. Bahdanau et al. [23] proposed RNNsearch for solving machine translation tasks. Although RNNsearch is not explicitly defined as an attention mechanism method in the original paper, its working principle is consistent with the characteristics of attention mechanism. In the previous work, using an encoder and decoder to build an end-to-end neural network is the basic solution to the sequence-to-sequence problem in machine translation tasks. However, the method based on an encoder and decoder has a large amount of computation for long sentences, and the translation accuracy is low because the correlation between words is not considered. RNNsearch combines translation and alignment together and uses the source sequence weighted sum of the encoder hidden layer when generating words in the target sequence, which greatly improves the translation accuracy. In subsequent research, the attention mechanism was extended from natural language processing to machine vision. Self-attention is a typical attention mechanism that was originally proposed in [24] . Self-attention constructs image features into specific vectors and computes the correlation. The correlation represents the relation-

ship between the global pixels of the image and is finally used to weigh image features, which can also be easily embedded in a variety of image task frameworks, including image segmentation [25], image classification [26], video action recognition [27], object detection [28], etc. Compared with traditional machine learning algorithms, the combination of attention mechanism and deep learning algorithm has more obvious advantages. Specifically, features in deep learning have achieved learnable extraction and optimization, while many traditional machine learning algorithms still rely on manual construction of features. In addition, the feature extraction and final decision in deep learning realize an end-to-end process, while traditional algorithms (e.g., traditional machine learning, feature extraction and decision) are two-stage processes. Finally, existing tasks usually have a huge amount of data. Deep learning networks can flexibly construct the network structure and the number of parameters, which provides more advantages for learning a large amount of data.

As an important branch of machine vision, a number of visual attention systems have been modified for medical images. Since 2017, deep learning attention mechanism has gradually appeared in medical image analysis. Hu et al. [29] designed a surgical tool recognition method with an attention mechanism, which included a global network and a local network. The global network can obtain the visual attention map and predictions, and the local network can refine the prediction results. Nie et al. [30] designed a semi-supervised segmentation network based on regional attention, which can incorporate unlabeled data into training. Xiao et al. [31] proposed an encoder-decoder structure with a weighted attention mechanism which is capable of processing small and fine blood vessels and improving the recognition ability of the optic disc region of the eye.

2.2. Technical Characteristics

In this part, we take RNNsearch (the first successful application in deep learning networks) and self-attention (a typical computer vision attention mechanism) as examples to introduce how the attention mechanism operates, affects deep learning networks and improves performance. The network structure and data flow of the two attention methods are shown in Figure 3.

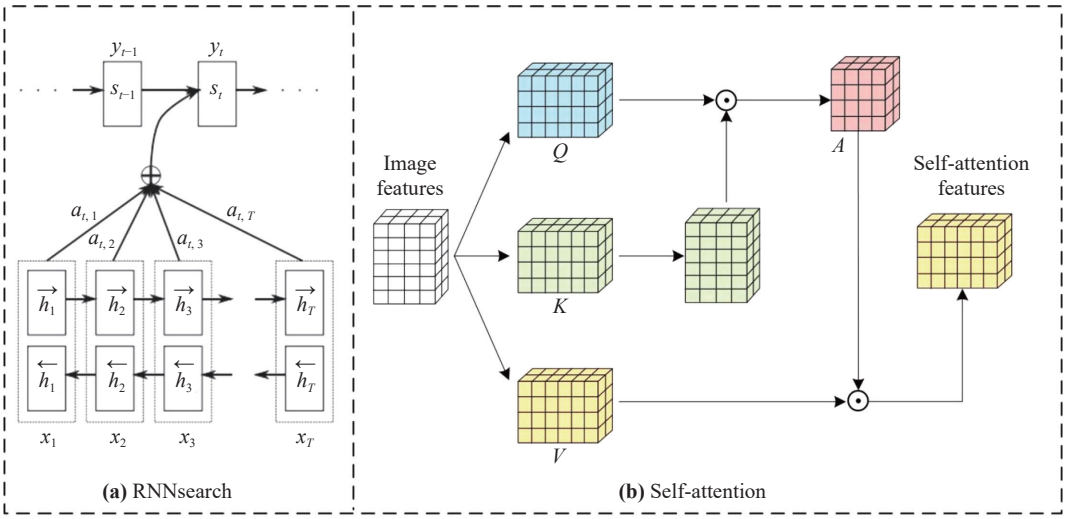


Figure 3. The structure of the RNNsearch and self-attention.

In the RNNsearch, the machine translation task is to translate a source sequence (sentence) x of length n into a target sequence (sentence) y of length m :

$$x = [x_1, x_2, \cdots, x_n] \tag{1}$$

$$y = [y_1, y_2, \cdots, y_m] \tag{2}$$

The encoder in RNNsearch uses bidirectional RNN and obtains the hidden state of forward propagation $\overrightarrow{h_i}$ and the hidden state of backward propagation $\overleftarrow{h_i}$. In order to obtain the context information of the sentence, RNNsearch splices the forward hidden state and the backward hidden state as the hidden layer state of the encoder:

$$h_i = [\overrightarrow{h_i}^T; \overleftarrow{h_i}^T]^T, i = 1, \ldots, n \tag{3}$$

The hidden layer state of the target sequence word at time t in the decoder network is:

$$s_t = f(s_{t-1}, y_{t-1}, c_t) \tag{4}$$

c_t is the context vector, computed by the decoder attention mechanism, which represents the context relationship between the current output and each word of the entire input sequence. c_t can be denoted as

$$c_t = \sum_{j=1}^{T} \alpha_{tj} h_j \tag{5}$$

where α_{tj} is the attention weight and $\alpha_{tj} = \frac{\exp(e_{tj})}{\sum_{k=1}^{T} \exp(e_{tk})}$. e is an alignment model that computes α_{tj} based on x_j (input at time j) and y_t (output at time t). α is the core of the method, which defines the weight of each output word for each input hidden state, so as to simulate the human attention mechanism.

In the self-attention mechanism, the input image feature is defined as Query (Q), Key (K), and Value (V), and the attention mechanism is to calculate the attention weight between the query and key, and then used to strengthen the value. Since Q, K, and V all come from the same image feature, it is called self-attention. To be specific, the image feature is X, and after a feature mapping function, K, Q, and V can be obtained. The features in the space of K, Q, and V roughly follow the same distribution. Then, K and Q can get the feature similarity matrix by matrix transpose and dot product operation. The similarity matrix is applied to the Softmax activation function to obtain the attention map between 0 and 1, which can be represented as:

$$\alpha_{ij} = Soft\max(\frac{QK^T}{\sqrt{d}}) \tag{6}$$

α_{ij} represents the contribution between the feature of the j_{th} position and the i_{th} position in the image feature, d is the length of the feature. The final output of attention is calculated as follows:

$$o_j = \sum_{j} \alpha_{ij} \times v_j \tag{7}$$

v_j is the j_{th} feature in the V, o_j is the output.

3. Channel Attention

In this section, we first introduce the technical development of the channel attention mechanism. Then the applications in medical image classification, segmentation, detection, and image enhancement are introduced. These methods are summarized in Table 1.

Table 1 Publication summary of the channel attention in medical image analysis

Method	Disease	Organ	Image	Dataset	Highlight	Performance
Classification task						
ResGANet [39]	Abnormal lesions in the digestive system	Intestine	CT	CAD-CAP WCE	Two branches for WCE image classification.	ACC = 93.17%
IL-MCAM [40]	Colorectal cancer	Colorectal	CT	HE-CRC-DS	Combines attention mechanisms into convolutional neural networks.	ACC = 99.78%
Yao et al. [41]	Breast cancer	Breast	X-Ray	BACH2018 Bioimaging2015	The parallel structure of deep neural network using CNN and RNN with an attention mechanism.	The state-of-the-art methods
MANet [42]	Brain tumor	Brain	MRI	BraTS'2018	Includes both spatial and cross-channel attention.	ACC = 96.51%
Segmentation task						
HDA-ResUNet [43]	Liver and lung tumors	Liver Lung	CT	LiTS 2017 DSB 2018 SBI 2012	Fully utilize advantages of U-Net, attention mechanism, and dilated convolution.	DICE = 97.03%
Sinha et al. [44]	Abdominal organs Cardiovascular	SpleenLiverKidneys	CT	CHAOS	The enhanced ability to model rich contextual dependencies over local features.	ACC(DSC) = 92.25% ACC(VS) = 97.85% ACC(MSD) = 92%
CA-Net [45]	Brain tumor	Fetal brain	CT	ISIC 2018	Proposing a joint spatial attention module to make the network focus more on the foreground region.	Dice = 95.88%
Detection task						
GCA-Net [47]	Cerebrovascular	Blood vessel	CTA	Private dataset	Combines a four-branch at the shallow feature.	Dice = 96.51%
Fan et al. [48]	COVID-19	Lung	X-Ray	COVID-19 Chest X-ray Dataset Initiative	A novel multi-kernel-size spatial-channel attention model.	ACC = 98.2%.
CAR-UNet [49]	Eye-related diseases	Retinal vessel	MRI	DRIVECHASE DB1 STARE	This method preserves performance while greatly reducing the network complexity in computer vision tasks.	The state-of-the-art performance
ResUNet ++ [50]	Colorectal cancer	Colon	Colonoscopy	Kvasir-SEG CVC-ClinicDB	The architecture takes advantage of the residual blocks, the squeeze and excitation block, ASPP, and the attention block.	Dice = 81.33%
Enhancement task						
DRAN [51]	Medical Image Denoising	Brain, skin	X-ray MRI CT	a.Radiology data b.Microscopy data c.Dermatology data	This study alleviates challenging denoising tasks by learning residual noise from a substantial extent of data samples.	a.SSIM = 0.0992 a.PSNR = 13.75 b.SSIM = 10.91 b.PSNR = 0.1137 c.SSIM = 11.17 c.PSNR = 0.3065
Rahman et al. [52]	Tomographic image reconstruction	Brain	MRI	Private dataset	Combines the multi-resolution features which lead to better preservation of the structural details.	SSIM = 85.48% PSNR = 93.83%
CA-Morph [53]	3D Medical Image Registration	Brain	MRI	a.OASIS b.LPBA40	Effectively improve the registration accuracy of 3D medical images.	a.Dice = 85.2% b.Dice = 71.2%
MedSRGAN [54]	Super-resolution	Brain	CTMRI	MOS	MedSRGAN not only preserves more texture details but also generates more realistic patterns on reconstructed SR images.	The average PSNR and SSIM outperform other methods

3.1. Channel Attention Technical Description

In the deep learning network, after multiple convolutional operations, the image will finally change to a multi-channel feature map. Deep learning feature maps of two-dimensional images generally have three dimensions, namely length, width, and channel. For example, an image with an input size of (H, W) will change to a feature map of size (H, W, 32) after passing through a convolutional layer with 32 convolution kernels. This calculation process is to decompose the input image into 32 information components, which have the key information of the original image

(different components contain different information). In the previous introduction, we mentioned that the attention mechanism is to mimic humans to find important information. The channel attention mechanism is to calculate the importance of channel component information and weighs the channels with key image information. In this section, we introduce the development of several channel attention mechanisms. From the original SE network to the PCA network, many new methods have been proposed to solve this problem, and the network structure development is shown in Figure 4.

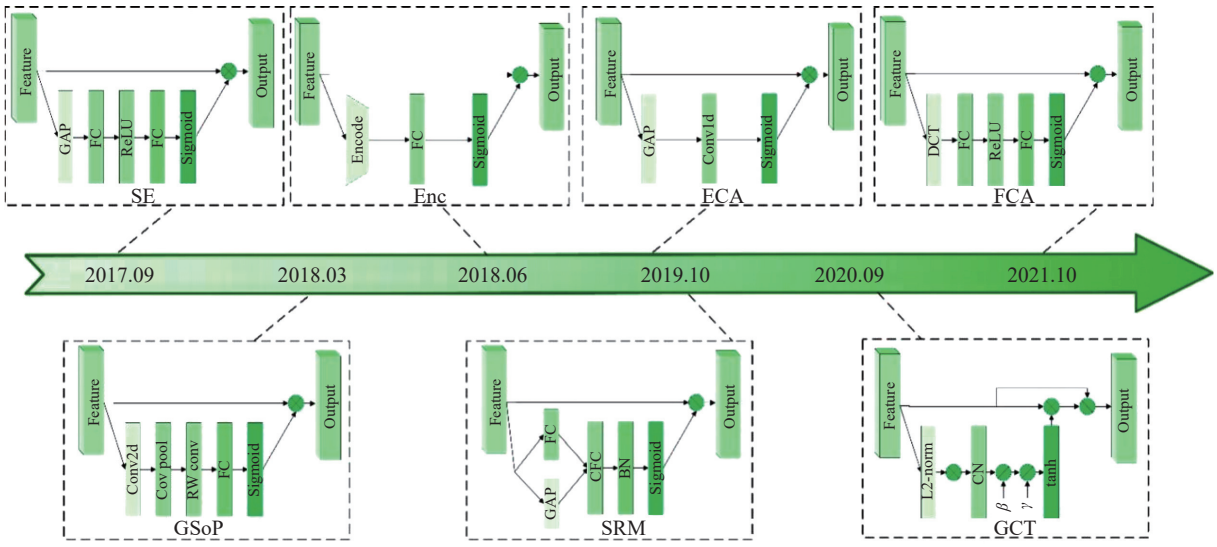


Figure 4. The Channel attention network structure development.

Squeeze-and-excitation network (SENet) [32] is the first one to use channel attention methods. The key part of SENet is an SE block mainly used to gather the global message, and obtain channel-wise relevance to enhance the ability of representation. The SE block compresses features into channel dimensions through the global pooling layer to obtain channel weights. This operation ignores the information in the width and height dimensions, reducing the ability to model higher-order statistics. To correct this defect, Gao et al. [33] designed a global second-order pooling (GSoP) method. GSoP uses covariance matrices to scale features on the channel dimension and to compute second-order statistics in features. Another advantage of GSoP is that the covariance matrix can be extended for spatial dimensions. Inspired by the SENet, Zhang et al. [34] proposed a new method named the context encoding module (CEM). The global context information of the shallow network is introduced into the loss function to strengthen the scene category information in the image segmentation task. Specifically, the shallow network features use the fully connected layer to encode the category information into the SE loss, and the other pathway uses a weighted scale to weigh the category of each channel. However, this method fails to achieve its goal of directly corresponding models between inputs and weight vectors. To solve this problem, Wang et al. [35] developed the efficient channel attention (ECA) method that mainly adopted a 1D convolution technique to control the channel interaction. Motivated by the ECA, Lee et al. [36] proposed the lightweight style-based recalibration module (SRM). This method successfully makes style transfer combined with the channel attention mechanism. Yang et al. [37] raised the gated channel transformation (GCT) block. This method achieves the goal to gather the information together and can exactly model the channel-wise relationships too. To prove mathematically that the traditional global average pool is a special case of eigendecomposition in the frequency domain. Qin et al. [38] designed a new Frequency Channel Attention Network (FcaNet), where the discrete cosine transform was used to compress the channel.

3.2. Channel Attention in Medical Image Classification Task

Some deep learning architectures have been raised and used in many fields such as classification, segmentation, and detection. Channel attention in medical image classification has been extensively used in medical image analysis and disease treatment, and many methods have been proposed to deal with this problem. Guo et al. [39] designed a semi-supervised ResGANet for recognizing bleeding, polyp, ulcer and other abnormal lesions in the digestive system. This method achieved 93.17% overall accuracy in fourfold cross-validation, thus proving efficient and convenient for image classification. Besides, Chen et al. proposed IL-MCAM [40] that was used for recognizing colorectal cancer. This method can be applied to solve colorectal cancer histopathology image classification tasks where attention mechanism was embeded into convolutional neural networks. Yao et al. [41] also proposed a new method that combines CNN and RNN to identify breast cancer, the proposed model performed perfectly on three datasets, showing the advantages of combined CNN and RNN. MANet [42] provided a way to utilize channel attention mechanisms to

recognize and classify brain tumor images too. This method achieves high accuracy in recognizing brain tumor images and makes better progress in several existing models for the tumor recognition task.

3.3. Channel Attention in Medical Image Segmentation Task

Channel attention in medical image segmentation is a crucial tool for analyzing a medical image and recognizing disease; and there are many methods raised to enrich this field. HDA-ResUNet [43] has been used for recognizing the disease of liver and tumor segmentation. This method fully combines the advantages of attention mechanism, U-Net, and dilated convolution. Besides, when facing fewer parameter problems, this method performs better with higher accuracy in the segmentation results than U-Net and also solves the slow convergence speed problem of U-Net. Multi-scale guided attention network for medical image segmentation proposed by Sinha et al. [44] is another method that has been used to solve this problem, and this method is mainly used for the treatment of cardiovascular structures, brain tumors, and abdominal organs. The model achieves a better performance than all previous methods qualitatively and quantitatively. This demonstrates the effectiveness of the approach of making accurate and reliable segmentations of the medical images. After that, a new method proposed by Gu et al. [45] named CA-Net was used for the recognition of brain tumors, where the collective spatial attention module was first used to let the model pay more attention to the foreground region and mainly focusing on enhancing the accuracy of segmentation which is hard to explain, it aims at designing a network with better comprehensive property, high accuracy, and efficiency at the same time. Mou et al. [46] proposed a new vascular structure segmentation network based on channel attention mechanism, which can obtain long-term dependency relationship and effectively use multi-channel space for normalization. It is effective in 2D and 3D vascular segmentation.

3.4. Channel Attention in Medical Image Detection Task

Besides the above methods mentioned in different application fields, channel attention has also been used in the medical image detection task. Many methods have been raised to solve this problem. Ni et al. [47] proposed a GCA-Net used for the detection task of liver and lung tumors. This method perfectly gains global context information which aims at preserving more details of the feature by implementing a four-branch method at the shallow feature, and its performance is better than previous methods. Multi-kernel-size spatial-channel attention method which has been used to justify whether this disease is COVID-19 or not raised by Fan et al. [48], is another method for COVID-19 detection. This method enables early diagnosis from X-ray images; and the method makes the representation and the accuracy of COVID-19's check and measure better. After that, Guo et al. raised a new approach named CAR-UNet [49] that was used for the detection of eye-related diseases and achieved better performance and greatly reduced the complexity of the model in the task. The result of this projected CAR-UNet performs better than previous on all three datasets. There is also a new method named ResUNet + +[50] that made great progress in the detection of colorectal cancer. The ResUNet ++ model based on the deep residual U-Net (ResUNet) is an architecture that absorbs the highlights of deep learning but also obtains the advantages of U-Net. As a result, this method performs well in the medical image detection task.

3.5. Channel Attention in Medical Image Enhancement Task

Sharif et al. [51] proposed the DRAN that can significantly reduce image noise and enhance the analysis outcomes of various medical images. The DRAN learns residual noises from a large number of data examples to achieve significant denoising tasks. Besides, another method named efficient U-Net with a symmetric decoder was proposed by Rahman et al. [52] and had been used for the construction of the tomographic image. This decoder model combines the multi-resolution features generated by the EfficientNet encoder using increased feature dilation tactics, which can better preserve the detail of the structure in reconstructed images. Yin et al. [53] also projected a creative method named CA-Morph to tackle the matter of 3D Medical Image Registration. Through the combination of the above techniques, the method greatly ameliorates the accuracy in the registration of 3D medical images. The method achieves the purpose of capturing more important features required for registration from many aspects. MedSRGAN proposed by Gu et al. [54] achieved the goal of realizing super-resolution of medical images. The method not only contains more details and key points but also obtains more features through reconstructed SR images. To summarize, this method reaches a better result than previous methods when reconstructing feature and texture details on data and the outcomes are much more similar to the actual images.

In all, with the development of the attention mechanism, channel attention has been extensively used in medical image classification, segmentation, detection, and enhancement task. All of them have achieved great progress in promoting the combination of medical images and deep learning.

4. Hybrid Attention

In this section, the technical development of the hybrid attention mechanism will be described in detail. Then,

the applications in medical image classification, segmentation, detection, and image enhancement are introduced. These methods are summarized in Table 2.

Table 2 Publication summary of the hybrid attention in medical image analysis

Method	Disease	Organ	Imgae	Dataset	Highlight	Performance
Classification task						
HIENet [62]	Adenocarcinoma	Endometrial	Histopathological	Histopathological image dataset for endometrial disease diagnosis	The proposed HIENet network was adopted to classify endometrial histopathological images and achieve high performance.	a. AUC = 0.96 b. AUC = 0.98
PCAM [63]	Macular	Retina	OCT	UCSD, NEH, and Duke	The hybrid attention is presented to extract inter-class discriminative features by focusing on the feature space of each level of CNN.	ACC = 99.79%
BASCNet [64]	Breast cancer	Breast	Mammogram	DDSM and INbreast	The BASCNet is proposed, which adaptively pays attention to discriminative features in spatial and channel dimensions.	a. ACC = 85.10% b. ACC = 90.51%
Segmentation task						
FFANet [70]	Abdominal disease and melanoma	Abdominal and skin	MRI and dermoscopic	CHAOS and ISIC2017	The hybrid attention module is adopted to consider the relevance of each space and channel.	a. Dice = 90.90%, b. Dice = 86.10%
TA-Net [46]	Diabetes, melanoma,	Retina, skin, intracranial vessels, cell	CT, dermoscopic, electron microscope sectioning	DRIVE, STARE, ISBI 2012, ISIC 2017, intracranial blood vessels	The TA-Net is proposed to explore the ability of attention mechanisms to simultaneously identify global contextual information in the channel, spatial, and feature interior domains.	a. Dice = 90.62% b. Dice = 86.67%
ω-Net [71]	Cancers of the kidney, pancreas and liver	Kidney, pancreas and liver	CT	KiTS19, Decathlon	The multidimensional self-attention is developed to highlight salient features and suppress irrelevant features continuously in spatial and channel dimensions.	a. Dice = 90.47% b. Dice = 64.95% c. Dice = 91.11%
Detection task						
Improved Faster RCNN [74]	Chapped tongue	Tongue	Original image	Private datasets	The hybrid attention mechanisms were used to emphasize local lesion information and suppress background tongue information.	Precision = 80.39%
EGCD [75]	Stomach cancer	Stomach	Gastroscopy images	Dataset provided by Chongqing University Cancer Hospital	Combined with CBAM to enhance the characteristic expression of tumor targets in channel and space, highlighting tumor targets and suppressing background information.	mAP = 94.16%
Attention-YOLO [76]	Potential diseases	Cell	Blood cell images	BCCD, RBCS	Attention-YOLO uses filtered and weighted feature vectors instead of original feature vectors for residual fusion to improve network detection accuracy.	mAP = 94.30%
Enhancement task						
DR-GAN [80]	Diabetes	Retina	Retinopathy images	EyePACS, FGADR	A multi-scale space and channel attention module were designed to improve the generation of synthetic small details.	ACC = 89.45 Kappa = 88.11
FA-GAN [81]	None	Cardiac, brain, knee, MMWHS	MRI	MICCAI 2013	Channel attention module, self-attention module, and fusion operations are used, aiming to enhance the important features of MR images.	a. PSNR = 38.58 b. PSNR = 44.11 c. PSNR = 36.58 d. PSNR = 39.78
RED-MAM [85]	None	Heart, head circumference	Ultrasound image	FH, HC18, CAMUS	Multi-attention fusion attention is introduced to make the model focus on ultrasound image texture structure information and enhance the denoising effect on ultrasound images.	PSNR = 38.39 MSE = 0.0013 MAE = 0.0229

4.1. Hybrid Attention Technical Description

The network structure development of hybrid attention is shown in Figure 5. In the process of employing deep learning technology to extract features from images, the contribution of images in different regions to the model task

is unbalanced, and the task-related regions contain key information that the model needs to focus on. Spatial attention only considers the spatial feature information [125,126], and performs the same processing on the features of different channels, resulting in the information interaction between different channels ignored. Similarly, channel attention only considers channel feature information, and performs the same processing on spatial features within the same channel, resulting in the information interaction between different spatial features ignored. Hybrid attention is usually composed of spatial attention and channel attention in series, parallel, or fusion improvement, and the advantages of both are fully utilized. To be specific, the parallel hybrid attention sets separate network paths for channel attention and spatial attention. Its advantage is that it fully considers the channel information and spatial information of feature maps. However, the channel and spatial dimension of the parallel hybrid attention do not interact sufficiently with each other, which reduces the critical spatial information within the important channels. Series hybrid attention achieves more reasonable attention weight distribution by sequentially computing the information in the spatial and channel dimensions. However, series hybrid attention destroys the independence of information in spatial and channel dimensions, and it is difficult to extract each channel interaction and spatial information within the channel. The two connection modes are widely used in most image tasks, and the connection method is usually selected according to the characteristics of the image.

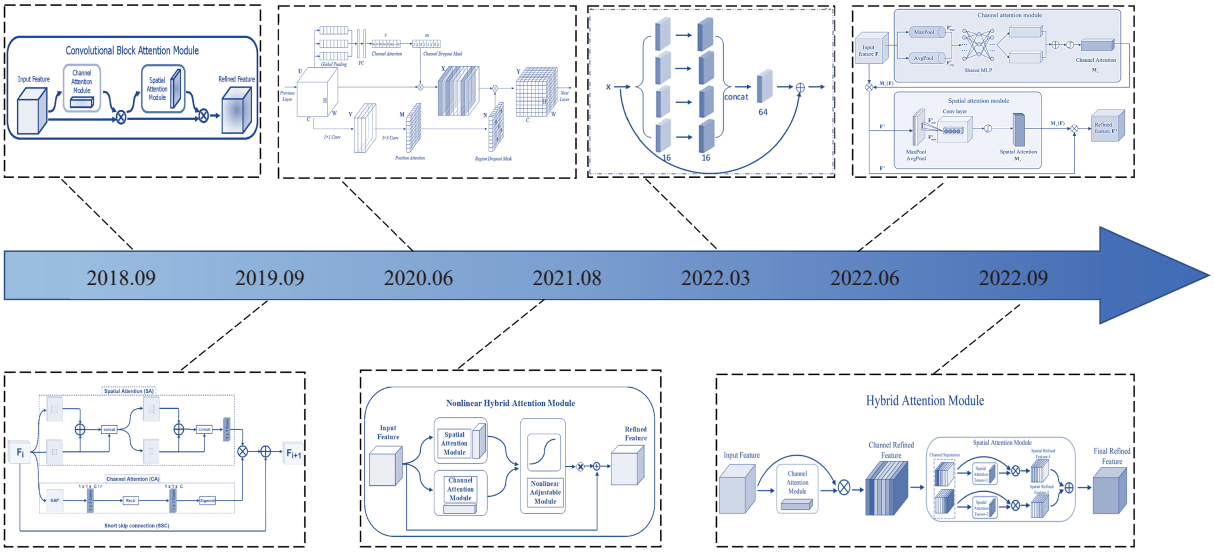


Figure 5. The hybrid attention network structure development.

In the previous section, channel attention has been described in detail and will not be repeated in this section. Next, spatial attention is briefly described. Spatial attention is employed to consider global information in spatial dimensions while assigning higher weights to key locations in task-relevant regions and suppressing redundant information with irrelevant regions. The computation process of spatial attention is shown as follows:

$$\mathrm{M_{sp}} = M \otimes \mathrm{F_{sp}}(\mathrm{M}) \tag{8}$$

where M and $\mathrm{M_{sp}}$ are the input and output features of spatial attention, respectively, and $\mathrm{F_{sp}}(\cdot)$ is the computational processes of spatial attention. Concatenating spatial attention and channel attention is a common hybrid attention method, and the output feature result is determined by the concatenation order of hybrid attention. The calculation process is given as follows:

$$\mathrm{M_h} = \mathrm{F_{sp}}(\mathrm{F_{ch}}(\mathrm{M})) \tag{9}$$

where M and $\mathrm{M_h}$ are the input and output features of hybrid attention, respectively, $\mathrm{F_{ch}}(\cdot)$ is the computational processes of channel attention. The computation of attention is implemented in various ways, such as similarity computation, training nonlinear network layers, global pooling combined with nonlinear network layers, etc. Furthermore, hybrid attention is constructed by parallel connections:

$$\mathrm{M_h} = \mathrm{Concat}\left(\mathrm{F_{sp}}(\mathrm{M}), \mathrm{F_{ch}}(\mathrm{M})\right) \tag{10}$$

where Concat(•) represents the process of feature concatenation. Finally, the fusion improvement strategy is a common way of constructing hybrid attention. Channels and spatial positions on the feature map are regarded as a whole, and hybrid attention is used to assign attention weights to it. The calculation process is given as follows:

$$\mathrm{M_h} = \mathrm{F_h}(\mathrm{M}) \tag{11}$$

where $F_h(\cdot)$ represents the calculation process of mixed attention based on the fusion improvement strategy.

Recently, hybrid attention has been widely developed due to its excellent feature extraction capability. The convolutional block attention module (CBAM) presented by Woo et al. [55] consists of spatial and channel attention in series, which effectively extracts the spatial and channel information of the image. Subsequently, the hybrid residual attention block (HRAB) was suggested by Muqeet et al. [56] that effectively integrates the spatial and channel attention modules in parallel. Liu et al [57] suggested a lightweight attention module called attention discard convolution module (ADCM), and the discard strategies in channel and location attention mechanisms were fully adopted by this module. Furthermore, a nonlinear hybrid attention mechanism (NHAM) was adopted by Guo et al. [58]. The parameters associated with each attentional branch of this hybrid attentional mechanism can be adjusted, which gives the attentional module better flexibility and self-adaptability. Sheng et al. [59] proposed a multi-scale residual attention module (MSRA), which uses null convolution at different scales to extract multi-scale information under different perceptual fields, enhancing the ability of the model to describe contextual information. The channel-space attention mechanism (CSAM) proposed by Zhang et al. [60] was applied to enhance the feature extraction capability of the model. In this module, adjacent spatial and channel attention modules are connected, and channel and spatial features are fused to construct the CSAM. A hybrid attentional module (HAM) was proposed by Li et al. [61]. The channel attention is used to generate channel refinement features, and spatial attention divides channel refinement features along channel axes to generate spatial attention descriptors. The spatial sub-module generates the final refinement features by applying spatial attention descriptors with adaptive emphasis on important regions.

4.2. Hybrid Attention in Medical Image Classification Task

Hybrid attention pays attention to the key regions of medical images and effectively improves the classification performance of the network. Therefore, hybrid attention has been developed in the field of medical image classification. The model based on combining hybrid attention and CNN was proposed by Sun et al. [62] for computer-aided diagnosis of endometrial histopathological images. Hybrid attention consists of positional and channel attention modules in parallel, which are adopted to extract the contextual information of images and the importance of different semantic information, respectively. The improved model achieves the best performance in the classification task of endometrioid adenocarcinoma. The perturbed composite attention model was presented by Mishra et al. [63]. The model consists of two modules: multi-level perturbation spatial attention (MPSA) and multi-dimensional attention (MDA). The perturbation mechanism in MPSA is used to process the spatial features of macular optical coherence tomography (OCT) images, and MDA is employed to process and encode channel information in different dimensions of the deep network. For the problem of breast density classification in mammograms, Zhao et al. [64] proposed a bilateral adaptive space and channel attention network. Adaptive spatial and channel attention modules are used to explore discriminative information for breast density classification, which achieves the best performance on the mentioned public datasets. A dual attention network was proposed by Wei et al. [65] to assist in learning skin lesion classification. This dual attention effectively improves the feature extraction capability of the model and highlights important local patterns in the skin lesion region. Furthermore, the attention module consists of spatial and channel attention modules. Spatial attention pays attention to the features of the skin damage area and reduces irrelevant artifactual features. Meanwhile, the global features of the lesion region are acquired by the channel attention module, which generates the feature channel complex weight vectors to extract the important local pattern features of the lesion region. The three-dimensional CNN with hybrid attention was proposed by Qin et al. [66] for the early diagnosis of Alzheimer's disease. The channel and spatial attention in the hybrid attention are fully utilized to improve the feature extraction capability of the network and also combined with the residual connectivity of the classification network to further improve the diagnostic accuracy of the model. The ISANET based on CNN and hybrid attention was proposed by Xu et al. [67] for the classification of lung cancer. Pathological regions are attended to by models with embedded channels and spatial attention, resulting in the superior performance of the model in classifying lung cancer.

4.3. Hybrid Attention in Medical Image Segmentation Task

Hybrid attention is widely adopted in the field of medical image segmentation due to its excellent channel and spatial feature extraction capabilities. An asymmetric U-Net model with hybrid attention was proposed by Chen et al. [68] for kidney ultrasound image segmentation. Furthermore, the hybrid attention guides the network to focus on regions related to the kidney, and extract more important kidney feature representations to improve the segmentation accuracy. For the lung tumor image segmentation task, Hu et al. [69] proposed a parallel deep learning algorithm with a hybrid attention mechanism. The hybrid attention consisting of spatial and channel attention modules enhances the feature extraction capability of the model while improving the segmentation accuracy. The feature fusion attention network proposed by Yu et al. [70] is used for medical image segmentation, and the similarity of each space and

channel is considered by the hybrid attention to obtain global dependencies and contextual features, which refine the segmentation results of the model in the up-sampling stage. Yang et al. proposed TA-Net, a triple attention network for medical image segmentation, whose attention simultaneously recognizes information in channel, spatial and feature domains. Furthermore, to address the long-range dependence of pixels, a channel self-attentive coding module is used to learn their latent features. To make the model focus on the location information of useful pixels, a spatial attention up-sampling module is used to fuse the network features of different layers. A dual-supervised medical image segmentation network with multi-dimensional hybrid attention was proposed by Xu et al. [71]. The multidimensional self-attention in the network highlights salient features and suppresses irrelevant information through two self-attention modules, and extracts dependencies between features in both channel and spatial dimensions. The medical image segmentation network embedded with hybrid attention was proposed by Chen et al. [72], which reuses inter-channel relations and spatial point features. Meanwhile, by embedding the hybrid attention module into the shrinking and expanding paths, respectively, the feature extraction ability and detail restoration effect of the network are enhanced. An attention mechanism fusion network MDAG-Net for the multi-object segmentation of tumors was proposed by Lu et al. [73]. The attention is able to extract multi-dimensional mixed features of the pancreas and pancreatic tumors and learn contextual information within the U-Net. At the same time, small object features are localized in spatial and channel dimensions to avoid the interference of redundant information in shallow features. In addition, hybrid attention is paid to enhance the feature representation of pancreatic tumors to boost segmentation capabilities.

4.4. Hybrid Attention in Medical Image Detection Task

The hybrid attention mechanism enhances the spatial and channel feature information and suppresses the background information to boost the detection precision of the detection model. Detection of localized lesions identified in the tongue helps determine the progression of the disease and the physical condition of the patient. The improved Faster RCNN based on ResNet50 and hybrid attention was proposed by Chen et al. [74]. The model employs hybrid attention to emphasize local lesion information and suppress background tongue information. Li et al. [75] proposed a detection model based on CNN and hybrid attention. Endoscopic early gastric cancer is irregular in shape and unclear in the boundary. To address this problem, CBAM was proposed to enhance the characteristic expression of tumor targets in the channel and space, highlight tumor information and suppress interference information. Blood cell counts play an important role in the field of clinical medicine diagnosis. To efficiently automate blood cell counting, a method based on Attention-YOLO was proposed by Jiang et al. [76], which was accomplished by introducing channel and spatial attention in the feature extraction network. Attention-YOLO replaces the original feature vectors with filtered and weighted feature vectors for residual fusion to boost the detection accuracy of the network. Most detection models cannot handle the scale variations of various acute pancreatitis lesions, leading to inaccurate detection and occasionally false-positive small lesions in the vicinity of large lesions. To overcome this problem, a dual-attention-based method for detecting acute pancreatitis lesions in CT images was proposed by Zhang et al. [77]. To be specific, channel attention was employed to capture relationships between channels of the feature map to downplay meaningless channels that are not relevant to the lesion, while spatial attention was utilized to motivate the network to pay attention to regions that are more relevant to the lesion. The multi-scale contextual information fusion cascade hybrid attention model was presented by Pan et al. [78] for the detection of nasopharyngeal lesions. Among them, the cascaded hybrid space and channel attention modules aim to transfer attention between the convolutional blocks of the front and back cascades. Simultaneously, the attention between different convolutional modules is fused to strengthen the effective features and inhibit the ineffective ones. A multi-scale CNN with channel and spatial attention was suggested by Zhao et al. [79] for lung nodule detection. A multi-scale feature extraction module and an attention mechanism were inserted into the residual block of the model, which allows better learning of the features of candidate nodes.

4.5. Hybrid Attention in Medical Image Enhancement Task

Hybrid attention has been rapidly developed in the fields of super-resolution reconstruction, image synthesis, and image registration of medical images. Diabetic retinopathy causes vision loss and even blindness in diabetic patients. Therefore, a generative adversarial network (GAN) with multiscale spatial and channel attention modules was presented by Zhou et al. [80]. This network was applied to synthesize high-resolution fundus images with fine-grained lesions to identify and grade the severity of diabetic retinopathy. High-resolution MRI images provide fine anatomical information, but acquiring the data requires long scanning times. Jiang et al. [81] introduced a framework called fused attention GAN for generating super-resolution magnetic resonance (MR) images from low-resolution MR images. Hybrid attention, including channel attention, self-attention, and fusion operations, was designed to strengthen the important features of MR images. The model based on the U-Net structure increases the computa-

tional speed while improving the registration effect. However, feature loss is prone to occur during the up-sampling of such structures. Therefore, an unsupervised generative adversarial network based on a fused dual attention mechanism was presented by Meng et al. [82] for the registration of deformable medical images. During the up-sampling of the proposed model, hybrid attention including both channel and location attention is introduced to boost the feature recovery. A dual attention network for super-resolution reconstruction of lung cancer images was suggested by Zhu et al. [83]. Hybrid attention effectively integrates channel attention and spatial attention, learns the relation between spatial regions and channel pixels, distinguishes important features from insignificant ones, and strengthens the reconstruction of high-frequency information. For the blurred edges and unclear texture of traditional computed tomography (CT) images, a super-resolution network of CT images based on hybrid attention and global feature fusion was presented by Chi et al. [84] for CT image restoration. The hybrid attention mechanism adaptively maps feature information from feature maps at different levels, and the connections between feature maps at different levels are established by the hybrid attention mechanism. Li et al. [85] suggested a residual network with multiple attention fusion for image denoising to enhance the clarity of ultrasound images. The abundant detailed information in ultrasound images is extracted by the model. In addition, important information from multiple feature domains is attended to by merging information from various attentions. Residual connectivity is inserted in the multi-attention fusion attention block to prevent missing important information during the ultrasound image reconstruction. The denoising capability of the network is verified on several public ultrasound datasets.

5. Transformer

In this section, we first introduce the technical development of the transformer. Then, the applications in medical image classification, segmentation, detection, and image enhancement are introduced. These methods are summarized in Table 3.

Table 3 Publication summary of the transformer in medical image analysis

Method	Disease	Organ	Image	Dataset	Highlight	Performance
Classification task						
He et al. [96]	Brain Age	Brain	MRI	BGSP, OASIS-3, NIH-PD, IXIABIDE-I, DLBS, CMI, CoRR	Image-level and patch-level attentive fusion.	-
Li et al. [97]	COVID-19	Lung	CT	CC-CCII,	Transformer combine multi-instance learning	ACC = 98.6%
Wang et al. [98]	Knee Cartilage Defect	Knee	MRI	-	Applying Transformers on 3D medical images	ACC = 70.2%
Verenich et al. [99]	COVID-19, viral pneumonia	Lung	X-ray	COVID-19 Radiology Dataset	Transformer combine CNN.	ACC = 94.0%
Wu et al. [100]	Melanocytic Lesions	Skin	Skin biopsy	MPATH-Dx	Representing multi-scale features with Transformers	ACC = 60.0%
Barhoumi et al. [101]	Intracranial Hemorrhage	Brain	MRI	RSNA intracranial hemorrhage dataset	Multi CNNs as feature generator, feature aggregation by Transformers.	ACC = 98.0%
Segmentation task						
Sun et al. [102]	Brain tissue segmentation	Brain	MRI	a.MRBrainS b.iSeg-2017	Multi-path strategy for multi-modal fusion	a. Dice = 83.47% b. Dice = 87.16%
Hatamizadeh et al. [103]	Brain tumor segmentation	Brain	MRI	BraTS2021 dataset	Swin Transformer	Dice = 91.3%
Gao et al. [104]	Cardiac segmentation	Cardiac	MRI	M&Ms	Hybrid Transformer combine CNN	Dice = 88.3%
Ji et al. [105]	a. Cell segmentation b. Polyp segmentation c. Skin lesion segmentation	a. Cell b. PolyP c. Skin	a. Pathology, b. Colonoscopy, c. Dermoscopy	a. Pannuke b. KCCEE c. ISIC2018	Multi-scale feature	a. Dice = 68.40% b. Dice = 92.30% c. Dice = 90.35%
Fu et al. [106]	a. Cardiac segmentation b. Multi-organ segmentation	a. Cardiac b. Multi-organ	a. MRI b. CT	a. ACDC b. Synapse multi-organ CT	Transformer combine U-Net	a. Dice = 91.72% b. Dice = 85.46%
Xie et al. [107]	Multi-organ segmentation	Multi-organ	CT	BCV dataset	Multi-scale feature, deformable Transformer combine CNN	Dice = 85.0%
Detection task						
Ma et al. [108]	Coronary arteries significant stenosis	Coronary arteries	MPR	-	Transformer combine CNN	ACC = 92.0%
Jiang et al. [109]	Caries	Tooth	Caries images	-	Transformer for feature extraction	mAP = 56.9%
Kong et al. [110]	Chest abnormality detection	Chest	X-ray	a. inbig Chest X-Ray dataset b. ChestX Det-10 dataset	Context-aware feature extractor, deformable Transformer detector	a. AP50 = 36.3 b. AP50 = 43.6
Tao et al. [111]	Vertebra detection and localization	Spine	CT	a. VerSe 2019 challenge b. MICCAI-CSI 2014 challenge	Spine-Transformer	a. Id-Rate = 97.22% b. Id-Rate = 92.2%
Chen et al. [112]	a. retina disease b. hemorrhage c. brain tumors	a. Retina b. Head c. Brain	a. OCT b. CT c. MRI	a. Retinal- OCT b. Head-CT c. Brain-MRI	Multi-scale feature	a. image-level AUROC = 98.38% b. image-level AUROC = 93.00% c. image-level AUROC = 95.81%
Enhancement task						
Feng et al. [113]	Brain MRI reconstruction & super-resolution	Brain	MRI	a. IXI b. Clinical dataset	Transformer combine multi-task	a. PSNR = 29.397 (scale = 2x) b. PSNR = 30.400 (scale = 2x)
Wang et al. [114]	Liver lesions CT denoising	Liver	CT	2016 NIH-AAPM-Mayo clinic LDCT Grand Challenge dataset	Symmetric token-to-token Transformer	SSIM = 0.9144
Korkmaz et al. [115]	Brain MRI reconstruct	Brain	MRI	a. IXI b. fastMRI	Cross-attention Transformers	a. PSNR = 30.7 (MRI T2- > T1) b. PSNR = 33.4 (MRI T2- > T1)
Shin et al. [116]	Brain image synthesis (MRI->PET)	Brain	MRI, PET	ADNI	Transformer combine GAN	PSNR = 47.52
Song et al. [117]	Brain image registration	Brain	MRI	OASIS	Transformer combine CNN	Dice = 74.16%

5.1. Transformer Development and Technical Description

Transformer [86] has dominated the field of NLP and made great contributions to many tasks, including speech translation [87], speech synthesis [88], and language generation [89]. Inspired via the success of Transformers in the NLP field, a number of attempts have been made to apply Transformers to visual tasks in the computer vision (CV) field, the most iconic of which are the detection transformer (DETR) [90], vision transformers (ViT) [91], data-efficient image transformers (DeiT) [92], pyramid vision transformer (PVT) [93], swin-transformer [94], and deformable attention transformer (DAT) [95]. Figure 6 briefly summarizes the representative transformer-based works in the deep learning aera.

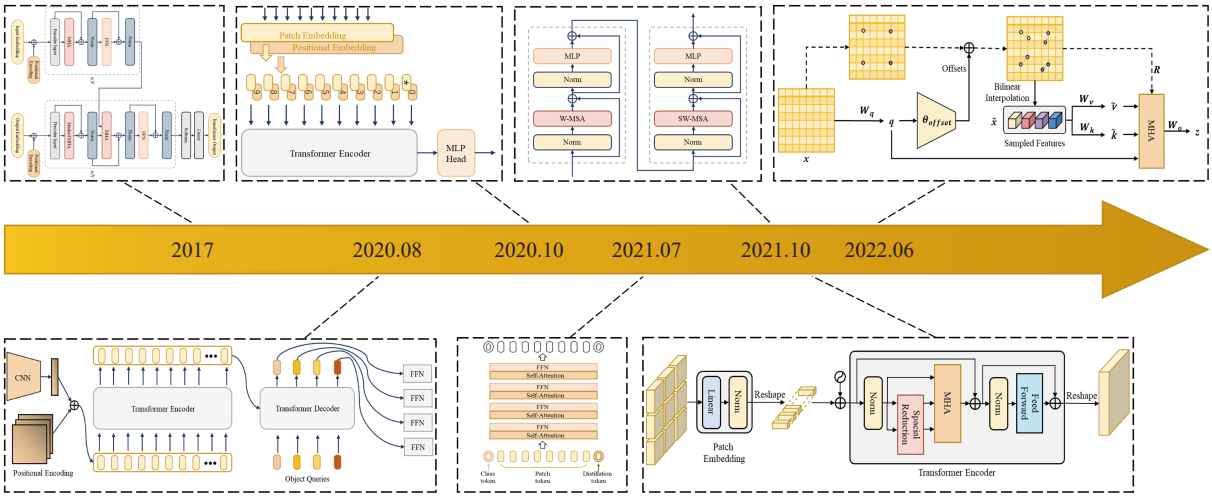


Figure 6. The transformer network structure development.

Transformer: The typical transformer with an encoder-decoder structure is free of convolution and only consists of the multi-head attention mechanism. In [86], the authors found that hierarchical features of the input can be better captured through multiple self-attentions, where a multi-head attention (MHA) mechanism was defined by concatenating the calculated attentions:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{12}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O \tag{13}$$

where W_i^Q, W_i^K, W_i^V, W^O denote the weights of the linear transformation matrix, which map Q, K, V, and calculated attention to different subspaces, respectively. The encoder and decoder in the transformer are concatenations of consecutive identical blocks. The blocks in the encoder are mainly composed of two parts and both parts use the residual connection, where the bottom part employs multi-head attention and layer normalization, and the top part employs a fully connected feed-forward network (which consists of two fully connected layers) and layer normalization. In a similar way to the encoder with some minor adaptations, the blocks in the decoder consist of three parts. Specifically, the bottom part has a similar structure to that of the encoder, except for the masked MHA; the middle part is similar to the bottom part of the encoder, which takes the encoder output as the input; and for the top part, its structure is the same as the corresponding part of the encoder. The residual connections are also applied to these three parts in the decoder.

DETR: DETR was proposed by Carion et al. [90] in 2020 to solve the task of object detection, where transformers were introduced to the CV field. DETR mainly consists of a backbone, transformer, and prediction heads. The input images first go through the backbone network for feature extraction. Then, a transformer encoder and decoder are, respectively, adopted to model the correlations between the features and generated object queries. Followed by a fully connected feed-forward network, the classes and generates suitable boxes for targets are identified.

ViT: ViT basically follows the design of the vanilla transformer and can be applied to solve large-scale image recognition tasks [91]. In ViT, the input image passes through the linear projection, transformer encoder, and multi-layer perception (MLP) three modules, and the class of image is predicted finally. Specifically, the image is divided into a series of patches, and a position embedding vector is constructed to represent the position information of these patches. Then, a learnable class token and patches are fed into the transformer encoder to learn patch embedding. By following this, an MLP is adopted to perform classification tasks based on the learned class token.

DeiT: The limitation of ViT is that it requires expensive training costs to obtain ideal generalization ability. For

DeiT, it does not need a large amount of dataset or computing resources to obtain classification performance as compared to (CNNs). Based on ViT, Touvron et al. [92] introduced knowledge distillation into the transformer and adopted a generic teacher-student framework for knowledge distillation, wherein a distillation token was added to allow the student to learn according to the knowledge distilled by the teacher.

PVT: To achieve better dense prediction, Wang et al. [93] introduced the pyramid structure into the transformer and designed PVT. Unlike ViT, which controls the size of patches to 16×16, PVT converts the images to a series of patches with smaller sizes and uses a progressive shrinking strategy to control the sizes of the output feature maps at different stages. Since PVT needs to deal with high-resolution features, the spatial-reduction attention (SRA) layer is designed to take place of the conventional MHA layer in the transformer encoder, which can reduce the computational complexity.

Swin-Transformer: In order to deal with the scale variations of visual entities in scene understanding tasks, Liu et al. [94] proposed a Swin-Transformer with a hierarchical design that includes shifted window operations. The shifted window operations contain two windows, namely, non-overlapping local windows and overlapping cross windows, both of which can restrict the computation of attention in local windows and model the cross-window relationships. Moreover, these shifted window operations introduce the locality of convolution operations and save computation.

DAT: Due to the use of sparse attention, both PVT and Swin-Transformer may have limitations on the long-range relations modeling. To mitigate the impact of these limitations on the classification or dense prediction tasks, Xia et al. [95] proposed DAT. The deformable self-attention module serves as the basis for DAT to inherit the idea of deformable convolution and therefore can pay more attention to the important regions in the feature maps than normal convolution layers. Under the guidance of offsets, the module can effectively model the correlations among tokens. In DAT, the vanilla MHA layers in the transformer encoder are replaced by deformable attention, which can be applied to various vision tasks that require multi-scale feature maps.

5.2. Transformer in Medical Image Classification Task

Classification, as one mainstream task in medical image analysis, is a process that classifies the given medical data into specific categories to assist in disease diagnosis or prognosis. CNN-based methods have been widely developed for this task and achieved superior results. Recently, transformer-based methods have been investigated in classification by many researchers. The work of He et al. [96] introduced a global-local transformer to perform rapid brain age assessment based on brain MRI. The global-local transformer mainly contains two flowing ways for features, namely, the global-pathway and the local-pathway. The former is to perform feature extraction from the input MRI, which can better focus on the global-context information. The latter is to capture the local fine granularity information from the local patches. Then, the global- and local-context information is fused via a global-local attention mechanism, followed by a revised global-local transformer to estimate brain age. Li et al. [97] combined multi-instance learning (MIL) with a transformer to model the global relations within the data. They proposed an explainable MIL strategy to consider the relation between local features and class representation, in which a transformer MIL pooling layer was applied to the instance-level feature maps produced by CNN to obtain embedding-level feature maps and bag representation. Thus, the interferences of false positive instances were reduced. Their work also explored the explainability of the deep learning-based method. They endowed the model with some interpretability by the attention mechanism in the explainable MIL strategy. Wang et al. [98] designed a 3D transformer to explore the application of a transformer on 3D medical images. They adopted 3D convolutional operations to extract features from 3D patches, instead of linear embedding. A teacher-student framework for knowledge distillation was used to learn the parameters of the transformer from a CNN teacher, which addressed the data-hungry issue. Verenich et al. [99] considered that ViTs lack translation invariance and equivariance. The CNNs have these two properties which are particularly useful for the tasks of detecting manifested abnormalities. As a result, they introduced ViT to CNN architectures, intending to combine the advantages of both, which enables the model to maintain spatial invariance and equivariance while extracting global correlations among salient features. Wu et al. [100] proposed the ScATNet to learn multi-scale representations for melanocytic skin lesions. The ScATNet has two transformers, in which one is used to learn inter-patch representations based on the local patch-wise embeddings captured by a CNN, and another is used to learn inter-scale representations from the multi-scale patch embeddings. Thus, the ScATNet can weigh multiple scale representations and automatically distinguish whether the information is relevant to the diagnosis or not. Barhoumi et al. [101] took an ensemble backbone consisting of several CNNs as a feature generator, which generates middle features as the inputs of ViTs. They stacked the features extracted by multiple CNNs and used the stacked features as the input of a ViT. Through this operation, the input features of ViT can be made more abundant than that extracted by a single CNN, claiming that 1) the proposed method can make the ViT focus on relevant features at multiple scales; and 2) ViT attends to calculate global attention among patches as the feature content gradu-

ally enriches.

5.3. Transformer in Medical Image Segmentation Task

For mainstream segmentation task in medical image analysis, there are also a number of transformer-based methods applied to various segmentation tasks. Sun et al. [102] proposed the HybridCTrm network for multi-modality brain segmentation based on transformers and CNNs. In this work, two different strategies were presented, including a single-path strategy and a multi-path strategy. In the single-path strategy, two input modalities MRI-T1 and MRI-T2 were combined as a multi-channel image and then fed into several convolution blocks and transformers for encoding and subsequent decoding. The single-path strategy mainly focuses on the interactions between different modalities. The multi-path strategy is designed to combine and utilize the information or features extracted from those modalities, which has a similar structure to the single-path strategy except for the different input. Inspired by the success of Swin-Transformer, Hatamizadeh et al. [103] presented a segmentation model termed Swin UNETR for 3D brain tumor segmentation. Swin UNETR is a U-shaped network design, wherein Swin-Transformer and CNN are respectively used as the encoder and decoder. Specifically, they reformulate the semantic segmentation as a sequence-to-sequence prediction, in which a hierarchical Swin-Transformer as the encoder is used to encode the 1D sequence of embedding. Feature maps with different resolutions were extracted by the Swin-Transformer encoder and connected to each resolution of the CNN-based decoder by skip connections. Gao et al. [104] proposed a hybrid transformer architecture termed UTNet that introduces self-attention into a UNet for the segmentation of enhancing the medical image. To model the remote dependency at different scales with minimum expenses, both the encoder and decoder in UTNet apply the self-attention mechanism. With the relative position encoding, the proposed self-attention modules can effectively model the global relations between different scales and reduce the complexity from $O(n^2)$ to approximate O(n). Ji et al. [105] presented a unified transformer network termed MCTrans to consider the dependencies of different pixels at different scales, the consistency of different feature representations at specific regions, and the correspondence of different labels. In MCTrans, multi-scale feature maps are embedded as a series of tokens, and both intra-scale and inter-scale self-attention are performed to construct the cross-scale dependencies. At the same time, a learnable embedding is introduced, and self-attention and cross-attention are used to model semantic relations and enhance feature representations. As a result, MCTrans is a unified framework that incorporates feature representation learning and semantic structure mining. Fu et al. [106] proposed a fully automated segmentation method that combines U-Net and transformer, wherein the transformer blocks replace the convolutional layers in the encoder and decoder of vanilla U-Net. In this work, CNN was used to extract features and encode spatial information of inputs, and a transformer was used for adding long-range dependencies to deep features and modeling multi-scale features. Xie et al. [107] proposed a 3D medical image segmentation framework termed CoTr that efficiently bridges a transformer and a CNN. In CoTr, a deformable transformer is designed to construct the remote dependencies on the features extracted by a CNN. The deformable transformer introduces the deformable self-attention mechanism to focus on key positions of feature maps.

5.4. Transformer in Medical Image Detection Task

For detection tasks in medical image analysis, the number of transformer-based methods is limited. Most of them adopt both transformers and CNN blocks as the main components of models, wherein the CNN performs feature extraction based on medical data and the transformer performs feature enhancement for the downstream detection task. Ma et al. [108] proposed the TR-Net that incorporates both CNN and transformer, and such a TR-Net was used for automatically detecting significant stenosis and completing the diagnosis of coronary artery disease. In this work, a shallow 3D CNN was used to efficiently perform local semantic feature extraction from coronary arteries. Then, these extracted semantic features of cubic volume were fed into a transformer, and the transformer can learn the correlations between features in each region of a multiplanar reformatted (MPR) image. Thus, based on local and global image information, significant stenosis in coronary arteries can be accurately detected by TR-Net. Jiang et al. [109] presented a caries detection framework based on YOLOv5s, where the transformer was incorporated in the backbone network to perform feature extraction from input data, and the FReLU activation function was adopted to the activation of visual-spatial information, which can achieve the accurate and efficient detection of caries. Kong et al. [110] proposed an end-to-end abnormality detection model for chest X-ray images termed CT-CAD, which contains context-aware transformers. They encoded context information at multiple scales and enlarged receptive fields through a context-aware feature extractor. Then, classification and location regression were performed by a deformable transformer detector. The deformable transformer blocks deal with a small set of key sampling points, which enables CT-CAD to focus on the feature subspace and speed up the convergence. Tao et al. [111] presented a 3D object detection method based on transformers to achieve auto-detection and location of vertebrae in arbitrary field-of-view (FOV) Spine CT, and suggested that auto-detection of vertebrae in arbitrary FOV CT can be seen as a

pairwise set prediction problem. They proposed a Spine-Transformer to model the relationships of vertebrae at different levels and to construct global context information. Based on Spine-Transformer, all positions of vertebrae can be obtained in parallel. Chen et al. [112] introduced a transformer-based anomaly detection framework termed UTRAD. In UTRAD, the word tokens were the pre-trained features that are generated by transformer-based autoencoders. A multi-scale pyramid hierarchy was incorporated into UTRAD to detect the structural and non-structural anomalies at multiple scales, and the residual connections were applied in the pyramid hierarchy.

5.5. Transformer in Medical Image Enhancement Task

Transformer models have also been verified to have powerful learning capabilities in various medical image enhancement tasks, including image super-resolution [113], denoising [114], reconstruction [115], synthesis [116], and registration [117]. Feng et al. [113] introduced an end-to-end task transformer network termed T2Net, allowing feature representations to be shared and transferred between MRI reconstruction and super-resolution tasks. They incorporated MRI reconstruction and super-resolution tasks into T2Net, and used two sub-branches to output the results of these tasks respectively. To promote joint learning and exchanging task-related information between two tasks, the task-specific features were extracted by two separate branches. After that, those extracted task-specific features were expressed as queries and keys, and a task transformer module took these features as input to calculate the relationship between the tasks. The work of Wang et al. [114] was the first one to introduce the transformer on low dose CT (LDCT) denoising task, which enlarged the categories of LDCT denoising algorithms. Wang et al. [114] proposed a TED-net, which is an encoder-decoder dilation network based on ViT and free of convolution. TED-net has a symmetric U-shaped architecture and its encoder-decoder blocks consist solely of the transformer. The TED-net has been verified that it can accomplish superior denoising results. Korkmaz et al. [115] presented an unsupervised reconstruction method termed SLATER to perform MRI reconstruction, which incorporated zero-shot learning and adversarial learning into transformers. SLATER is an unconditional deep adversarial network with cross-attention transformers, which consists of a synthesizer, a mapper, and a discriminator. The cross-attention transformers, as the key component of the synthesizer, are designed to capture long-range spatial dependencies without computational burden. Shin et al. [116] combined generative adversarial network (GAN) with BERT to perform PET synthesizing from MRI images. They supposed that synthesizing PET from MRI can benefit from the self-attention over these two modality images on the premise of sufficient training samples, and thus they introduced BERT to consider and calculate the self-attention over these two modalities for better image synthesis. Song et al. [117] combined transformer and CNN, and proposed a U-shaped hybrid network termed TD-Net. In TD-Net, a CNN is used for feature extraction from MRI images, and a transformer encoder is employed to capture global information, wherein the encoding features are concatenated to the corresponding decoding blocks. Thus, the decoding process can utilize the CNN to fuse the local and global information extracted by a transformer.

6. Key Challenges and Potential Solutions

(1) Deep learning models are often regarded as black-box models, and the main users of medical image analysis models are doctors, most of whom have no engineering research backgrounds [118, 119]. Interpretable deep learning models are more friendly to doctors in clinical applications. Existing deep learning attention methods can provide interpretability with limitations, namely visual interpretation. These methods show the important parts of the model decisions in the form of saliency mapping. For example, Seo et al. [120] proposed a multiscale 3D super-voxel model for Alzheimer's disease (AD) diagnosis, and its significance map explained the regional information of the model in discriminating Alzheimer's diseased patients and normal human brains. However, most of the existing deep learning attention models only provide qualitative interpretable analysis, which only answers the question "where does the model focus on?". These methods lack quantitative interpretable research and do not answer the question "how much does the model focus on?". Therefore, a potential research direction is the quantitative deep learning attention models for medical image analysis.

(2) The execution process of the deep learning model requires a large number of computing resources, such as advanced GPU devices[121]. Since radiologists can't always use expensive high-performance computing equipment during the diagnosis process, the lightweight models are a potential research direction. At present, the common design idea is to use the attention module to replace multiple groups of convolutional layers in the model, and use the attention module to improve the performance. For example, Zhao et al. [122] designed a lightweight feature extraction module combining deep separable convolution, residual connection, and attention mechanism, which replaced the convolution layer with high computational demand and was successfully applied to the medical image segmentation task. Unfortunately, this design approach only focuses on the deletion or replacement of part layers of the model, without considering the model as a whole. Distillation learning is expected to achieve the lightweight design of the overall model, which uses the knowledge learned from the large model to guide the training of the small model so

that (a) the small model has the same performance as the large model; and (b) the number of parameters is greatly reduced, achieving model compression and acceleration.

(3) There are various types of medical images, including computed tomography (CT), ultrasound (US), magnetic resonance imaging (MRI), optical coherence tomography (OCT), etc. In clinical diagnosis, doctors may need to make judgments based on a variety of data. The attention model is a good way to analyze the multi-modality data, which relies on its powerful screening and weighting functions for the multi-modality data. Therefore, multi-modality medical image analysis based on deep learning attention is also a potential research direction.

(4) The fourth potential research direction is to use the prior knowledge of doctors to improve the performance of deep learning models [123]. In the existing models, prior knowledge is summarized and injected into the model once and for all. For example, Saha et al. [124] proposed a multi-stage 3D prostate cancer diagnostic model in which the deep attention network detects lesion structures and then probabilistic anatomical priors encode the spatial prevalence and region of cancer. However, in the actual clinical process, the doctor's clinical knowledge will change with the progress of diagnosis. The method based on human-in-the-loop can realize multiple injections of prior knowledge. The essence of the human-in-the-loop is an interactive human-machine model. The prior knowledge of doctors can be injected into the deep learning network in batches, and each injection will update the model to improve the performance. Each prior knowledge injection is an effective introduction to human judgment, which can be summarized as adding the right knowledge at the right time.

7. Conclusion

Deep learning technology and attention mechanism have demonstrated powerful performance in medical image analysis. This paper has reviewed the application of existing deep attention networks in medical image analysis. We have introduced the technical details of the channel attention network, hybrid attention network, and transformer according to the categories of attention. Then, recent methods have been introduced about the applications of these techniques in solving the medical image classification task, segmentation task, detection task, and enhancement task. The remaining challenges and potential directions have been also discussed, including interpretability, lightweight network design, multimodal learning, and prior knowledge.

Author Contributions: **Xiang Li:** Investigation, data curation and writing—original draft preparation; **Yuchen Jiang:** data curation; **Pengfei Yan:** writing—original draft preparation; **Guanyi Li:** writing—original draft preparation; **Minglei Li:** writing—review and editing; **Shen Yin:** writing—review and editing; **Hao Luo:** supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Longoni, C.; Bonezzi, A.; Morewedge, C. K. Resistance to medical artificial intelligence. *J. Consum. Res.*, **2019**, *46*: 629−650.
2. Li, X.; Jiang, Y.C.; Rodriguez-Andina, J. J.; *et al*. When medical images meet generative adversarial network: Recent development and research opportunities. *Discov. Artif. Intell.*, **2021**, *1*: 5.
3. Li, X.; Jiang, Y.C.; Li, M.L.; et al. MSFR‐Net: Multi‐modality and single‐modality feature recalibration network for brain tumor segmentation. *Med. Phys*. **2022**, in press. doi: 10.1002/mp.15933
4. Chen, L.; Zhao, L.; Chen, C.Y.C. Enhancing adversarial defense for medical image analysis systems with pruning and attention mechanism. *Med. Phys.*, **2021**, *48*: 6198−6212.
5. Wang, Q.C.; Wu, T.Y.; Zheng, H.; et al. Hierarchical pyramid diverse attention networks for face recognition. In *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020*; IEEE: Seattle, USA, 2020; pp. 8326–8335. doi: 10.1109/CVPR42600.2020.00835
6. Wang, X.; Lv, R.R.; Zhao, Y.; et al. Multi-scale context aggregation network with attention-guided for crowd counting. In *Proceedings of 2020 15th IEEE International Conference on Signal Processing, Beijing, China, 6–9 December 2020*; IEEE: Beijing, China, 2020; pp. 240–245. doi: 10.1109/ICSP48669.2020.9321067
7. Li, X.; Jiang, Y.C.; Zhang, J.S.; *et al*. Lesion-attention pyramid network for diabetic retinopathy grading. *Artif. Intell. Med.*, **2022**, *126*: 102259.
8. Li, X.; Jiang, Y.C.; Liu, Y.L.; *et al*. RAGCN: Region aggregation graph convolutional network for bone age assessment from x-ray images. *IEEE Trans. Instrum. Meas.*, **2022**, *71*: 4006412.
9. Corbetta, M.; Shulman, G. L. Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.*, **2002**, *3*: 201−215.
10. Rensink, R. A. The dynamic representation of scenes. *Visual Cognit.*, **2000**, *7*: 17−42.
11. Noton, D.; Stark, L. Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision Res*. **1971**, *11*, 929–942, IN3–IN8. doi: 10.1016/0042-6989(71)90213-6
12. Hayhoe, M.; Ballard, D. Eye movements in natural behavior. *Trends Cogn. Sci.*, **2005**, *9*: 188−194.
13. Ahmad, S. VISIT: A neural model of covert visual attention. In *Proceedings of the 4th International Conference on Neural Infor-*

mation Processing Systems, Denver Colorado, 2–5 December 1991; Morgan Kaufmann Publishers Inc.: Denver, USA, 1991; pp. 420–427.
14. Zhang, W.; Yang, H.Y.; Samaras, D.; et al. A computational model of eye movements during object class detection. In *Proceedings of the 18th International Conference on Neural Information Processing Systems, Vancouver British Columbia Canada, 5–8 December 2005*; MIT Press: Vancouver British, Canada, 2005; pp. 1609–1616.
15. Larochelle, H.; Hinton, G. Learning to combine foveal glimpses with a third-order Boltzmann machine. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems, Vancouver British Columbia Canada, 6–9 December* 2010; Curran Associates Inc.: Vancouver British, Canada, 2010; pp. 1243–1251.
16. Bazzani, L.; de Freitas, N.; Larochelle, H.; et al. Learning attentional policies for tracking and recognition in video with deep networks. In *Proceedings of the 28th International Conference on Machine Learning, Bellevue Washington USA, 28 June 2011–2 July 2011*; Omnipress: Bellevue, USA, 2011; pp. 937–944.
17. Fukushima, K. Neural network model for selective attention in visual pattern recognition and associative recall. *Appl. Opt.*, **1987**, *26*: 4985−4992.
18. Milanese, R.; Wechsler, H.; Gill, S.; et al. Integration of bottom-up and top-down cues for visual attention using non-linear relaxation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 21–23 June 1994*; IEEE: Seattle, USA, 1994; pp. 781–785. doi: 10.1109/CVPR.1994.323898
19. Paletta, L.; Fritz, G.; Seifert, C. Q-learning of sequential attention for visual object recognition from informative local descriptors. In *Proceedings of the 22nd International Conference on Machine Learning, Bonn Germany, 7–11 August 2005*; ACM: Bonn, Germany, 2005; pp. 649–656. doi: 10.1145/1102351.1102433
20. Postma, E. O.; van den Herik, H.J.; Hudson, P.T.W. SCAN: A scalable model of attentional selection. *Neural Networks*, **1997**, *10*: 993−1015.
21. Stollenga, M.F.; Masci, J.; Gomez, F.; et al. Deep networks with internal selective attention through feedback connections. In *Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, Canada, 8–13 December 2014*; MIT Press: Montreal, Canada, 2014; pp. 3545–3553.
22. Salah, A.A.; Alpaydin, E.; Akarun, L. A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, **2002**, *24*: 420−425.
23. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015*; San Diego, 2015. http://arxiv.org/abs/1409.0473(accessed on 10 October 2022).
24. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 4–9 December 2017*; Curran Associates Inc.: Long Beach, USA, 2017; pp. 6000–6010.
25. Li, X.; Jiang, Y.C.; Li, M.L.; *et al*. Lightweight attention convolutional neural network for retinal vessel image segmentation. *IEEE Trans. Industr. Inform.*, **2021**, *17*: 1958−1967.
26. Xiao, T.J.; Xu, Y.C.; Yang, K.Y.; et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015*; IEEE: Boston, 2015; pp. 842–850. doi: 10.1109/CVPR.2015.7298685
27. Song, S.J.; Lan, C.L.; Xing, J.L.; et al. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, California, USA, 4–9 February 2017*; AAAI Press: San Francisco, USA, 2017; pp. 4263–4270.
28. Xie, Q.; Lai, Y.K.; Wu, J.; et al. MLCVNet: Multi-level context VoteNet for 3D object detection. In *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020*; IEEE: Seattle, USA, 2020; pp. 10447–10456. doi: 10.1109/CVPR42600.2020.01046
29. Hu, X.W.; Yu, L.Q.; Chen, H.; et al. AGNet: Attention-guided network for surgical tool presence detection. In *Proceedings of the 3rd International Workshop on Deep Learning in Medical Image Analysis, Québec City, QC, Canada, 14 September 2017*; Springer: Québec City, Canada, 2017; pp. 186–194. doi: 10.1007/978-3-319-67558-9_22
30. Nie, D.; Gao, Y.Z.; Wang, L.; et al. ASDNet: Attention based semi-supervised deep networks for medical image segmentation. In *Proceedings of the 21st International Conference on Medical Image Computing and Computer-Assisted Intervention, Granada, Spain, 16–20 September* 2018; Springer: Granada, Spain, 2018; pp. 370–378. doi: 10.1007/978-3-030-00937-3_43
31. Xiao, X.; Lian, S.; Luo, Z.M.; et al. Weighted res-UNet for high-quality retina vessel segmentation. In *Proceedings of the 9th International Conference on Information Technology in Medicine and Education, Hangzhou, China, 19–21 October 2018*; IEEE: Hangzhou, China, 2018; pp. 327–331. doi: 10.1109/ITME.2018.00080
32. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: Salt Lake City, USA, 2018; pp. 7132–7141. doi: 10.1109/CVPR.2018.00745
33. Gao, Z.L.; Xie, J.T.; Wang, Q.L.; et al. Global second-order pooling convolutional networks. In *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019*; IEEE: Long Beach, USA, 2019; pp. 3024–3033. doi: 10.1109/CVPR.2019.00314
34. Zhang, H.; Dana, K.; Shi, J.P.; et al. Context encoding for semantic segmentation. In *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: Salt Lake City, USA, 2018; pp. 7151–7160. doi: 10.1109/CVPR.2018.00747
35. Wang, Q.L.; Wu, B.G.; Zhu, P.F.; et al. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020*; IEEE: Seattle, USA, 2020; pp. 11531–11539. doi: 10.1109/CVPR42600.2020.01155
36. Lee, H.; Kim, H.E.; Nam, H. SRM: A style-based recalibration module for convolutional neural networks. In *Proceedings of 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), 27 October 2019-2 November 2019*; IEEE: Seoul, Korea (South), 2019; pp. 1854–1862. doi: 10.1109/ICCV.2019.00194
37. Yang, Z.X.; Zhu, L.C.; Wu, Y.; et al. Gated channel transformation for visual recognition. In *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020*; IEEE: Seattle, USA, 2020; pp. 11794–11803. doi: 10.1109/CVPR42600.2020.01181
38. Qin, Z.Q.; Zhang, P.Y.; Wu, F.; et al. FcaNet: Frequency channel attention networks. In *Proceedings of 2021 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021*; IEEE: Montreal, Canada, 2021; pp. 783–792. doi: 10.1109/ICCV48922.2021.00082

39. Guo, X.Q.; Yuan, Y. X. Semi-supervised WCE image classification with adaptive aggregated attention. *Med. Image Anal.*, **2020**, *64*: 101733.
40. Chen, H.Y.; Li, C.; Li, X.Y.; *et al*. IL-MCAM: An interactive learning and multi-channel attention mechanism-based weakly supervised colorectal histopathology image classification approach. *Comput. Biol. Med.*, **2022**, *143*: 105265.
41. Yao, H.D.; Zhang, X.J.; Zhou, X.B.; *et al*. Parallel structure deep neural network using CNN and RNN with an attention mechanism for breast cancer histology image classification. *Cancers*, **2019**, *11*: 1901.
42. Shaik, N.S.; Cherukuri, T. K. Multi-level attention network: Application to brain tumor classification. *Signal Image Video Process.*, **2022**, *16*: 817−824.
43. Wang, Z.K.; Zou, Y.N.; Liu, P. X. Hybrid dilation and attention residual U-Net for medical image segmentation. *Comput. Biol. Med.*, **2021**, *134*: 104449.
44. Sinha, A.; Dolz, J. Multi-scale self-guided attention for medical image segmentation. *IEEE J. Biomed. Health Inform.*, **2021**, *25*: 121−130.
45. Gu, R.; Wang, G.T.; Song, T.; *et al*. CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE Trans. Med. Imaging*, **2021**, *40*: 699−711.
46. Li, Y.; Yang, J.; Ni, J.J.; *et al*. TA-Net: Triple attention network for medical image segmentation. *Comput. Biol. Med.*, **2021**, *137*: 104836.
47. Ni, J.J.; Wu, J.H.; Wang, H.Y.; *et al*. Global channel attention networks for intracranial vessel segmentation. *Comput. Biol. Med.*, **2020**, *118*: 103639.
48. Fan, Y.Q.; Liu, J.H.; Yao, R.X.; *et al*. COVID-19 detection from X-ray images using multi-kernel-size spatial-channel attention network. *Pattern Recognit.*, **2021**, *119*: 108055.
49. Guo, C.L.; Szemenyei, M.; Hu, Y.T.; et al. Channel attention residual U-net for retinal vessel segmentation. In *Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, ON, Canada, 6–11 June 2021*; IEEE: Toronto, Canada, 2021; pp. 1185–1189. doi: 10.1109/ICASSP39728.2021.9414282
50. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; et al. ResUNet++: An advanced architecture for medical image segmentation. In *Proceedings of 2019 IEEE International Symposium on Multimedia* (*ISM*), *San Diego, CA, USA, 9–11 December 2019*; IEEE: San Diego, CA, USA, 2019; pp. 225–2255. doi: 10.1109/ISM46123.2019.00049
51. Sharif, S.M.A.; Naqvi, R.A.; Biswas, M. Learning medical image denoising with deep dynamic residual attention network. *Mathematics*, **2020**, *8*: 2192.
52. Rahman, T.; Bilgin, A.; Cabrera, S. Asymmetric decoder design for efficient convolutional encoder-decoder architectures in medical image reconstruction. In *Proceedings of the Multimodal Biomedical Imaging XV II 2022, San Francisco, United States, 22 Jan 2022-27 Jan 2022*; SPIE: San Francisco, USA, 2022; pp. 7–14.
53. Yin, X.W.; Qian, W.H.; Xu, D.; et al. An unsupervised dual attention method for 3D medical image registration. In *Proceedings of 2021 7th International Conference on Computer and Communications, Chengdu, China, 10–13 December 2021*; IEEE: Chengdu, China, 2021; pp. 975–979. doi: 10.1109/ICCC54389.2021.9674730
54. Gu, Y.C.; Zeng, Z.T.; Chen, H.B.; *et al*. MedSRGAN: Medical images super-resolution using generative adversarial networks. *Multimed. Tools Appl.*, **2020**, *79*: 21815−21840.
55. Woo, S.; Park, J.; Lee, J.Y.; et al. CBAM: Convolutional block attention module. In *Proceedings of the 15th European Conference on Computer Vision, Munich, Germany, 8–14 September 2018*; Springer: Munich, Germany, 2018; pp. 3–19. doi: 10.1007/978-3-030-01234-2_1
56. Muqeet, A.; Iqbal, M.T.B.; Bae, S. H. HRAN: Hybrid residual attention network for single image super-resolution. *IEEE Access*, **2019**, *7*: 137020−137029.
57. Liu, Z.G.; Du, J.; Wang, M.; *et al*. ADCM: Attention dropout convolutional module. *Neurocomputing*, **2020**, *394*: 95−104.
58. Guo, N.; Gu, K.; Qiao, J.F.; *et al*. Improved deep CNNs based on Nonlinear Hybrid Attention Module for image classification. *Neural Networks*, **2021**, *140*: 158−166.
59. Sheng, J.C.; Lv, G.Q.; Du, G.; *et al*. Multi-scale residual attention network for single image dehazing. *Digital Signal Process.*, **2022**, *121*: 103327.
60. Zhang, S.; Liu, Z.W.; Chen, Y.P.; et al. Selective kernel convolution deep residual network based on channel-spatial attention mechanism and feature fusion for mechanical fault diagnosis. *ISA Trans*. **2022**, in press. doi: 10.1016/j.isatra.2022.06.035
61. Li, G.Q.; Fang, Q.; Zha, L.L.; *et al*. HAM: Hybrid attention module in deep convolutional neural networks for image classification. *Pattern Recognit.*, **2022**, *129*: 108785.
62. Sun, H.; Zeng, X.X.; Xu, T.; *et al*. Computer-aided diagnosis in histopathological images of the endometrium using a convolutional neural network and attention mechanisms. *IEEE J. Biomed. Health Inform.*, **2020**, *24*: 1664−1676.
63. Mishra, S.S.; Mandal, B.; Puhan, N. B. Perturbed composite attention model for macular optical coherence tomography image classification. *IEEE Trans. Artif. Intell.*, **2022**, *3*: 625−635.
64. Zhao, W.W.; Wang, R.Z.; Qi, Y.L.; *et al*. BASCNet: Bilateral adaptive spatial and channel attention network for breast density classification in the mammogram. *Biomed. Signal Process. Control*, **2021**, *70*: 103073.
65. Wei, Z.H.; Li, Q.; Song, H. Dual attention based network for skin lesion classification with auxiliary learning. *Biomed. Signal Process. Control*, **2022**, *74*: 103549.
66. Qin, Z.W.; Liu, Z.; Guo, Q.H.; *et al*. 3D convolutional neural networks with hybrid attention mechanism for early diagnosis of Alzheimer's disease. *Biomed. Signal Process. Control*, **2022**, *77*: 103828.
67. Xu, Z.W.; Ren, H.J.; Zhou, W.; *et al*. ISANET: Non-small cell lung cancer classification and detection based on CNN and attention mechanism. *Biomed. Signal Process. Control*, **2022**, *77*: 103773.
68. Chen, G.P.; Zhao, Y.; Dai, Y.; *et al*. Asymmetric U-shaped network with hybrid attention mechanism for kidney ultrasound images segmentation. *Expert Syst. Appl.*, **2023**, *212*: 118847.
69. Hu, H.X.; Li, Q.Q.; Zhao, Y.F.; *et al*. Parallel deep learning algorithms with hybrid attention mechanism for image segmentation of lung tumors. *IEEE Trans. Industr. Inform.*, **2021**, *17*: 2880−2889.
70. Yu, J.K.; Yang, D.D.; Zhao, H. S. FFANet: Feature fusion attention network to medical image segmentation. *Biomed. Signal Process. Control*, **2021**, *69*: 102912.
71. Xu, Z.H.; Liu, S.J.; Yuan, D.; *et al*. ω-net: Dual supervised medical image segmentation with multi-dimensional self-attention and diversely-connected multi-scale convolution. *Neurocomputing*, **2022**, *500*: 177−190.
72. Chen, J.D.; Chen, W.R.; Zeb, A.; *et al*. Segmentation of medical images using an attention embedded lightweight network. *Eng. Appl. Artif. Intell.*, **2022**, *116*: 105416.
73. Cao, L.Y.; Li, J.W.; Chen, S. Multi-target segmentation of pancreas and pancreatic tumor based on fusion of attention mechanism.

Biomed. Signal Process. Control, **2023**, *79*: 104170.
74. Chen, P.H.; Men, S.Y.; Lin, H.B.; et al. Detection of local lesions in tongue recognition based on improved faster R-CNN. In *Proceedings of 2021 6th International Conference on Computational Intelligence and Applications, Xiamen, China, 11–13 June 2021*; IEEE: Xiamen, China, 2021; pp. 165−168. doi: 10.1109/ICCIA52886.2021.00039
75. Li, X.Y.; Chai, Y.; Zhang, K.; et al. Early gastric cancer detection based on the combination of convolutional neural network and attention mechanism. In *Proceedings of 2021 China Automation Congress, Beijing, China, 22–24 October 2021*; IEEE: Beijing, China, 2021; pp. 5731−5735. doi: 10.1109/CAC53003.2021.9728413
76. Jiang, Z.F.; Liu, X.; Yan, Z.Z.; *et al*. Improved detection performance in blood cell count by an attention-guided deep learning method. *OSA Continuum*, **2021**, *4*: 323−333.
77. Zhang, J.Y.; Zhang, D.Q. Dual-attention network for acute pancreatitis lesion detection with CT images. In *Proceedings of 2021 International Conference on Medical Imaging and Computer-Aided Diagnosis* (*MICAD 2021*); Su, R.D.; Zhang, Y.D.; Liu, H., Eds.; Springer: Singapore, 2021; pp. 238–250. doi: 10.1007/978-981-16-3880-0_25
78. Pan, X.Y.; Liu, X.X.; Bai, W.D.; et al. Detection model of nasolaryngology lesions based on multi-scale contextual information fusion and cascading mixed attention. In *Proceedings of 2021 16th International Conference on Intelligent Systems and Knowledge Engineering, Chengdu, China, 26 –28 November 2021*; IEEE: Chengdu, China, 2021; pp. 464 –470. doi: 10.1109/ISKE54062.2021.9755353
79. Zhao, Y.Y.; Wang, J.X.; Wang, X.M.; et al. A new pulmonary nodule detection based on multiscale convolutional neural network with channel and attention mechanism. In *Signal and Information Processing, Networking and Computers*; Sun, J.D.; Wang, Y.; Huo, M.Y.; Xu, L.X., Eds.; Springer: Singapore, 2023; pp. 1004–1010. doi: 10.1007/978-981-19-3387-5_120
80. Zhou, Y.; Wang, B.Y.; He, X.D.; *et al*. DR-GAN: Conditional generative adversarial network for fine-grained lesion synthesis on diabetic retinopathy images. *IEEE J. Biomed. Health Inform.*, **2022**, *26*: 56−66.
81. Jiang, M.F.; Zhi, M.H.; Wei, L.Y.; *et al*. FA-GAN: Fused attentive generative adversarial networks for MRI image super-resolution. *Comput. Med. Imaging Graph.*, **2021**, *92*: 101969.
82. Li, M.; Wang, Y.W.; Zhang, F.C.; et al. Deformable medical image registration based on unsupervised generative adversarial network integrating dual attention mechanisms. In *Proceedings of 2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, Shanghai, China, 23–25 October 2021*; IEEE: Shanghai, China, 2021; pp. 1–6. doi: 10.1109/CISP-BMEI53629.2021.9624229
83. Zhu, D.M.; Sun, D.G.; Wang, D. B. Dual attention mechanism network for lung cancer images super-resolution. *Comput. Methods Programs Biomed.*, **2022**, *226*: 107101.
84. Chi, J.N.; Sun, Z.Y.; Wang, H.; *et al*. CT image super-resolution reconstruction based on global hybrid attention. *Comput. Biol. Med.*, **2022**, *150*: 106112.
85. Li, Y.C.; Zeng, X.H.; Dong, Q.; *et al*. RED-MAM: A residual encoder-decoder network based on multi-attention fusion for ultrasound image denoising. *Biomed. Signal Process. Control*, **2023**, *79*: 104062.
86. Vaswani, A.; Ramachandran, P.; Srinivas, A.; et al. Scaling local self-attention for parameter efficient visual backbones. In *Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021*; IEEE: Nashville, TN, USA, 2021; pp. 12889–12899.
87. Vila, L.C.; Escolano, C.; Fonollosa, J.A.R.; et al. End-to-end speech translation with the transformer. In *Proceedings of the Fourth International Conference, IberSPEECH 2018, Barcelona, Spain, 21–23 November 2018*; ISCA: Barcelona, Spain, 2018; pp. 60–63.
88. Chen, L.W.; Rudnicky, A. Fine-grained style control in transformer-based text-to-speech synthesis. In *Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, Singapore, 23–27 May 2022*; IEEE: Singapore, Singapore, 2022; pp. 7907–7911. doi: 10.1109/ICASSP43922.2022.9747747
89. Egonmwan, E.; Chali, Y. Transformer and seq2seq model for paraphrase generation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation @EMNLP-IJCNLP 2019, Hong Kong, China, 4 November 2019*; ACL: Hong Kong, China, 2019; pp. 249–255.
90. Carion, N.; Massa, F.; Synnaeve, G.; et al. End-to-end object detection with transformers. In *Proceedings of the 16th European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; Springer: Glasgow, UK, 2020; pp. 213–229. doi: 10.1007/978-3-030-58452-8_13
91. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations, Virtual Event, Austria, 3 –7 May 2021*; OpenReview.net, 2021. Available online: https://arxiv.org/abs/2010.11929 (accessed on 10 October 2022).
92. Touvron, H.; Cord, M.; Douze, M.; et al. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning, 18–24 July 2021*; PMLR, 2021; pp. 10347−10357. Available online: https://arxiv.org/abs/2012.12877 (accessed on 10 October 2022).
93. Wang, W.H.; Xie, E.Z.; Li, X.; et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of 2021 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021*; IEEE: Montreal, Canada, 2021; pp. 568–578. doi: 10.1109/ICCV48922.2021.00061
94. Liu, Z.; Lin, Y.T.; Cao, Y.; et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 10–17 October 2021*; IEEE: Montreal, Canada, 2021; pp. 10012–10022. doi: 10.1109/ICCV48922.2021.00986
95. Xia, Z.F.; Pan, X.R.; Song, S.J.; et al. Vision transformer with deformable attention. In *Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022*; IEEE: New Orleans, USA, 2022; pp. 4794–4803. doi: 10.1109/CVPR52688.2022.00475
96. He, S.; Grant, P.E.; Ou, Y. M. Global-Local transformer for brain age estimation. *IEEE Trans. Med. Imaging*, **2022**, *41*: 213−224.
97. Li, M.L.; Li, X.; Jiang, Y.C.; *et al*. Explainable multi-instance and multi-task learning for COVID-19 diagnosis and lesion segmentation in CT images. *Knowl. Based Syst.*, **2022**, *252*: 109278.
98. Wang, S.; Zhuang, Z.X.; Xuan, K.; et al. 3DMeT: 3D medical image transformer for knee cartilage defect assessment. In *Proceedings of the 12th International Workshop on Machine Learning in Medical Imaging, Strasbourg, France, 27 September 2021*; Springer: Strasbourg, France, 2021; pp. 347–355. doi: 10.1007/978-3-030-87589-3_36
99. Verenich, E.; Martin, T.; Velasquez, A.; et al. Pulmonary disease classification using globally correlated maximum likelihood: An auxiliary attention mechanism for convolutional neural networks. arXiv preprint arXiv: 2109.00573, 2021. Available online: https://arxiv.org/abs/2109.00573 (accessed on 10 October 2022).
100. Wu, W.J.; Mehta, S.; Nofallah, S., *et al*. Scale-aware transformers for diagnosing melanocytic lesions. *IEEE Access*, **2021**, *9*: 163526−163541.

101. Barhoumi, Y.; Ghulam, R. Scopeformer: N-CNN-ViT hybrid model for intracranial hemorrhage classification. arXiv preprint arXiv: 2107.04575, 2021. Available online: https://arxiv.org/abs/2107.04575 (accessed on 10 October 2022).
102. Sun, Q.X.; Fang, N.H.; Liu, Z.; *et al*. HybridCTrm: Bridging CNN and transformer for multimodal brain image segmentation. *J. Healthc. Eng.*, **2021**, *2021*: 7467261.
103. Hatamizadeh, A.; Nath, V.; Tang, Y.C.; et al. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *Proceedings of the 7th International MICCAI Brainlesion Workshop, Strasbourg, France, 27 September-1 October 2021*; Springer: Strasbourg, France, 2022; pp. 272–284 doi: 10.1007/978-3-031-08999-2_22
104. Gao, Y.H.; Zhou, M.; Metaxas, D.N. UTNet: A hybrid transformer architecture for medical image segmentation. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021*; Springer: Strasbourg, France, 2021; pp. 61–71. doi: 10.1007/978-3-030-87199-4_6
105. Ji, Y.F.; Zhang, R.M.; Wang, H.J.; et al. Multi-compound transformer for accurate biomedical image segmentation. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021*; Springer: Strasbourg, France, 2021; pp. 326–336. doi: 10.1007/978-3-030-87193-2_31
106. Fu, Z.Y.; Zhang, J.; Luo, R.Y.; *et al*. TF-Unet: An automatic cardiac MRI image segmentation method. *Math. Biosci. Eng.*, **2022**, *19*: 5207−5222.
107. Xie, Y.T.; Zhang, J.P.; Shen, C.H.; et al. CoTr: Efficiently bridging CNN and transformer for 3D medical image segmentation. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021*; Springer: Strasbourg, France, 2021; pp. 171–180. doi: 10.1007/978-3-030-87199-4_16
108. Ma, X.H.; Luo, G.N.; Wang, W.; et al. Transformer network for significant stenosis detection in CCTA of coronary arteries. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021*; Springer: Strasbourg, France, 2021; pp. 516–525. doi: 10.1007/978-3-030-87231-1_50
109. Jiang, H.; Zhang, P.L.; Che, C.; *et al*. RDFNet: A fast caries detection method incorporating transformer mechanism. *Comput. Math. Methods Med.*, **2021**, *2021*: 9773917.
110. Kong, Q.R.; Wu, Y.R.; Yuan, C.; et al. CT-CAD: Context-aware transformers for end-to-end chest abnormality detection on X-rays. In *Proceedings of 2021 IEEE International Conference on Bioinformatics and Biomedicine, Houston, TX, USA, 9–12 December 2021*; IEEE: Houston, USA, 2021; pp. 1385–1388. doi: 10.1109/BIBM52615.2021.9669743
111. Tao, R.; Zheng, G.Y. Spine-transformers: Vertebra detection and localization in arbitrary field-of-view spine CT with transformers. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021*; Springer: Strasbourg, France, 2021; pp. 93–103. doi: 10.1007/978-3-030-87199-4_9
112. Chen, L.Y.; You, Z.Y.; Zhang, N.; *et al*. UTRAD: Anomaly detection and localization with U-Transformer. *Neural Networks*, **2022**, *147*: 53−62.
113. Feng, C.M.; Yan, Y.L.; Fu, H.Z.; et al. Task transformer network for joint MRI reconstruction and super-resolution. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021*; Springer: Strasbourg, France, 2021; pp. 307–317. doi: 10.1007/978-3-030-87231-1_30
114. Wang, D.Y.; Wu, Z.; Yu, H.Y. TED-Net: Convolution-free T2T vision transformer-based encoder-decoder dilation network for low-dose CT denoising. In *Proceedings of the 12th International Workshop on Machine Learning in Medical Imaging, Strasbourg, France, 27 September 2021*; Springer: Strasbourg, France, 2021; pp. 416–425. doi: 10.1007/978-3-030-87589-3_43
115. Korkmaz, Y.; Dar, S.U.H.; Yurt, M.; *et al*. Unsupervised MRI reconstruction via zero-shot learned adversarial transformers. *IEEE Trans. Med. Imaging*, **2022**, *41*: 1747−1763.
116. Shin, H.C.; Ihsani, A.; Mandava, S.; et al. GANBERT: Generative adversarial networks with bidirectional encoder representations from transformers for MRI to PET synthesis. arXiv preprint arXiv: 2008.04393, 2020. Available online: https://arxiv.org/abs/2008.04393 (accessed on 10 October 2022).
117. Song, L.; Liu, G.X.; Ma, M. R. TD-Net: Unsupervised medical image registration network based on Transformer and CNN. *Appl. Intell.*, **2022**, *52*: 18201−18209.
118. Jiang, Y.C.; Yin, S.; Li, K.; *et al*. Industrial applications of digital twins. *Philos. Trans. Roy. Soc. A: Math. Phys. Eng. Sci.*, **2021**, *379*: 20200360.
119. Li, X.; Jiang, Y.C.; Liu, C.L.; *et al*. Playing against deep-neural-network-based object detectors: A novel bidirectional adversarial attack approach. *IEEE Trans. Artif. Intell.*, **2022**, *3*: 20−28.
120. Seo, D.; Oh, K.; Oh, I. S. Regional multi-scale approach for visually pleasing explanations of deep neural networks. *IEEE Access*, **2020**, *8*: 8572−8582.
121. Yin, S.; Rodriguez-Andina, J.J.; Jiang, Y. C. Real-time monitoring and control of industrial cyberphysical systems: With integrated plant-wide monitoring and control framework. *IEEE Ind. Electron. Mag.*, **2019**, *13*: 38−47.
122. Zhou, Q.; Wang, Q.W.; Bao, Y.C.; *et al*. LAEDNet: A Lightweight Attention Encoder–Decoder Network for ultrasound medical image segmentation. *Comput. Electr. Eng.*, **2022**, *99*: 107777.
123. Jiang, Y.C.; Li, X.; Luo, H.; *et al*. Quo vadis artificial intelligence? *Discov. Artif. Intell.*, **2022**, *2*: 4.
124. Saha, A.; Hosseinzadeh, M.; Huisman, H. End-to-end prostate cancer detection in bpMRI via 3D CNNs: Effects of attention mechanisms, clinical priori and decoupled false positive reduction. *Med. Image Anal.*, **2021**, *73*: 102155.
125. Sun, K.; He, M.J.; He, Z.C.; *et al*. EfficientNet embedded with spatial attention for recognition of multi-label fundus disease from color fundus photographs. *Biomed. Signal Process. Control*, **2022**, *77*: 103768.
126. Shen, N.Y.; Wang, Z.Y.; Li, J.; *et al*. Multi-organ segmentation network for abdominal CT images based on spatial attention and deformable convolution. *Expert Syst. Appl.*, **2023**, *211*: 118625.

Citation: Li, X.; Li, M.; Yan, P.; *et al*. Deep learning attention mechanism in medical image analysis: basics and beyonds. *International Journal of Network Dynamics and Intelligence*. 2023, 2(1): 93−116. doi: 10.53941/ijndi0201006