Journal of Radiation Research and Applied Sciences 16 (2023) 100680

Contents lists available at ScienceDirect

Journal of Radiation Research and Applied Sciences

journal homepage: www.journals.elsevier.com/journal-of-radiation-research-and-applied-sciences

Transformers and their application to medical image processing: A review

Dongmei Zhu, Dongbo Wang *

College of Information Management, Nanjing Agricultural University, Nanjing, 210095, China

ARTICLE INFO

Keywords:
Transformer
Image processing
Image classification
Image segmentation
Image reconstruction

ABSTRACT

Transformers perform well in natural language processing tasks and have made many breakthroughs in computer vision. In medical image processing, transformers are successfully used in image segmentation, classification, reconstruction, and diagnosis. In this paper, we mainly expound on the transformer principle and its application in medical imaging. Specifically, we first introduce the basic principles and model structure of transformers. Then, we summarize the improvement mechanism of the transformer's network including combining the transformer with the Unet network, creating a transformer lightweight variant network, strengthening the cross-fast link mechanism, and building a large model with the transformer as the skeleton. Second, extensive discussion is given to medical image segmentation, reconstruction, classification, and other applications. Finally, the main challenges transformers face in the medical image processing field and future development prospects. Furthermore, we systematically summarize the latest research progress of transformers and their application in medical image processing, which has significant reference value for transformer research in the medical field.

1. Introduction

Transformer (Vaswani et al., 2017) is a well-known deep learning model that is widely adopted in many fields, such as natural language processing (NLP), computer vision (CV), and speech processing. The transformer was initially proposed as a sequence-to-sequence model for machine translation, and research shows that the transformer-based pre-training model (PTM) can achieve the most advanced performance on various tasks. Therefore, transformer has become the architecture of choice for NLP, especially PTMs (Xie & C, 2022). In addition to language-related applications, transformer has been adopted in CV, audio processing, and other disciplines, such as chemistry and life sciences. An extended model of the Transformer model, as shown in Fig. 1, the model above is an improved model of the transformer, and the following is its variant model in the medical field.

In 2017, the Google Machine Translation Team (Vaswani et al., 2017) published "Attention is All You Need", which proposed the transformer model, abandoning network structures such as recurrent neural networks (RNN) and convolutional neural networks (CNN), and only using the attention mechanism (Hao et al., 2023) to carry out machine translation tasks and achieving good results. Due to its favorable experimental results, the attention mechanism has gained significant attention.

In 2018, Radford et al. (Radford et al., 2018) presented the Generative Pre-trained Transformer (GPT) model, a semi-supervised method that utilizes unsupervised pre-training with large amounts of unlabeled text data. Through fine-tuning with supervised learning, GPT achieved optimal performance across various NLP tasks, setting new records in 12 out of 21 benchmarks. In the same year, Devlin et al. (Devlin et al., 2018) proposed BERT, a bidirectional transformer encoder model that surpassed previous representation models in sentence-level and block-level tasks, establishing 11 new records for NLP performance.

In February 2019, Radford et al. (Radford et al., 2019) proposed GPT-2, an upgraded version of GPT, whose significant difference is more scale and more training data. GPT is a 12-layer transformer, BERT is the deepest 24-layer transformer, and GPT-2 is a 48-layer, the experimental results show that the current model is still underfitting. In October 2019, Sanh et al. (Sanh et al., 2019) proposed a compression model of BERT characterized by smaller, faster, cheaper, and lighter. In the pre-training phase, the number of parameters in the model was reduced by 40% and the reasoning speed of the model was increased by 60% while maintaining 97% of the model language comprehension.

In 2020, Carion et al. (Carion et al., 2020) proposed a DERT model and introduced a transformer for target detection tasks. Transform the target detection task into a sequence prediction task, use the transformer encoder-decoder structure and the bilateral matching method, and directly obtain the prediction result sequence from the input image. In August 2020, Croce et al. (Croce et al., 2020) proposed a

* Corresponding author.
E-mail address: db.wang@njau.edu.cn (D. Wang).

https://doi.org/10.1016/j.jrras.2023.100680
Received 6 July 2023; Received in revised form 12 September 2023; Accepted 13 September 2023
Available online 3 October 2023

semi-supervised learning method based on GAN-BERT to classify images and extended the fine-tuning stage by introducing discriminator generator settings. This method can generate positron emission computed tomography (PET) images from magnetic resonance imaging (MRI) images with a wide intensity range without manual adjustments in pre-processing or post-processing, which is a scalable and deployable approach. In October 2020, Dosovitskiy et al. (Dosovitskiy et al., 2020) proposed the ViT model, where the image is partitioned into patches of uniform size, corresponding to tokens in NLP, which are then represented in embedding and become the input to the transformer. Experiments have verified that ViT can still achieve good results if pre-trained on a large amount of sufficient data. However, if the training data is insufficient, the effect of ViT will be significantly reduced.

In August 2021, Wang et al. (Wang et al., 2021a) proposed a PVT Transformer model suitable for pixel-level image tasks and achieved considerable performance in visual tasks such as object detection and semantic segmentation. In August 2021, Liu et al. (Liu et al., 2021a) introduced the Swin Transformer model, which achieved a hierarchical structure in CNN through the shifted windows operation. Experimental results demonstrated that the Swin Transformer model outperformed previous architectures in multiple tasks such as classification, detection, and segmentation. As the transformer is used in an increasingly wide range of applications, its successful applications are in medical image processing, such as medical image segmentation (Yanping et al., 2023), detection, reconstruction, and diagnosis. In October 2019, Xiong et al. (Xiong et al., 2019) proposed the Reinforced Transformer for Medical Image Captioning (RTMIC) report generation model that improves the performance of the Reinforced Transformer model in BLEU-1 by more than 50% compared to other state-of-the-art image captioning methods.

In 2021, Valanarasu et al. (Valanarasu et al., 2021) proposed a gated axial attention model, which extends existing architectures by incorporating an additional control mechanism in the self-attention module. It is experimentally demonstrated that it is possible to train medical images better and with good results on convolution architecture and other related transformers. The same year, Xie et al. (Xie et al., 2021) proposed a new framework for accurate 3D medical image segmentation that effectively connects convolutional neural networks and transformers (CoTr). Experimental results show that our CoTr brings considerable performance improvement in 3D multi-organ segmentation tasks compared with other CNN and transformer hybrid approaches. Due to the limitation that convolutional operations do not learn global and remote semantic information interactions well, Cao et al. (Cao et al., 2022a) proposed Swin-Unet, a pure transformer similar to Unet for medical image segmentation. The tokenized image blocks are fed into the transformer's U-shaped en-decoder architecture via a jump connection for local global semantic feature learning.

Dalma et al. (Dalmaz et al., 2022) designed ResViT, a novel generative adversarial method for medical image synthesis, which exploits the context sensitivity of visual transformers with the precision of convolution operators and the realism of adversarial learning. To better balance the accuracy and efficiency of the transformer module for quick and precise medical picture segmentation, Xu et al. (Xu et al., 2021a) introduced LeViT-UNet, which incorporates the LeViT transformer module into the U-Net architecture and employs LeViT as the encoder of LeViT-UNet. The results demonstrate that LeViT-UNet obtains an equivalent Dice Similarity Coefficient (DSC) to Swin-UNet and TransUNet. For example, the LeViT-UNet-192 and LeViT-Unet-384 reach 90.08% and 90.32% DSC, respectively.

Considering the small dataset of medical images, it is impossible to apply pure Transformer in the deep learning algorithm of clinical medicine. Therefore, Dai et al. (Dai et al., 2021a) proposed to use of TransMed for multimodal medical image classification, which combines the advantages of CNN and transformer and can effectively extract low-level features of images and establish long-term dependency between modes. TransMed achieves state-of-the-art performance costs while requiring fewer parameters and computational effort. Luthra et al. (Luthra et al., 2021) proposed Eformer, an edge-enhanced transformer based on a novel architecture that uses the transformer module to build an encoding-decoding network for medical image denoising. It combines the learnable Sobel-Feldman operator to improve image edges and proposes an efficient method to connect them in the architecture's intermediate layers. On the AAPM Mayo Clinic Low Dose CT Grand Challenge dataset, the Eformer model achieved state-of-the-art performance of 43.487 dB peak signal-to-noise ratio (PSNR) and 0.9861 structural similarities (SSIM). Current 2D-based medical image segmentation only considers attention encoding within a single slice and does not take advantage of the axis information that 3D volumes naturally provide. Yan et al. (Yan et al., 2022) proposed the axial fusion transformer UNet (AFTer-UNet), which exploits both the ability of convolutional layers to extract detailed features and the advantages of the transformer in modeling long sequences. Experiments on three multi-organ segmentation datasets show that this method is superior to the current most advanced method.

To overcome the challenge of obtaining a sufficient amount of high-quality data with high-cost annotation. Language meets Vision Transformer (LViT) is a new visual, linguistic medical image segmentation model proposed by Li et al. (Li et al., 2023a) in 2023. In the experiments, the Language-Vision (LV)loss is designed to supervise the training of unlabeled images using textual information. The results show that our proposed LViT has better segmentation performance in both total and semi-complementary cases. Aiming at the problem that the receptive field of convolution in the previous medical image segmentation network is too small and the feature loss of the transformer, Yang et al. (Yang et al., 2023) proposed an end-to-end lightweight contextual transformer medical image segmentation network (CoT-TransUNet). For the input image, the encoder uses a hybrid module of CoTNet-Transformer, and CoTNet is used as a feature extractor to generate feature maps. In experiments on multi-organ segmentation tasks, CoT-TransUNet achieves better performance than other networks.

In this paper, we comprehensively summarize the research status of transformer networks. Specifically, we explain the basic principles of a transformer, from the network framework and model structure, then summarize the improvement mechanism of the transformer's network including combining the transformer with the Unet network, creating a transformer lightweight variant network, strengthening the cross-fast link mechanism, and building a large model with the transformer as the skeleton. Finally, we discuss the challenges faced by Transformer in the medical imaging field and look at the future development direction. Our main contributions are as follows.

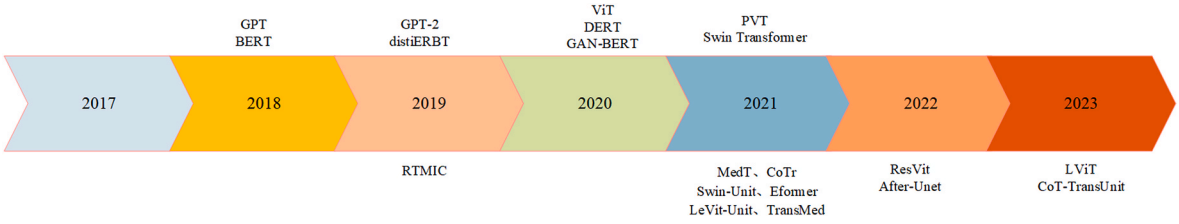

Fig. 1. Development of the Transformer variant model.

1) Summarize the progress of transformers in the medical field, which is conducive to quickly understanding and familiarizing with the application and development status of the transformer network for medical image processing in medical images.
2) Classify the improvement of transformer methods in the medical field, and introduce different types of transformer improvement methods and their characteristics.
3) Summarize the current challenges in the field of medical image processing space, and prospecte the future development trends in this field.

2. Transformers

Vaswanid et al. (Vaswani et al., 2017) first proposed the transformer due to its unique design, which allows it to handle indefinitely long inputs, capture long-distance dependencies, and have sequence-to-sequence (seq2seq) properties. Transformer mainly includes a decoder and encoder. Each encoder includes position encoding, multi-head attention mechanism, layer normalization (LN), feed forward network (FFN), and skip connections. The decoder is the same as the encoder, except that a masked multi-head attention mechanism is added to the input layer (Fu et al., 2022), and the traditional transformer structure is shown in Fig. 2.

2.1. Encoder-decoder architecture

2.1.1. Encoder model

The Encoder is composed of six identical layers, as shown in Fig. 2. On the left side is "Nx", named N6, and each layer includes two parts: a multi-head attention mechanism and a feedforward neural network. A residual connection between the two sub-layers is used, followed by layer normalization. Each encoder is divided into two sub-layers: first

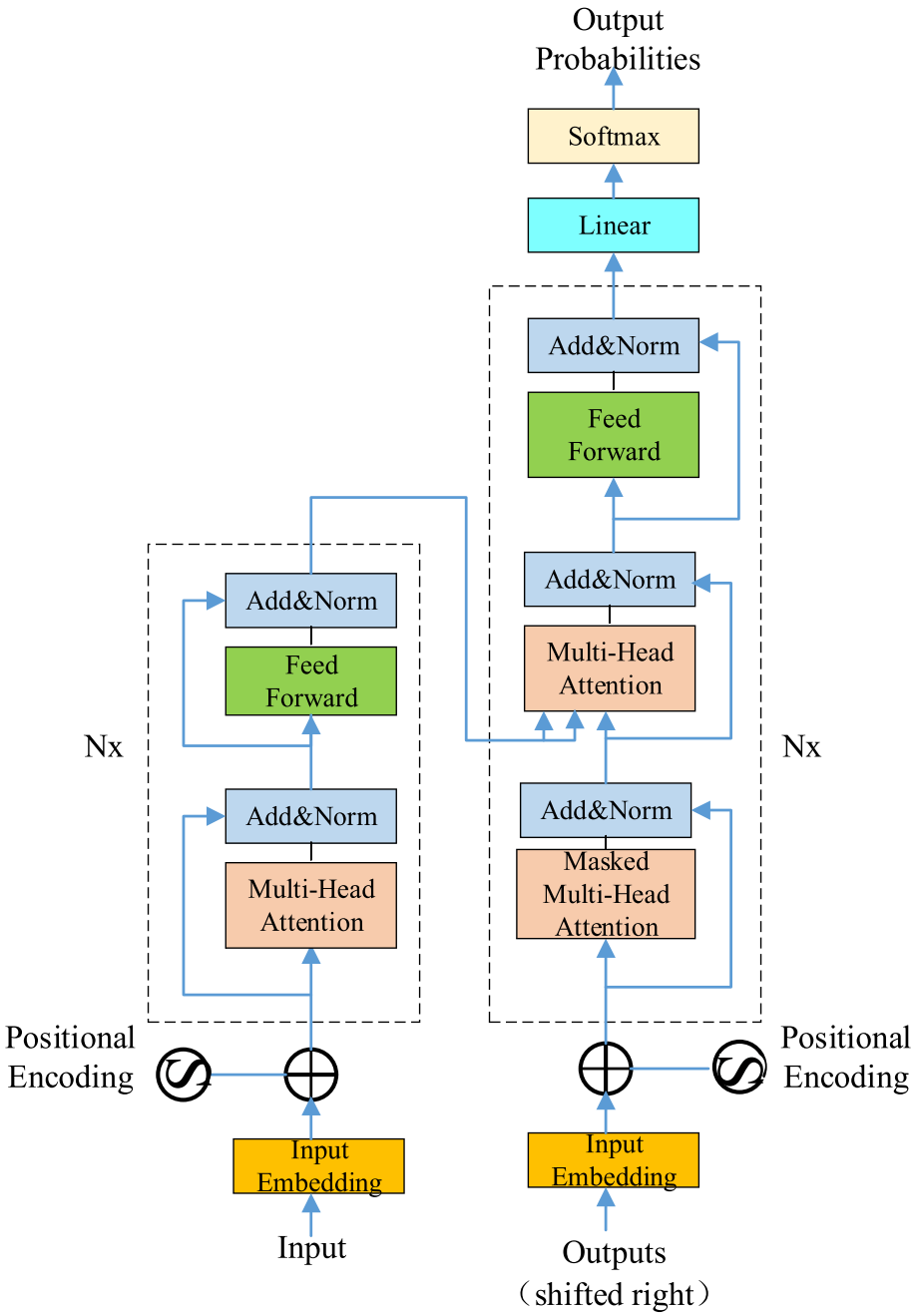

Fig. 2. Transformer model diagram.

flows through the self-attention layer, which helps the encoder look at other words when encoding a particular word. The output of the self-attention layer is then passed to a feedforward neural network layer, which is identical at each location, but their roles are independent and do not share weights.

2.1.2. Decoder model

The structure of the decoder and encoder is similar, and there is a sublayer that performs multi-head attention on the output of the encoder stack, called the encoding-decoding self-attention mechanism. First, it uses the embedding algorithm to convert the input word into a vector. The bottom encoder's input is the embedding vector. Inside each Encoder, the input vector goes through self-attention, and then through the feed-forward layer, the output vector of each encoder is the input to the encoder just above it. The size of the vectors is a hyperparameter, usually set to the length of the longest sentence in the training set.

The input representation x of the word in the transformer is obtained by adding the word embedding and the position embedding. There are many ways to get word embeddings, such as Word2Vec, Glove, and other algorithms pre-trained (Hong, 2022, pp. 125–128) or trained in the transformer. In transformer, in addition to the word embedding, we also need to use the position Embedding to indicate the word's position in the sentence. This is due to the fact that Transformer does not choose the basic structure of RNN, but only the global correlation information, and cannot choose the information of word order. The word position is important for NLP, so the position Embedding is used in Transformer to save the relative or absolute position of the word in the sequence. Furthermore, absolute locations are adopted in Transformer, and the formula is as follows:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{1}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{2}$$

Among them, pos represent the position of the word in the sentence, d represents the dimension of PE (same as the word embedding), 2i represents the actual dimension, and 2i+1 represents the odd dimension (i.e., $2i \leq d$, $2i+1 \leq d$).

2.2. Self-attention architecture

The multi-head attention mechanism adopted by the transformer structure enables the model to learn different semantic features in other sub-layer spaces and adjust the weights and different projection methods during the projection process to make the model more generalizable. The structure of self-attention is shown in Fig. 3, which requires the use of matrices query (Q), key (K), and value (V) in the calculation. First, it calculates the similarity between Q and K to obtain the weight, and divides the scaling layer by the parameter D_k to adjust the scaling, control the inner product not to be too large, and then uses the softmax function to normalize the similarity weight. Finally, the attention output is calculated by summing the normalized and corresponding weights, and the equation for calculating the output vector of the self-attention mechanism is as follows:

$$\text{attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

The multi-head attention mechanism module is shown in Fig. 3, essentially an integration of h self-attention mechanisms, and its module structure is not very complicated. All self-attention mechanisms pay attention to the same Q, K, and V, but each module only corresponds to a subspace in the final output sequence. The output sequences are independent of each other, which means the multi-head attention mechanism module can be used for different positions; simultaneous attention is achieved by additional information in the representation subspace. In the case of self-attention mechanisms, normalization suppresses this information.

When calculating, initialize h groups of Q, K, and V vectors, and the weight parameter W of each group of Q, K, and V is different, as shown in Equation (4). The multi-head attention mechanism module can learn more information in the representation subspace by introducing

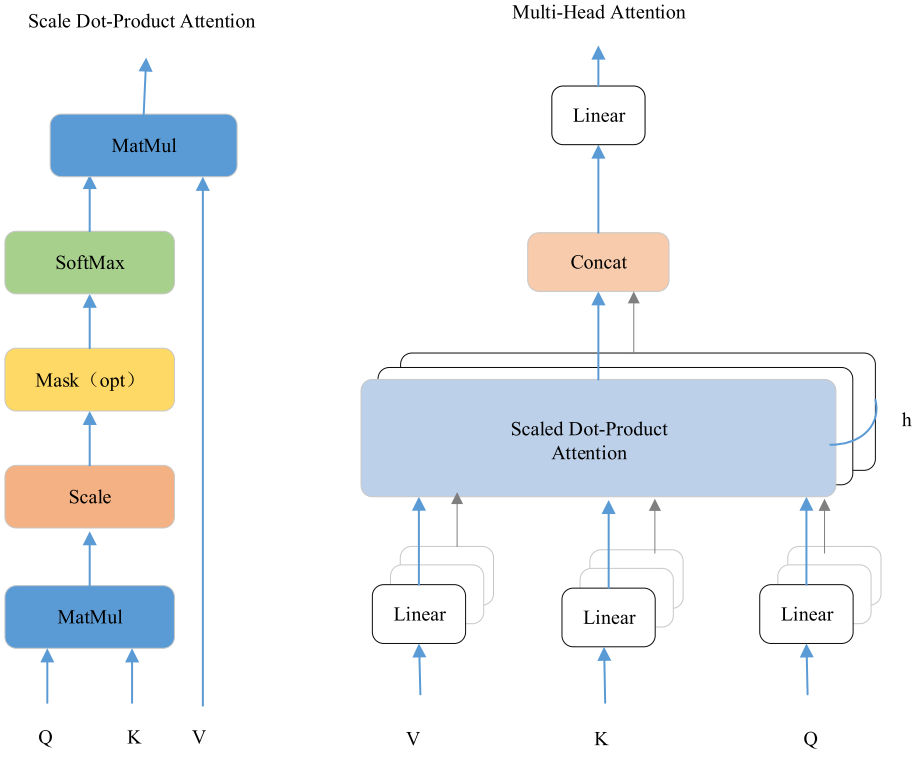

Fig. 3. Dot product attention mechanism in transformer.

different weights. Then, the self-attention mechanism is calculated for each group, and the obtained self-attention mechanism output results are connected and then multiplied by a weight vector W^0 to obtain the final output vector of the multi-head attention mechanism module, as shown in Equation (5). The calculation formula of the multi-head attention mechanism module is as follows:

$$\text{head}_i = \text{attention}\left(QW_i^Q, KW_i^K, VW_i^v\right) \tag{4}$$

$$\text{multi head}(Q, K, V) = \text{concat}(\text{head}_1, \cdots \text{head}_h)W^0 \tag{5}$$

The transformer model completely avoids the recurrent structure (Fan et al., 2022, pp. 25–36) and adopts a multi-head attention mechanism to achieve global dependency estimation of input and output. Since each attention head can learn to perform a different task, the multi-head attention mechanism can lead to more interpretive models.

3. Transformer network improvement mechanism

Transformer network has made remarkable achievements in the field of natural language processing with its self-attention mechanism and global modeling ability. However, with the in-depth research on transformers, researchers began to apply it to other fields and proposed a series of improvement mechanisms. These improvements include combining the transformer with the Unet network (Qiu et al., 2021a), creating a transformer lightweight variant network, strengthening the cross-fast link mechanism, and building a large model with the transformer as the skeleton, as shown in Fig. 4. These improvements are aimed at further improving the performance of the transformers in computer vision, medical image analysis and other fields, and expanding its application range. By combining the characteristics of the transformer with other network structures and optimization methods, these improvements provide new possibilities for solving practical problems, and also accelerate the development of fields such as medicine and computer vision. For researchers and practitioners in the medical and health field, understanding the improvement mechanism of these transformer networks will help to better apply and promote this advanced technology. Provide better solutions for tasks such as medical diagnosis and image analysis.

3.1. Combination of transformer and unet network

Li et al. (Li et al., 2021a) proposed a dual encoder-decoder-based X-Net (X-shaped). In the encoding stage, local and global features are extracted simultaneously by convolutional downsampling and transformer, merged through skip connections. In the decoding stage, the input image is reconstructed with a variational autoencoder branch to mitigate the effect of insufficient data. Compared with the U-shaped symmetric structure of TransUNet with only encoder and decoder branches, Chang et al. (Chang et al., 2021) designed the Claw U-NET with Transformers (TransClaw) network, a network structure with three branches of the encoder, upsampling, and decoder. Skip connections connect the multi-scale feature maps of each part. The convolution and transformation operations are combined in the encoding function, and the convolution part is used to extract shallow spatial features to restore image resolution after upsampling. The decoding part preserves the bottom upsampling structure for better-detailed segmentation performance. The transformer part is used to encode the patches, while the self-attention mechanism is used to obtain global information between sequences. Although the implementation results show that the model has not improved significantly on the Dice indicator, it performs well on the Hausdorff Distance (HD) indicator.

Influenced by GoogLeNet (Szegedy et al., 2015) and Swin Transformer, Liang et al. (Liang et al., 2022a) proposed a Transforming Conversion parallel network (TransConvert). It replaces the multi-branch structure in GoogLeNet with the transformer module and convolution module and the filter splicing layer of GoogeNet with Cross-Attention Fusion with Global and Local Feature (CAFGL) module based on the cross-attention mechanism to obtain Transformer Conversion Insertion (TC-Inception). Meanwhile, an improved skip connection Structure Called the Cross-Attention Fusion skip connection (SCCAF) mechanism can alleviate the semantic difference between encoder and decoder features, leading to better feature fusion. The CNN and Swin Transformer models exchange the detailed features and global background information of the 3D brain image through the cross-attention module, which not only improves the accuracy of tumour segmentation but also reduces the computational load of the model and improves the efficiency of model training.

Most of the networks mentioned above focus on improving model accuracy, ignoring model speed to some extent. To balance the speed and accuracy of the segmentation model, Xu et al. (Xu et al., 2021b) used LeViT (Graham et al., 2021) as the encoder of Vision Transformer based U-Net (LeViT UNet), which better weighs the accuracy and efficiency of the transformer block. LeViT can speed up model reasoning and extract global context information from feature maps. Keep the decoder part the same as UNet to help the model obtain global features from feature maps with spatial priors after convolution operations. LeViT-UNet's segmentation accuracy on the Synapse dataset exceeds that of most models, and it is particularly noteworthy that LeViT-UNet had the best segmentation performance among the fast segmentation networks of the time.

Li et al. (Li et al., 2021b) proposed a Segtran model based on a compression expansion transformer for the decoder. Compression-expansion transformer includes compressed attention block, expanded attention block, and learnable sinusoidal position code. Among them, the close attention module comes from the Induced Squeezed Attention Block (ISAB) in SetTransformer (Lee et al., 2019), which is specially used to deal with unordered set features. ISAB condenses the critical information of X (matrix of type $n \times d$) ($n \gg m$) through the transition feature map I (matrix of type $m \times d$), which can significantly reduce the complexity of the attention module. For the

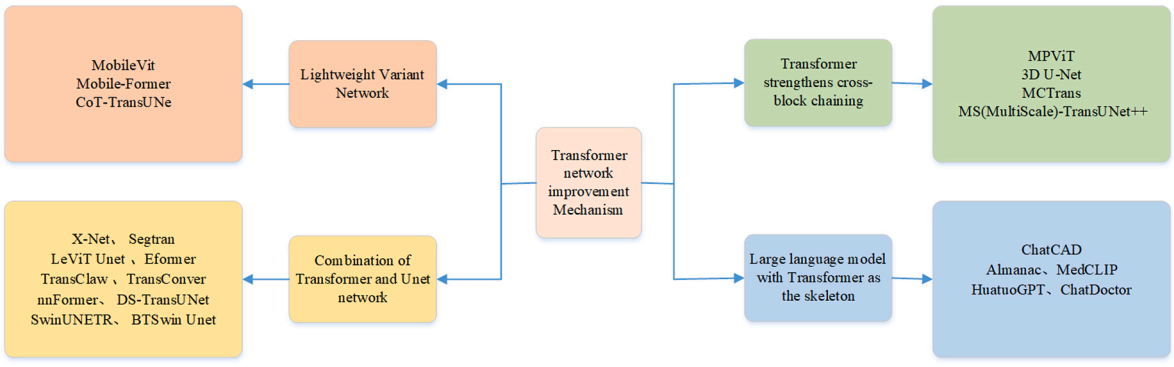

Fig. 4. Transformer network improvement graph.

extended attention module, replacing the multi-head attention mechanism with multiple single-head transformers is proposed to adapt to the diversity of data and obtain more differentiated sample features. In the position-coding part, to get pixel locality and semantic continuity, the author proposes a learnable sinusoidal position coding based on the sinusoidal position coding in the original transformer. Experimental results show that learnable position codes and features extracted by multiple transformers can improve model performance to a certain extent.

The work introduced above only improves the encoder or decoder. Next, this paper will discuss the transformer to simultaneously improve the network structure in the encoder and decoder. Tang et al. (Tang et al., 2022) proposed Swin UNet TRansformers (SwinUNETR) based on Swin Transformer to propose a self-supervised pre-training segmentation model with a layered encoder for self-supervised pre-training. The model is pre-trained with SwinTransformer modules on three proxy tasks: contrastive learning, masked voxel blocks, and random data augmentation on 5050 non-target CT images. These three proxy tasks can help the pre-training model learn Region of Interest (ROI) information, adjacent voxel information, and prior structural knowledge. In the target task, the fine-tuned SwinTransformer module combined with the convolutional layer performs excellently in the 3D medical image segmentation task.

Zhou et al. (Zhou et al., 2021) proposed that no other transFormer (nnFormer) alternately uses a transformer and CNN in the network and extracts feature information of each scale for multi-scale supervised learning to ensure that the multi-scale feature expression is as accurate as possible, and the nnFormer model framework is shown in Fig. 5. The encoder consists of an embedding module, a transformer block, and a downsampling layer, and the decoder comprises a transformer block, an upsampling layer, and an expansion block. Due to the introduction of multiple transformers, the computational load will be greatly increased. So the author pre-trained the transformer in Imagenet in advance, fixed the attention module and the multi-layer perceptron (Multi-Layer Perceptron, MLP) layer parameters, and the other parts performed new learning according to the target task. In addition, inspired by Swin-Transformer, the author replaces the original two-dimensional window with a three-dimensional window and performs self-attention calculations in the window. Compared with the three-dimensional multi-head attention mechanism, the calculation amount is reduced by more than 90%. The size of the 3D window is designed according to the size of the 3D image. The purpose is to avoid the mismatch between the 3D window and the 3D image, which will cause redundant information to be filled during calculation. The authors also propose that the learned embedding layers with successive, small convolutional layers have richer positional information and help reduce model complexity.

Benefiting from the patch embedding operation of ViT, Luthra et al. (Luthra et al., 2021) proposed the transformer based on edge enhancement, which uses transformer blocks to build a new architecture of a coder-decoder network. The convolutional feature maps pass through a local enhancement window LeWin Transformer block at each encoding and decoding stage. This block includes a non-overlapping Window Multi-head Self-Attention module (W-MSA) and a Local Enhancement Feed-Forward network (LeFF) to capture local context information effectively. The model was evaluated on the AAPM Mayo Clinic Low-dose CT Grand Challenge dataset and achieved state-of-the-art performance. LIN et al. (Lin et al., 2022) proposed the Dual Swin Transformer U-Net (DS-TransUNet) model. DS-TransUNet uses a SwinTransformer-based dual-scale encoder sub-network to extract coarse-grained and fine-grained feature representations of different language scales. Dual SwinTransformer branches constitute the encoder such that each department has different image slice sizes to obtain more varied multi-scale features. The Transformer Interactive Fusion module (TIF) is used to successfully establish the global dependencies between the characteristics of each scale through the self-attention mechanism to obtain rich multi-scale features.

Another U-shaped network BTSwin Unet model based on Swin-Transformer and Swin-Unet (Cao et al., 2022b) was proposed by Liang et al. (Liang et al., 2022b). It is a 3D U-shaped symmetrical brain tumour segmentation network composed of a coder-decoder and skip connection. This network also builds a self-supervised learning framework to pre-train the model encoder with reconstruction tasks. The non-overlapping patch embedding and downsampling modules in Swin Transformer are replaced by overlapping patch embedding and downsampling to enhance the locality of the segmentation network. Experimental results and ablation studies on the BraTS 2018 and BraTS 2019 datasets show that the method achieves comparable Dice scores and Hausdorff distances. The above model demonstrates the potential of SwinTransformer applied to medical image datasets. The above model reflects the potential of SwinTransformer in medical image data. Because SwinTransformer is lighter than a transformer when pre-training a large amount of data, it is more suitable for medical image segmentation tasks. Therefore, using SwinTransformer will help solve the problem that medical image datasets limit model progress.

3.2. *Lightweight variant network*

Most smart terminals' hardware is unsuitable for running large-scale deep neural network models. To combine the advantages of CNN's lightweight and Vit's global representation, Mehta (Mehta & Rastegari, 2021) proposed a light, general-purpose visual network for mobile devices: MobileVit. MobileVit network design aims to design lightweight, general-purpose, and low-latency networks for mobile vision tasks. The MobileVit network combines CNN and transformer, using CNN to extract local features and the transformer to extract global features. Unlike ViT and its variants, MobileViT learns international representations from different perspectives, using transformers to replace regional modeling in convolutions with global modeling. MobileViT block has the properties of CNN and ViT, which helps it learn better representations with fewer parameters and easy training. The downside is that MobileViT and other Transformer models are still slower than MobileNet, mainly due to the dedicated CUDA core on the GPU for convolution and other device-level optimizations on the CNN network.

Chen et al. (Chen et al., 2022) proposed Mobile-Former, a two-way bridge between MobileNet and the transformer designed in parallel, and the structure is shown in Fig. 6. Unlike recent work on vision transformers, the Mobile-Former transformer contains few randomly initialized markers to learn global prior, resulting in lower computational cost. Combined with the proposed lightweight cross-attention bridge model, Mobile-Former is computationally efficient and has more vital representation capabilities. The disadvantage is that the parallel design is not efficient in parameter sharing. While the Former model is computationally efficient due to the small number of tokens, it does not save on the number of parameters.

Aiming at the problem that the receptive field of convolution is too small and the features of the transformer are lost in the previous medical image segmentation network, an end-to-end lightweight context transformer medical image segmentation network CoT-TransUNet (Yang et al., 2023) was proposed. The network consists of three parts: encoder, decoder, and skip connections. In the encoder part of CoT-TransUNet, CoTNet is used as the feature extractor to extract richer features, and the number of transformer layers is increased to 16. The transformer block encodes the feature map into an input sequence and sends it to the decoder for upsampling. In the decoder part, the CARAFE operator with a larger receptive field is used in the upsampler, which can upsample based on content while maintaining light weight. Finally, the feature aggregation of the encoder and decoder at different resolutions is achieved through skip connections. The network's performance has been dramatically improved through the above improvements combined with the advantages of TransUNet. In the segmentation task of 8 abdominal organs: aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, spleen and stomach, DSC and HD were used as evaluation

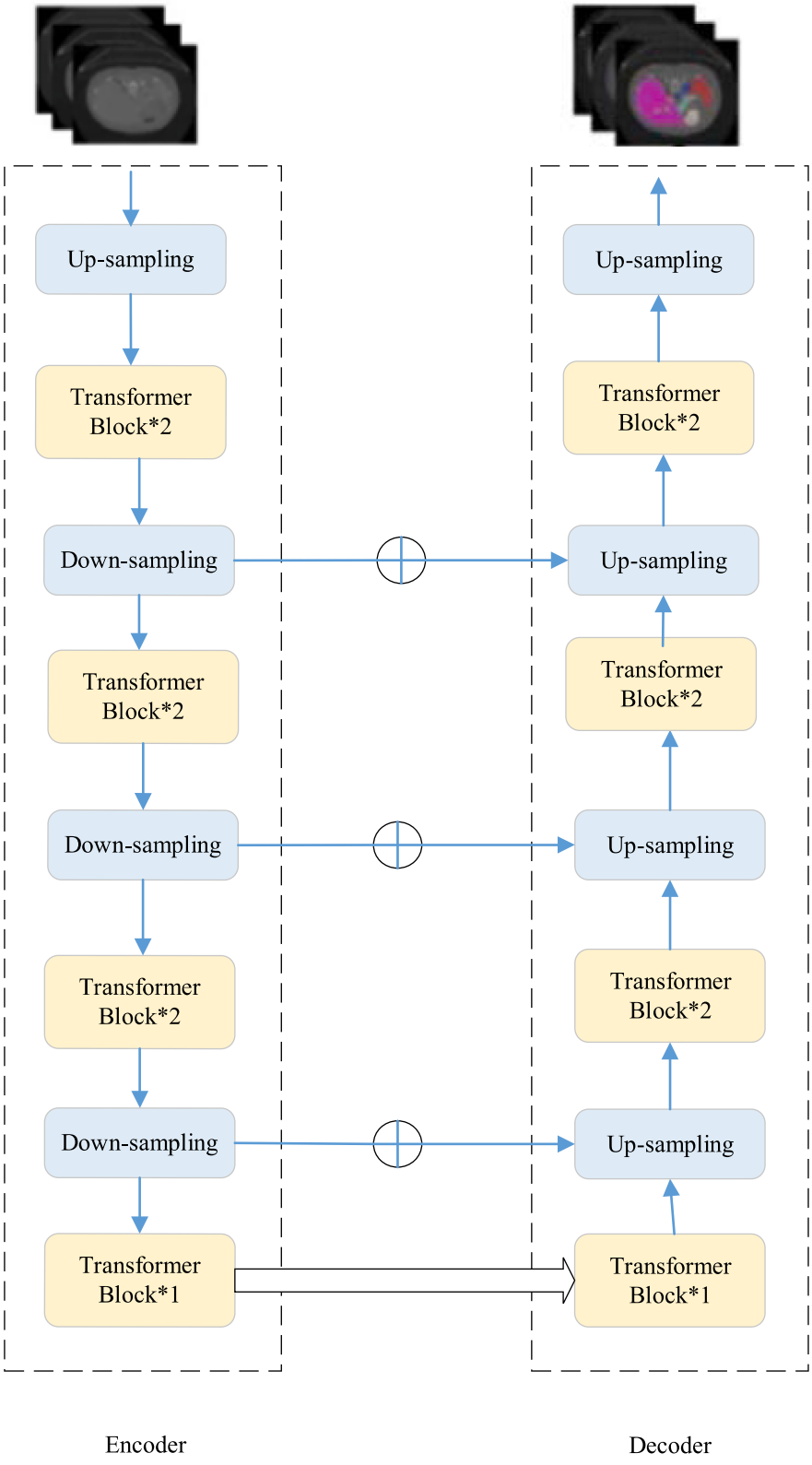

Fig. 5. Nnformer network structure diagram.

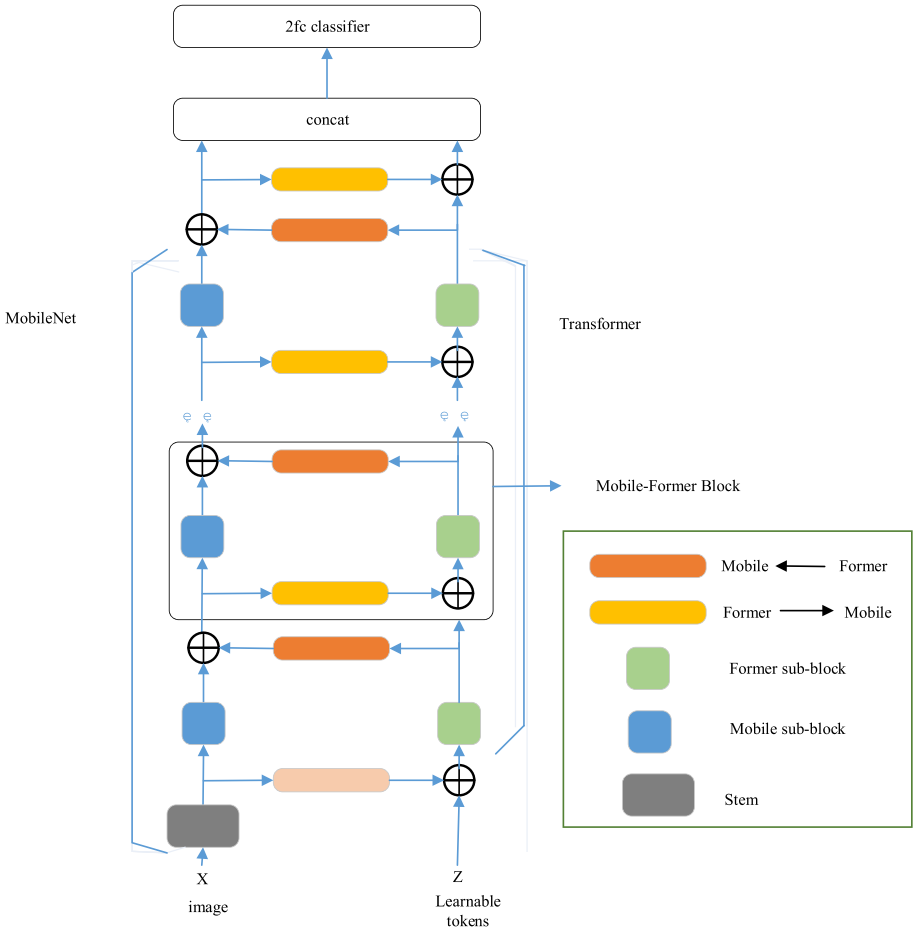

Fig. 6. Structure of mobile former.

indexes, and the 8 abdominal organs were labeled with different colors in the visualization results. Compared with TransUNet, CoT-TransUNet has improved DSC by about 0.9% and HD by about 5.7%. Its shortcoming is that all experiments are only segmented on 2D images. However, most medical image data are 3D.

3.3. *Transformer strengthens cross-block chaining*

Lee et al. (Lee et al., 2022) proposed the MPViT model for a multi-path visual transformer for dense prediction. By exploring multi-scale block embedding and multi-path structure, combined with overlapping convolutional block embedding, MPViT can simultaneously embed and aggregate features of different scales and the same sequence length. Tokens of different scales are sent to other transformer modules to construct coarse-grained and fine-grained features at the same feature level. Unlike the local-to-global feature extraction of previous models, the author introduces global-to-local feature interaction (GLI) to simultaneously utilize the convolution's local connectivity and the transformer's global context to represent features. Compared with existing models, CrossViT (Chen et al.) exploits different patch sizes and dual paths in a single-level structure, such as ViT (Dosovitskiy et al., 2020) and XCiT (Ali et al., 2021).

However, the interactions between the branches of CrossViT only happen through tokens, while MPViT allows patches of all different scales to interact. Furthermore, unlike CrossViT in classification, MPViT explores multi-dimensional paths and performs densely prediction of multi-stage structures. Experimental results show that MPVIT outperforms the state-of-the-art Vit Transformer while having fewer parameters and fewer operations. To solve the problem of computational and space complexity in the attention module of 3D medical images, Wang et al. (Wang et al., 2022a) proposed Multi-Scale (MS)-TransUNet++. The deep-wise convolutional layer reduces the feature space of K and V, thereby reducing the cost of calculating the attention score and superimposing multiple efficient transformer modules at the bottom of the model. Furthermore, in the multi-scale feature fusion module, adding network and dense connections can strengthen the feature connection between the encoder and decoder, allowing for better restoration of the detailed representation lost during downsampling. Most medical image segmentation models use DICE plus multi-classification cross-loss or binary cross-loss as the network loss function. In contrast, MS-TransUNet++ uses Focal (Lin et al., 2017), multi-scale structural similarity (MS-SSIM) (Wang et al., 2003), and Jaccard (Yu et al., 2016) Constitutes a global loss to supervise the segmentation results. Experiments prove that MS-TransUNet++ can perform competently in prostate MR and liver CT image segmentation.

Liu et al. (Liu et al., 2021b) designed a gated attention mechanism (Zhang et al., 2020) and transformer layers based on the structure of 3D U-Net to optimize the extracted image features. In addition, spatial pyramid pooling blocks (Zhai et al., 2022) and deep supervision (Qin et al., 2021) are integrated to improve performance. It filters out redundant information output by each layer in the encoder. To further optimize the segmentation network, the model does not change the composition of the loss function but supervises the intermediate features and the final result simultaneously to ensure feature consistency between the various parts of the model. The model solves the 3D segmentation challenge at different sizes ABVS data has low data quality. Experimental results show that this model can perform the best tumour segmentation.

Ji et al. (Ji et al., 2021a) also used a deformable transformer to improve model efficiency. Multi-Compound Transformer (MCTrans) embedded Transformer Self Attention (TSA) and Transformer Cross Attention (TCA) in the transition module. They facilitated TSA to acquire CNN output feature maps with a deformable attention mechanism. For the entire TCA module, the added learnable auxiliary embedding matrix is denoted by Q, and the feature maps from TSA are denoted by K and V. At the end of the TCA module, the temporary multi-classification results are obtained through linear mapping. The auxiliary loss is calculated with labels, which guides TSA to learn the difference between feature representations between different classes and the connection between feature representations of the same category, ensuring intra-class consistency and inter-class degree of distinction. The MCTrans model is evaluated on six segmentation datasets of three types: cell segmentation (Gamper et al., 2019), PolyP segmentation (Bernal et al., 2012, 2015; Jha et al., 2021; Silva et al., 2014), and skin lesion segmentation (Codella et al., 2019) demonstrating the generality of the proposed MCTrans on various segmentation tasks sex. Compared with the UNet baseline, MCTrans, with almost the same parameters and slightly increased computation, achieves a significant improvement of 3.64%. Other top methods, such as UNet++, outperform MCTrans in mathematics but with lower performance.

3.4. Large language model with transformer as the skeleton

Hiesinger et al. (Hiesinger et al., 2023) developed the Almanac large-scale language model framework, a framework for safely deploying large-scale language models in healthcare settings, aimed at more accurately answering clinical queries across specialties. Its language model fine-tuning the generative pre trained transformer architecture. The model is responsible for extracting relevant information from the scoring context returned by the retriever, and formulating an answer by combining context and Chain of Thought (CoT) reasoning cues. In the field of biomedicine, LLM models (such as LLaMa, and ChatGLM) perform poorly due to the lack of certain medical professional knowledge corpus. Zhang et al. (Zhang et al., 2023) constructed a Chinese medical instruction dataset through a medical knowledge graph and GPT3.5 API. And fine-tuning the LLaMa model to obtain an intelligent consultation model HuaTuo for the medical field. Compared with the original LLaMa which has not been fine-tuned by medical data instructions, the HuaTuo model performs well at the level of intelligent consultation. The core of HuatuoGPT is to use the refined data of ChatGPT and the real data of doctors in the supervised fine-tuning stage. Experimental results show that Huatuo Medical has reached the state-of-the-art in medical consultation among open-source large language models on GPT-4 evaluation, human evaluation, and medical benchmark datasets. Especially by using additional real medical data and RLAIF, HuatuoGPT outperforms other ChatGPT models in most cases.

Li et al. (Li et al., 2023b) proposed that the ChatDoctor model was based on the LLaMA model and added medical field data for instruction fine-tuning pre-training. At the same time, a sample of fine-tuning instructions for generating structured medical knowledge maps using chatGPT is given, which is representative of fine-tuning training of large medical models. Fine-tuning training is carried out based on the LLaMA model. The sample data uses 100 k online real doctor-patient dialogues, and at the same time adds the ability of independent knowledge retrieval. For example, from Wikipedia or disease databases, fine-tuning the training model significantly improves understanding of patient needs and providing suggestions, and the ability to autonomously retrieve knowledge can access authoritative information in real-time, and prompt the accuracy of the model's answers. This is important in the medical field where error tolerance is low.

Since the number of existing medical image-text datasets is several orders of magnitude lower than the general images and captions from the Internet, Wang et al. (Wang et al., 2022c) proposed the MedCLIP model. It decouples images and text for multimodal contrastive learning, thus scaling the available training data at a low-cost combined magnitude. And propose to replace InfoNCE loss with medical knowledge-based semantic matching loss to eliminate false negatives in contrastive learning. Results show that it outperforms state-of-the-art methods in zero-shot prediction, supervised classification and image-to-text retrieval. Wang et al. (Wang et al., 2023b) proposed the ChatCAD model, combining the advantages of LLM's medical domain knowledge and logical reasoning with the visual understanding ability of existing medical image CAD models. Create a system for patients that is more user-friendly and understandable than traditional CAD systems. In this way, patients can better understand their conditions, reduce consultation expenses for patients, and enhance the feasibility of online medical services. The ChatCAD network structure diagram is shown in Fig. 7. Firstly, the examination images (such as X-rays) are fed into the trained disease classification model, image segmentation model, and reporter model. Then, the report generation model is improved and output is based on the results of the disease classification model and image segmentation model. Then, the tensors of these outputs are converted into natural language, and the language model is used to summarize the results and draw the final conclusion. Although the above-mentioned large models have achieved certain results in the medical field, they also face some challenges and limitations. The training and application of large models require a large amount of computing resources and data support, which may be an expensive investment for some medical institutions and research teams. In addition, due to the complexity and diversity of the medical field, these large models may suffer from insufficient generalization when dealing with domain-specific problems, resulting in poor performance on specific tasks.

4. Combination of transformer and other models

The transformer network is widely used in medical images due to its excellent performance. It has achieved good results in the clinical auxiliary diagnosis of major diseases such as benign and malignant tumors, lung cancer, breast cancer, skin diseases, and cardiovascular and cerebrovascular diseases. Due to the limited performance of a single network, more and more studies have found that the combination of transformer and other network models is a significant development direction. In this paper, we examine the segmentation, classification, reconstruction, denoising, and fusion of medical diagrams from the perspective of a Transformer combined with other network models. With the progress of the research in this field, it is expected that the further improvement and adaptation of the Transformer model will promote the latest technologies in the medical image field, and ultimately benefit various clinical applications and promote better medical diagnosis and treatment.

4.1. Field of medical image segmentation

The deep convolution neural network shows good performance in image segmentation. Due to the limitation of the CNN receptive field, it only obtains local features and cannot capture long-distance dependence. Moreover, The fixed size and form of the convolution kernel prevent it from adequately adapting to the type of input image, which restricts the convolution's generalization space and lessens the generalization of the segmentation model (Fu et al., 2023). Medical images also have problems such as blurred borders, low contrast, different object sizes, and diverse modalities. To effectively solve the above issues, it is essential to obtain critical global context information. Therefore, the transformer, which uses the self-attention mechanism to get global features from the NLP field, is used to optimize the automated segmentation technology of medical images. Xie et al. proposed the DeTrans (Xie et al., 2021) model, which effectively connects the convolutional neural network and transformer (CoTr) for accurate 3D

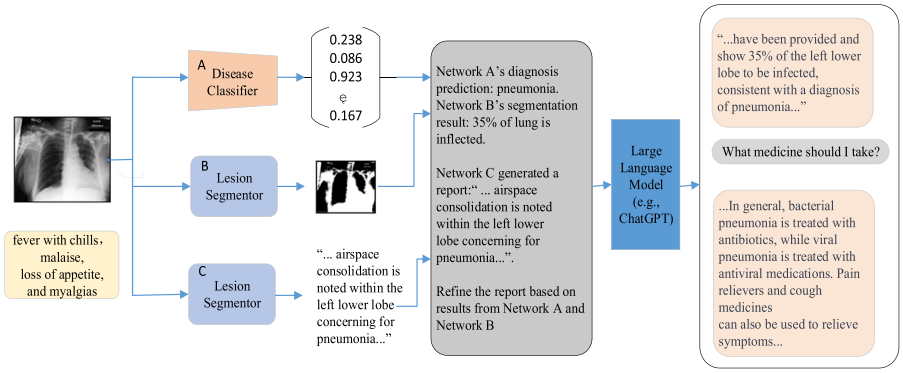

Fig. 7. Overview of ChatCAD.

medical image segmentation.

Under this framework, CNN is constructed to extract feature representations, and an efficient Deformable Transformer (DeTrans) is built to model long-range dependencies on the extracted feature maps. The TransBTS (Wang et al., 2021b) model was proposed by Wang et al. The encoder employs 3DCNN to extract 3D spatial feature maps to acquire local 3D context information. At the same time, the feature map is carefully modified, and the processed vector is input into the transformer for global feature modeling. The decoder uses the features embedded by the transformer to perform progressive upsampling to predict detailed segmentation maps. Valanarasu et al. (Valanarasu et al., 2021) propose the medical transformer framework, including a gated axial attention model, which extends existing architectures by introducing additional control mechanisms in the self-attention module. To effectively train the model on medical images, a local-global training strategy is proposed to improve the performance further. Ji et al. (Ji et al., 2021b) proposed the MCTrans model, which can construct cross-scale contextual dependencies and appropriate semantic relations and be inserted into a UNet-like network, and the results show that the six challenging datasets reach the best performance.

Before the transformer was applied to medical image segmentation, the segmentation model based on CNN or a fully convolutional network showed excellent performance in the downstream tasks of major image segmentation, especially the U-Net (Qiu et al., 2022). However, Chen et al. (Chen et al., 2021) proposed Transformer sand U-Net (TransUNet) as a transformer shines in NLP tasks. The emergence of this model opens the application of transformers in the field of medical image segmentation. Since transformer performs better on large datasets, but most medical image data are small, research on further improving transformer modules to make them suitable for medical image processing has become one of the hottest research directions. Among them, one of the most effective methods is to combine the transformer with the U-shaped network. Using the U-shaped network to reduce the amount of calculation as much as possible can also effectively capture the characteristics of important information and fully tap the potential of the transformer and U-shaped network. The network structure diagram of TransUNet is shown in Fig. 8, the model takes advantage of the transformer's advantage of obtaining long-distance dependencies in low-resolution (LR) images and a symmetrical encoder-decoder structure to improve the model's automatic segmentation performance. However, since TransUNet directly uses the NLP transformer model, the image block size in its sequence is fixed. Due to a large number of attention calculations, the segmentation efficiency of TransUNet needs to be further improved. LeViT-UNet (Xu et al., 2021a), Transformer-Unet (Sha et al., 2021), BiTr-Unet (Jia & Shu, 2021), AFTer-UNet (Yan et al., 2022), and nnFormer (Zhou et al., 2021) have been proposed one after another. They are based on the combination and improvement of Transformer + Unet, basically focusing on the following point: 1) Global plus local to improve performance; 2) Reduce complexity and improve speed and performance; 3) Combine traditional feature extraction with CNN and

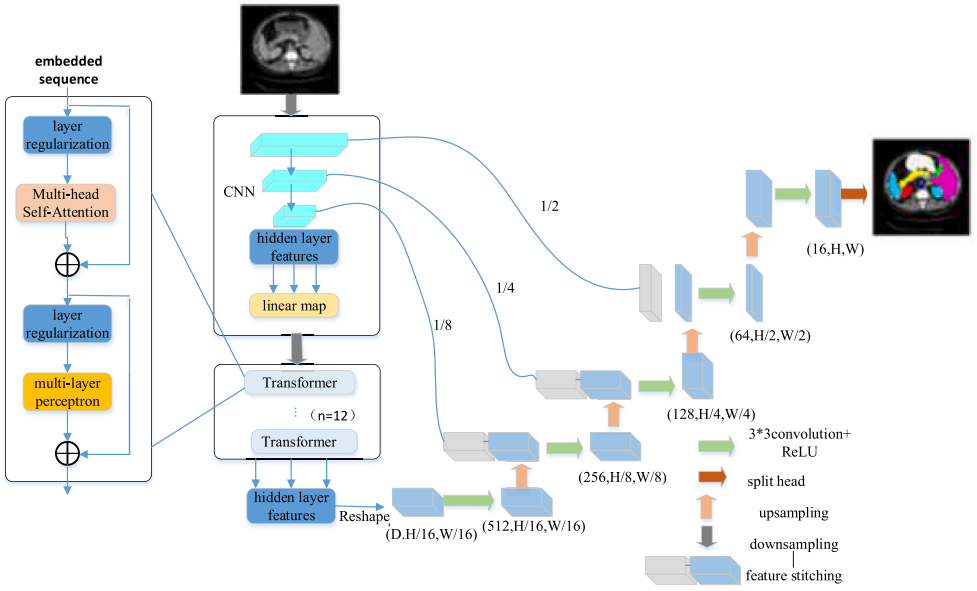

Fig. 8. Structure of TransUnit.

Transformer networks. Table 1 summarizes the application in the field of medical image segmentation in terms of literature, models, datasets, characteristics, and results.

4.2. Field of medical image reconstruction

High-resolution (HR) (Zhu et al., 2022b) medical images can provide richer lesion information and improve the accuracy of diagnosis. Due to the influence of many factors, the acquisition process of HR medical images is complex. In addition to the potential limitations of imaging hardware, medical images are more susceptible to physical and acquisition time constraints (Qiu et al., 2023b). For example, movement due to patient fatigue and organ pulsation further degrades image quality and results in lower signal-to-noise ratio (SNR) images. Therefore, medical imaging super-resolution (SR) methods are now becoming extremely important (Qiu et al., 2021b, 2023a; Zhu & Qiu, 2021). Yang et al. (Yang et al., 2022) proposed an LDCT denoising network, the Sinogram Inner Structure Transformer (SIST), to reduce noise by exploiting the internal structure in the sinogram domain. They also proposed a sinogram transformer module to extract sinogram features better. The transformer architecture using the self-attention mechanism can exploit the interrelationships between projections from different viewpoints to achieve superior performance in sinogram denoising. Image reconstruction and SR are two critical technologies in MRI, but many people ignore the commonalities between the two. Feng et al. (Feng et al., 2021) proposed an end-to-end task transformation network T2Net for combined MRI image reconstruction and SR, which allows representation and feature transfer to be shared across multiple tasks to obtain higher quality, SR and motion-artifact free images from highly under-sampled and degraded MRI data. This multi-task model yields better results than sequential combinations of state-of-the-art MRI reconstruction and SR models.

Table 1
Domain models in medical image segmentation.

Model	Dataset	Highlights	Experimental results
DeTrans (Qiu et al., 2021a)	BCV	The deTrans model effectively connects the convolutional neural network and transformer.	BCV: Average Dice 85
TransBTS (Wang et al., 2021b)	BraTs 2019/2020	Proposed TransBTS based on the encoder-decoder structure.	Avrege Dice (ET 78.92 WT 90.23 TC 81.19)
MedT (Valanarasu et al., 2021)	BrainUS, GlaSandMoNuSeg	Proposed a gated axial-attention model that extends the existing architectures by introducing an additional control mechanism in the self-attention module.	F1 88.84
MCTrans (Ji et al., 2021b)	Pannuke, CVC-Clinic, CVC-Colon, Etis, Kavirs, ISIC2018	MCTrans embeds the multi-scale convolutional features as a sequence of tokens, and performs intra- and inter-scale self-attention.	CVC-Clinic: DSC 92.30 CVC-Colon: DSC 86.58 Etis: DSC 83.69 Kavirs: DSC 86.20 ISIC 2018: DSC 90.35
TransUNet (Chen et al., 2021)	Synapse, ACDC	TransUNet merits both Transformers and U-Net.	Synapse: average DSC 77.48 ACDC: average DSC 89.71
LeViT-UNet (Xu et al., 2021a)	Synapse, ACDC	LeViT-UNet integrates a LeViT Transformer module into the U-Net architecture.	Synapse: average DSC 78.53 ACDC: average DSC 90.32
Transformer-Unet (Sha et al., 2021)	CT-82	Transformer-Unet by adding transformer modules in raw images instead of feature maps in Unet.	Dice 79.66
BiTr-Unet (Jia & Shu, 2021)	BraTS2021	CNN-Transformer combined model.	Median Dice 93.35
AFTer-UNet (Yan et al., 2022)	BCV, Thorax-85, SegTHOR	AFTer-UNet takes both advantage of the convolutional layers' capability of extracting detailed features and the transformers' strength on long sequence modeling.	Thorax-85: DSC 92.32 BCV: DSC 81.02 SegTHOR: DSC 92.10
nnFormer (Zhou et al., 2021)	Synapse, ACDC	nnFormer exploits the combination of interleaved convolution and self-attention operations.	Synapse: average DSC 87.40 ACDC: average DSC 91.78

Although SR methods based on convolutional neural networks perform well, many CNN-based MRI methods ignore the internal priors of MRI images (Ali et al., 2023). Zhang et al. (Zhang et al., 2022) proposed a 3D cross-scale feature transformer network (CFTN) to exploit cross-scale priors in MR features. This network consists of a plug-in mutual projection feature enhancement module (MFEM) to extract target scale features with HR (Zhu et al., 2022a) clues. The other is a spatial attention fusion module (SAFM) to adaptively adjust and fuse the target scale features and upsampling features extracted by MFEM and trunk, respectively. The CFTN network shows excellent results on the Kirby21 and BRATS (Menze et al., 2014) datasets. When Kirby21 is magnified by $\times 2$ largment, PSNR 39.70, SSIM0.9847, and BRATS are magnified by $\times 2$ largment. The results are better PSNR of 40.13 dB and SSIM of 0.9910. Li et al. (Li et al., 2022) suggested a unique transformer-enabled multi-scale context matching and aggregation network for multi-contrast MRI SR reconstruction. By using multi-scale context matching and combining reference features at several scales, this technique gives target LR characteristics enough supplemental information. The experimental results show that the method in this paper is superior to the current multi-contrast MRI super-resolution method (Qiu et al., 2021c), and it is expected to be applied in clinical medicine.

At present, there are two major challenges in the application of Transformer in the field of computer vision: First, the visual target changes greatly, and the performance of Transformer in different scenes may not be very good. Second, if the image resolution is relatively high and there are many pixels, the calculation based on global self-attention in Transformer will lead to a large amount of calculation (Ye et al., 2022). In response to the above two problems, SwinTransformer (Liu et al., 2021c) proposed a method that includes sliding window operation and hierarchically builds the Transformer. By limiting the attention calculation to one window, the sliding window operation can introduce the local perception of CNN convolution operation on the one hand, and save the calculation amount on the other hand. Furthermore, Swin-Transformer's strategy of controlling the computation area to the unit of Windows greatly reduces the computation of the network and reduces the complexity to a linear scale of the image size. Therefore, Swin-Transformers' variation was applied to medical image SR reconstruction and achieved good results.

Chao et al. (Yan et al., 2021) proposed a magnetic resonance imaging reconstruction model SMIR based on SwinTransformers. The model consists of two modules: a multi-layer feature extraction module and a reconstruction module, and combines the loss of frequency domain and spatial domain to reconstruct image details better. Pan et al. (Pan et al., 2022) proposed a multi-domain ensemble SwinTransformer network that combines rich domain features of data, residual data, images, and residual images. It can capture the global and local parts of the reconstructed image with better reconstructed image quality, feature

recovery, and edge protection. The advantage is that the features of the reconstructed image are wholly restored, and the disadvantage is that the features and edges require more computing and storage overhead. The transformer-based depth model has strict requirements on the size of the input image, which limits the flexibility of reconstruction.

Kai et al. (Jin et al., 2022) proposed SwiniPASSR, which employs SwinTransformer as the backbone while combining it with a Bidirectional Parallax Attention Module (biPAM) to maximize the auxiliary information provided by the binocular mechanism. A transformation layer is introduced in EvenTransformer and Parallax Attention Mechanism (PAM) to solve the integration problem. A progressive training strategy is adopted to learn other correspondence through gradually expanding receptive fields. Extensive experiments demonstrate the effectiveness of the proposed method, which achieves the state-of-the-art performance of PSNR of 24.13 dB and SSIM of 0.7579 on the validation set of Flickr1024 (Wang et al., 2019). Meanwhile, in the NTIRE2022 Stereo Image Super-Resolution Challenge (Wang et al., 2022b), SwiniPASSR achieves a PSNR of 23.71 dB and SSIM of 0.7295, ranking second on the leaderboard. The above network is shown in Table 2, which summarizes the field of medical image reconstruction in terms of literature, models, datasets, features, and results.

4.3. Field of medical image classification

In the field of medical image processing, image classification technology is widely used. With the development of computer vision technology and the increasing amount of medical image data, computer-aided doctor diagnosis and treatment has become an indispensable part of medicine. In the direction of medical image classification, Wang et al. (Wang et al., 2021c) designed a new hybrid model TransPath by combining a convolutional neural network and transformer. This model is pre-trained on the TCGA (Tomczak et al., 2015) and PAIP (Kim et al., 2021a) datasets using self-supervised learning and further fine-tuned on the MHIST (Wei et al., 2021), NCT-CRCHE (Nikolas et al., 2022), and PatchCamelyon (Veeling et al., 2018) datasets. The accuracy indicators reach 89.68%, 95.85%, and 89.91%, respectively. Chen et al. (Chen, 2022) proposed a model for classifying benign and malignant pulmonary nodules based on a multi-headed self-attentive mechanism. According to the small-scale characteristics of the medical image dataset, the general visual self-attention architecture of SwinTransformer in the visual transformer is adopted. The SwinTransformerBlock consists of a window-based multi-headed self-attentive layer (W-MSA) and a shift-window-based multi-headed self-attentive layer (SW-MSA). Compensates for the problem of no information interaction between windows caused by dividing fixed windows through a shift-window-based multi-headed self-attentive mechanism.

Dai et al. (Dai, Gao, & Liu, 2021) combined CNN and transformer to extract the image's low-level features, established the modality's long-distance dependence and proposed the TransMed model. The authors claimed this is the first work applying transformers to multimodal medical image classification, and they applied TransMed to the PGT dataset and the MRNet dataset (Bien et al., 2018) and divided the training set, validation set, and test set at a ratio of 7:1:2 in the PGT dataset; 1130:120:120 on the MRNet dataset training set, validation set, and test set. Finally, it reaches 88.9% accuracy on the PGT dataset and 85% on the MRNet dataset. Leamons et al. (Leamons et al., 2022) developed three different deep-learning models to detect the presence of invasive ductal carcinoma, the most common form of breast cancer. These models include convolutional neural networks (CNNs), residual neural networks (RNNs), and visual transformers (VTs) used as baselines. The experimental results of the three models using breast cancer tissue image datasets for training and experiments show that the VT model is superior to CNN and RNN in different tasks, with a classification accuracy of 93%. In contrast, the highest classification rate of other models is 87%. Jang et al. (Jang & Hwang, 2022) proposed a 3D medical image classifier M3T. It consists of a transformer, a 3D convolutional neural network, and a 2D convolutional neural network. On the ADNI (Petersen et al., 2010), AIBL (Ellis et al., 2009), and OASIS datasets (Marcus et al., 2007), the accuracy achieved results of 3.21%, 93.27%, and 85.26%. Today, medical image classification is still limited to applicable diseases and needs to be expanded to more conditions to assist medical diagnosis. The above network is shown in Table 3, which summarizes the literature, models, datasets, characteristics, and results in medical image classification.

Table 2
Domain models in medical image reconstruction.

Model	Dataset	Highlights	Experimental results
LDCT (Yang et al., 2022)	LDCT, Simulated	LDCT denoising network to reduce the noise by utilizing the inner structure in the sinogram domain.	LDCT: PSNR 41.80, SSIM 0.916 Simulated: PSNR 41.80, SSIM 0.916
T2Net (Feng et al., 2021)	IXI, Clinical	propose an end-to-end task transformer network (T2Net) for joint MRI reconstruction and SR.	IXI:PSNR 29.397, SSIM 0.872 Clinical: PSNR 30.400, SSIM 0.841
CFTN (Zhang et al., 2022)	Kirby21, BRATS	3D residual channel attention blocks (RCABs) + mutual-projection feature enhancement module (MFEM) + spatial attention fusion module (SAFM)	Kirby21:PSNR 39.70, SSIM 0.9847 BRATS: PSNR 40.13, SSIM 0.9910
McMRSR (Li et al., 2022)	pelvic, brain, fastMRI	Proposed a new multi-scale contextual matching method to capture corresponding contexts from reference features at different scales.	Pelvic: PSNR 36.23, SSIM 0.96 Brain: PSNR 36.07, SSIM 0.95 fastMRI: PSNR33.28, SSIM0.90
SMIR (Yan et al., 2021)	HCP	SMIR consists of two modules: a multi-level feature extraction module and a reconstruction module.	SSIM 0.91, MAE 25.82
MIST-net (Pan et al., 2022)	Real cardiac clinical datasets	SwinTransformer + U-Net	48- views: SNR 42.83, SSIM 0.9818 64-views: PSNR 43.64, SSIM 0.9844
SwiniPASSR (Jin et al., 2022)	Flickr1024, Stereo Image	SwinTransformer + biPAM	Flickr1024: PSNR 24.13, SSIM 0.7579 Stereo Image: PSNR 23.71, SSIM 0.7295

4.4. Other fields in medical imaging

In medical image denoising, Wang et al. (Wang et al., 2023a) proposed a model CTformer for low-dose CT denoising and conducted experiments on the MayoLDCT dataset (McCollough et al., 2017). SSIM achieved an excellent effect of 0.9121. CTformer is the first pure transformer model for low-dose CT denoising. Luthra et al. (Luthra et al., 2021) combined the learnable Sobel-Feldman operator and transformer and proposed a new architecture Eformer. Tested on the AAPM dataset and obtained a competitive result with SSIM of 0.9861.

In the direction of multimodal fusion, Li et al. (Li et al.) proposed a framework that can be used for a wide range of vision and language modeling. It aims to capture the rich semantics in images and related texts and has achieved good results in tasks such as visual question answering and visual reasoning, which belongs to the world's advanced level. Khare et al. (Khare et al., 2021) proposed a multimodal

Table 3
Medical image classification.

Model	Dataset	Highlights	Experimental results
TransPath (Wang et al., 2021c)	MHIST, NCT-CRCHE, PatchCamelyon	Proposed hybrid model (TransPath) and a token-aggregating and excitation (TAE) module.	MHIST: ACC 89.68% NCT-CRCHE: ACC 95.85% Patch Camel yon:ACC 89.91%
Swin-B (Chen, 2022)	LIDC-IDRI	W-MSA + SW-MSA	LIDC-IDRI: ACC 98.33%
TransMed (Dai, Gao, & Liu, 2021)	PGT, MRNet	TransMed combines the advantages of CNN and transformer.	PGT: ACC 88.9% MRNet: ACC 85%
VT (Leamons et al., 2022)	Breast cancer tissue images	CNN + RNN + VT	VT: ACC 93%
M3T (Jang & Hwang, 2022)	ADNI, AIBL, OASIS	3DCNN+2DCNN + Transformer	ADNI: ACC 3.21% AIBL: ACC 93.27% OASIS: ACC 85.26%

pre-training model for medical question-answering charges, which reduces the dependence on image annotations in a self-supervised manner. The model achieves state-of-the-art medical question-answering tasks.

Kim et al. (Kim et al., 2021b) proposed a visual-language transformer model for processing visual input without convolution, dramatically reducing the complexity of visual feature extraction and improving the model's speed. Current research lacks the content of medical multimodal pre-trained models, which can help encode images or text better. Modality fusion has organically combined multimodality, especially text and image data. These tasks are either used to enhance a modality-related task or to complete multimodal tasks. In representation learning, multimodal fusion has also promoted the development of this field, breaking through some previous limits. The comparative understanding between multiple modalities also provides new ideas for the future development of this segmented field.

5. Developments in the field of medical image processing

As a powerful deep learning model, the transformer has also played an important role in the field of medical image processing in recent years. The self-attention mechanism of the transformer model enables it to better understand the relationship between pixels in the image, which has advantages in image processing tasks. In tasks such as medical image classification, segmentation, and reconstruction, the transformer has shown its potential in processing complex medical images and improving the accuracy of analysis. There are still many problems to be solved in this field, and the relevant problems are manifested in the following aspects.

1) **Number of medical image datasets is often limited in size.** The gradual maturity of medical image processing technology is inseparable from supporting a large amount of medical image data because computer-aided diagnosis and treatment are inseparable from image feature extraction. A large number of extracted image features extensively help image processing technology. The more and more accurate the extracted image features are, the greater the impact on image processing. Therefore, medical image processing research is mainly based on big data. Unlike ordinary image data, medical images often involve the personal privacy of patients, which poses a problem for workers who collect medical images. Taking full advantage of the transformer's benefits in capturing long-distance dependencies requires a specific sample size, the majority of medical image datasets do not meet the requirements.
2) **Dimension of medical image.** With the rapid development of medical imaging technology, two-dimensional medical images can no longer meet the needs of doctors for diagnosis and treatment. The effect of purely using two-dimensional images for diagnosis and treatment is not apparent and sometimes misdiagnosed. Therefore, three-dimensional images are the mainstream and radiological images that doctors primarily study at this stage. Two-dimensional images are usually used to deal with medical image problems based on deep learning, so the introduction of three-dimensional images will inevitably bring significant challenges to deep learning researchers.
3) **Definition of medical images.** Medical images will inevitably be affected by human tissues, organs, and imaging equipment during imaging, so the generated medical images have low definition. Using such images for medical diagnosis will bring a significant burden to doctors. The effect of using deep learning to process such medical images will not be perfect, so the clarity of medical images significantly impacts image processing results using deep learning. Transformer was originally used to process sequence tasks in natural language. If it is used to process image tasks, images need to be serialized. However, medical images have high resolution and many pixels; after serialization, an excessively long sequence will be formed. Although ViT proposes a series of image patches, applying the sequence after HR medical images still results in much computation.

Combined with the current status of transformer network development and the challenges it faces, the following suggestions and prospects are put forward for future research.

1) **Semi-supervised or unsupervised learning.** Since the Transformer can extract global key features from large data sets, it can be used to conduct auxiliary task training in big data or learn features of existing images to generate high-confidence pseudo tags. This can alleviate the problem of small medical image datasets.
2) **Combine CNN with transformer.** The transformer has powerful global modeling capabilities, and long-distance information will not be weakened, but at the same time, its generalization performance is not as good as CNN due to the lack of correct inductive bias. Convolution is good at extracting details and can effectively process low-level features. However, to capture global information, it is often necessary to stack many convolutional layers, and the Transformer is good at grasping the whole, so combining CNN can enhance the locality of the Transformer. Future work based on the combination of the two will promote more breakthroughs in medical image tasks.
3) **Multimodal fusion of transformers.** Before Transformer, the multimodal task was mainly to extract image features through CNN, extract text features through RNN, and then fuse the two modalities. The emergence of Vision Transformer breaks the barriers between CV and NLP models, and the same model can be used for both images and text. Then to deal with multi-modal tasks, you can directly use the two modes to input into this model, and then connect your own downstream tasks, which saves a lot of time. Transformer is powerful in the multimodal field because its self-attention structure can adapt to various types of data. It makes all kinds of data perform better in terms of schema alignment.

The application of transformers in medical image processing is of great significance. As a powerful deep learning model, the transformer plays a key role in the fields of medical image classification, segmentation and reconstruction. With the continuous improvement of deep learning and transformer models, it is expected that more optimized versions for medical image processing will appear, further promoting innovation and progress in the field of medical image processing. Its

application in medical imaging will bring great potential for medical diagnosis and research, and provide patients with better medical services and health management.

Declaration of competing interest

The authors declare that they have no conflicts of interest.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 71673143 and 18ZDA327).

References

Ali, A., et al. (2021). Xcit: Cross-covariance image transformers. *Advances in Neural Information Processing Systems, 34*, 20014–20027. https://proceedings.neurips.cc/paper_files/paper/2021/hash/a655fbe4b8d7439994aa37ddad80de56-Abstract.html.

Ali, A., et al. (2023). Evaluation of awareness and knowledge regarding MRI safety among students in the faculty of applied medical science at Jazan University. *Journal of Radiation Research and Applied Sciences*, 1687–8507. https://doi.org/10.1016/j.jrras.2023.100669

Bernal, J., et al. (2012). Towards automatic polyp detection with a polyp appearancemodel. *Pattern Recognition, 45*(9), 3166–3182. https://doi.org/10.1016/j.patcog.2012.03.002

Bernal, J., et al. (2015). WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics, 43*, 99–111. https://doi.org/10.1016/j.compmedimag.2015.02.007

Bien, N., et al. (2018). Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Medicine, 15*(11), Article e1002699. https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002699.

Cao, H., et al. (2022a). Swin-unet: Unet-like pure transformer for medical image segmentation. In *Europeanconference on computer vision* (pp. 205–218). Cham: Springer Nature Switzerland. https://link.springer.com/chapter/10.1007/978-3-031-25066-8_9.

Cao, H., et al. (2022b). Swin-unet: Unet-like pure transformer for medical image segmentation. In *Europeanconference on computer vision* (pp. 205–218). Cham: Springer Nature Switzerland. https://link.springer.com/chapter/10.1007/978-3-031-25066-8_9.

Carion, N., et al. (2020). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213–229). Cham: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-030-58452-8_13.

Chang, Y., et al. (2021). Transclaw u-net: Claw u-net with transformers for medical image segmentation. arXiv preprint arXiv:2107.05188 https://doi.org/10.48550/arXiv.2107.05188.

Chen, J. M. (2022). *Research on classification method of benign and malignant pulmonary nodules based on attention mechanism.* Tianjin Normal University. https://doi.org/10.27363/d.cnki.gtsfu.2022.001052

Chen, C. F. R., et al. Crossvit: Cross-attention multi-scale vision transformer for image classification. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 357-366).https://openaccess.thecvf.com/content/ICCV2021/html/Chen_CrossViT_Cross-Attention_Multi-Scale_Vision_Transformer_for_Image_Classification_ICCV_2021_paper.html.

Chen, J., et al. (2021). *Transunet: Transformers make strong encoders for medical image segmentation.* https://doi.org/10.48550/arXiv.2102.04306. arXiv preprint arXiv: 2102.04306.

Chen, Y., et al. (2022). Mobile-former: Bridging mobilenet and transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5270–5279). https://openaccess.thecvf.com/content/CVPR2022/html/Chen_Mobile-Former_Bridging_MobileNet_and_Transformer_CVPR_2022_paper.html.

Codella, N., et al. (2019). *Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)*. https://doi.org/10.48550/arXiv.1902.03368. arXiv preprint arXiv:1902.03368.

Croce, D., et al. (2020). *GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples*. https://www.amazon.science/publications/gan-bert-generative-adversarial-learning-for-robust-text-classification-with-a-bunch-of-labeled-examples.

Dai, Y., et al. (2021a). Transmed: Transformers advance multi-modal medical image classification. *Diagnostics, 11*(8), 1384. https://doi.org/10.3390/diagnostics11081384

Dai, Y., Gao, Y., & Liu, F. (2021). Transmed: Transformers advance multi-modal medical image classification. *Diagnostics, 11*(8), 1384. https://www.mdpi.com/2075-4418/11/8/1384.

Dalmaz, O., et al. (2022). ResViT: Residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging, 41*(10), 2598–2614. https://doi.org/10.1109/TMI.2022.3167808

Devlin, J., et al. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. https://doi.org/10.48550/arXiv.1810.04805. arXiv preprint arXiv: 1810.04805.

Dosovitskiy, A., et al. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. https://doi.org/10.48550/arXiv.2010.11929. arXiv preprint arXiv:2010.11929.

Ellis, K. A., et al. (2009). The Australian imaging, biomarkers and lifestyle (AIBL) study of aging: Methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of alzheimer's disease. *International Psychogeriatrics, 21*(4), 672–687. https://doi.org/10.1017/S1041610209009405

Fan, J., et al. (2022). *A review of Transformer-based single-channel speech enhancement models[J]*. Computer Engineering and Applications, 2022 https://kns-cnki-net-s.vpn2.njau.edu.cn:8118/kcms2/article/abstract?v=3uoqIhG8C44YLTlOAiTRKibYlV5Vjs7iJTKGjg9uTdeTsOI_ra5_XQaS6ni5SKIh4qXfqDnJnwSEtppawT-nNE0wnaXVDUR8&uniplatform=NZKPT.

Feng, C. M., et al. (2021). Task transformer network for joint MRI reconstruction and super-resolution. In *Medical image computing and computer assisted intervention–MICCAI 2021:24th international conference, strasbourg, France, september 27–october 1, 2021, proceedings, Part VI 24* (pp. 307–317). Springer InternationalPublishing. https://link.springer.com/chapter/10.1007/978-3-030-87231-1_30.

Fu, L. Y., et al. (2022). A review of U-shaped medical image segmentation network based on Transformer. computer application. https://kns-cnki-net-s.vpn2.njau.edu.cn:8118/kcms2/article/abstract?v=3uoqIhG8C44YLTlOAiTRKibYlV5Vjs7ioT0BO4yQ4m_mOgeS2ml3UBV8gSVxETk7YVBPSWr7OISQbhTKsC50znNOrEQhyl4m&uniplatform=NZKPT.

Fu, L., Yin, M., & Yang, F. (2023). Transformer based U-shaped medical image segmentation network:a survey. *Journal of Computer Applications, 43*(5), 1584. http://www.joca.cn/EN/10.11772/j.issn.1001-9081.2022040530.

Gamper, J., et al. (2019). Pannuke: An open pan-cancer histology dataset for nuclei instance segmentationand classification. In *Digital pathology: 15th European congress, ECDP 2019* (Vol. 15, pp. 11–19). Warwick, UK: Springer International Publishing. April 10–13, 2019,Proceedings https://link.springer.com/chapter/10.1007/978-3-030-23937-4_2.

Graham, B., et al. (2021). Levit: A vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 12259–12269). https://openaccess.thecvf.com/content/ICCV2021/html/Graham_LeViT_A_Vision_Transformer_in_ConvNets_Clothing_for_Faster_Inference_ICCV_2021_paper.html.

Hao, H., et al. (2023). Renal ultrasound image segmentation method based on channel attention and GL-UNet11. *Journal of Radiation Research and Applied Sciences, 16*(3). https://doi.org/10.1016/j.jrras.2023.100631

Hiesinger, W., et al. (2023). *Almanac: Retrieval-Augmented language models for clinical medicine. Res sq [preprint]*. May 2:rs.3.rs-2883198. doi: 10.21203/rs.3.rs-2883198/v1. PMID: 37205549; PMCID: PMC10187428.

Hong, J. F. (2022). *A review of Transformer research status*. Information system engineering. https://kns-cnki-net-s.vpn2.njau.edu.cn:8118/kcms2/article/abstract?v=3uoqIhG8C44YLTlOAiTRKibYlV5Vjs7iJTKGjg9uTdeTsOI_ra5_XVMK3R9LcrUtqNMFJ5P0pQCGABTJnSeW9QPNWR_rqTec&uniplatform=NZKPT.

Jang, J., & Hwang, D. (2022). M3T: Three-dimensional medical image classifier using multi-plane and multi-slice transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20718–20729). https://openaccess.thecvf.com/content/CVPR2022/html/Jang_M3T_Three-Dimensional_Medical_Image_Classifier_Using_Multi-Plane_and_Multi-Slice_Transformer_CVPR_2022_paper.html.

Jha, D., et al. (2021). A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation. *IEEE journal of biomedical and health informatics, 25*(6), 2029–2040. https://ieeexplore.ieee.org/abstract/document/9314114.

Jia, Q., & Shu, H. (2021). Bitr-unet: A cnn-transformer combined network for mri brain tumorsegmentation. In *International MICCAI brainlesion workshop* (pp. 3–14). Cham: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-031-09002-8_1.

Ji, Y., et al. (2021a). Multi-compound transformer for accurate biomedical image segmentation. In *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference* (pp. 326–336). Strasbourg,France: Springer International Publishing. September 27–October 1, 2021, Proceedings,Part I 24 https://link.springer.com/chapter/10.1007/978-3-030-87193-2_31.

Ji, Y., et al. (2021b). Multi-compound transformer for accurate biomedical image segmentation. In *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference* (pp. 326–336). Strasbourg,France: Springer International Publishing. September 27–October 1, 2021, Proceedings,Part I 24 https://link.springer.com/chapter/10.1007/978-3-030-87193-2_31.

Jin, K., et al. (2022). SwiniPASSR: Swin transformer based parallax attention network for stereo image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 920–929). https://openaccess.thecvf.com/content/CVPR2022W/NTIRE/html/Jin_SwiniPASSR_Swin_Transformer_Based_Parallax_Attention_Network_for_Stereo_Image_CVPRW_2022_paper.html.

Khare, Y., et al. (2021). Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)* (pp. 1033–1036). IEEE. https://doi.org/10.1109/ISBI48211.2021.9434063.

Kim, Y. J., et al. (2021a). Paip 2019: Liver cancer segmentation challenge. *Medical Image Analysis, 67*, Article 101854. https://doi.org/10.5114/wo.2014.47136

Kim, W., et al. (2021b). Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning* (pp. 5583–5594). PMLR https://proceedings.mlr.press/v139/kim21k.html.

Leamons, R., et al. (2022). Vision transformers for medical images classifications. In *Proceedings of SAIIntelligent systems conference* (pp. 319–325). Cham: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-031-16075-2_22.

Lee, Y., et al. (2022). Mpvit: Multi-path vision transformer for dense prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7287–7296). openaccess.thecvf.com/content/CVPR2022/html/Lee_MPViT_Multi-Path_Vision_Transformer_for_Dense_Prediction_CVPR_2022_paper.html?ref= https://githubhelp.com.

Lee, J., et al. (2019). Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning* (pp. 3744–3753). PMLR http://proceedings.mlr.press/v97/lee19d.html.

Liang, J., et al. (2022a). TransConver: Transformer and convolution parallel network for developing automatic brain tumor segmentation in MRI images. *Quantitative Imaging in Medicine and Surgery, 12*(4), 2397. https://doi.org/10.21037/2Fqims-21-919

Liang, J., et al. (2022b). Btswin-unet: 3d u-shaped symmetrical swin transformer-based network for brain tumor segmentation with self-supervised pre-training. *Neural Processing Letters*, 1–19. https://link.springer.com/article/10.1007/s11063-022-10919-1.

Li, L. H., et al. Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557. https://doi.org/10.48550/arXiv.1908.03557.

Li, Y., et al. (2021a). X-net: A dual encoding–decoding method in medical image segmentation. In *The visual computer* (Vols. 1–11). https://link.springer.com/article/10.1007/s00371-021-02328-7.

Li, S., et al. (2021b). *Medical image segmentation using squeeze-and-expansion transformers*. https://doi.org/10.48550/arXiv.2105.09511. arXiv preprint arXiv:2105.09511.

Li, G., et al. (2022). Transformer-empowered multi-scale contextual matching and aggregation for multi-contrast MRI super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20636–20645). https://openaccess.thecvf.com/content/CVPR2022/html/Li_Transformer-Empowered_Multi-Scale_Contextual_Matching_and_Aggregation_for_Multi-Contrast_MRI_Super-Resolution_CVPR_2022_paper.html.

Li, Z., et al. (2023a). Lvit: Language meets vision transformer in medical image segmentation. In *IEEE transactions on medical imaging*. https://doi.org/10.1109/TMI.2023.3291719

Li, Y., et al. (2023b). ChatDoctor: A medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. *Cureus, 15*(6). https://assets.cureus.com/uploads/original_article/pdf/152858/20230724-24731-1v47a9.pdf.

Lin, T. Y., et al. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988). https://openaccess.thecvf.com/content_iccv_2017/html/Lin_Focal_Loss_for_ICCV_2017_paper.html.

Lin, A., et al. (2022). Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement, 71*, 1–15. https://ieeexplore.ieee.org/abstract/document/9785614.

Liu, Z., et al. (2021a). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022). https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.

Liu, Y., et al. (2021b). 3D deep attentive u-net with transformer for breast tumor segmentation from automated breast volume scanner. In *2021 43rd annual InternationalConference of the IEEE engineering in medicine & biology society (EMBC)* (pp. 4011–4014). IEEE. https://doi.org/10.1109/EMBC46164.2021.9629523.

Liu, Z., et al. (2021c). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022). https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.

Luthra, A., et al. (2021). *Eformer: Edge enhancement based transformer for medical image denoising*. https://doi.org/10.48550/arXiv.2109.08044. arXiv preprint arXiv: 2109.08044.

Marcus, D. S., et al. (2007). OpenAccess series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience, 19*(9), 1498–1507. https://doi.org/10.1162/jocn.2007.19.9.1498

McCollough, C. H., et al. (2017). Low-dose CT for the detection and classification of metastatic liver lesions: Results of the 2016 low dose CT grand challenge. *Medical Physics, 44*(10), e339–e352. https://aapm.onlinelibrary.wiley.com/doi/full/10.1002/mp.12345.

Mehta, S., & Rastegari, M. (2021). *Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer*. https://doi.org/10.48550/arXiv.2110.02178. arXiv preprint arXiv:2110.02178.

Menze, B. H., et al. (2014). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging, 34*(10), 1993–2024. https://doi.org/10.1109/TMI.2014.2377694

Nikolas, K. J., et al. (2022). 100, 000 Histological images of human colorectal cancer and healthy tissue (v0.1). https://doi.org/10.5281/zenodo.1214456.

Pan, J., Zhang, H., Wu, W., Gao, Z., & Wu, W. (2022). Multi-domain integrative swin transformer network for sparse-view tomographic reconstruction. *Patterns, 3*(6). https://www.cell.com/patterns/pdf/S2666-3899(22)00083-6.pdf.

Petersen, R. C., et al. (2010). Alzheimer's disease neuroimaging initiative (ADNI): Clinical characterization. *Neurology, 74*(3), 201–209. https://doi.org/10.1212/WNL.0b013e3181cb3e25

Qin, C. B., Song, Z. Y., Zeng, J. Y., Tian, L. F., & Li, F. (2021). Deeply supervised breast cancer segmentation with joint multi-scale and attention-residual[J]. *Optical Precision Engineering, 29*, 877–895, 04 https://kns-cnki-net-s.vpn2.njau.edu.cn:8118/kcms2/article/abstract?v=3uoqIhG8C44YLTlOAiTRKibYlV5Vjs7iy_Rpms2pqwbFRRUtoUImHcHuj-q1wNidDag0JyDChQ_01qQj28H8Qw1yekrz92CT&uniplatform=NZKPT.

Qiu, D., Cheng, Y., & Wang, X. (2021a). Progressive U-net residual network for computed tomography images super-resolution in the screening of COVID-19. *Journal of Radiation Research and Applied Sciences, 14*(1), 369–379. https://doi.org/10.1080/16878507.2021.1973760

Qiu, D., Cheng, Y., & Wang, X. (2022). Dual U-Net residual networks for cardiac magnetic resonance images super-resolution. *Computer Methods and Programs in Biomedicine, 218*, Article 106707. https://doi.org/10.1016/j.cmpb.2022.106707

Qiu, D., et al. (2021b). Multiple improved residual networks for medical image super-resolution. *Future Generation Computer Systems, 116*, 200–208. https://doi.org/10.1016/j.future.2020.11.001

Qiu, D., et al. (2021c). Gradual back-projection residual attention network for magnetic resonance imagesuper-resolution. *Computer Methods and Programs in Biomedicine, 208*, Article 106252. https://doi.org/10.1016/j.cmpb.2021.106252

Qiu, D., et al. (2023a). Residual dense attention networks for COVID-19 computed tomography images super-resolution. *IEEE Transactions on Cognitive and Developmental Systems, 15*(2), 904–913. https://ieeexplore.ieee.org/document/9837306.

Qiu, D., et al. (2023b). Progressive feedback residual attention network for cardiac magnetic resonance imaging super-resolution. *IEEE Journal of Biomedical and Health Informatics, 27*(7), 3478–3488. https://ieeexplore.ieee.org/abstract/document/10113677.

Radford, A., et al. (2018). *Improving language understanding by generative pre-training*. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf.

Radford, A., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*(8), 9. https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf.

Sanh, V., et al. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. https://doi.org/10.48550/arXiv.1910.01108. arXivpreprint arXiv: 1910.01108.

Sha, Y., Zhang, Y., Ji, X., & Hu, L. (2021). *Transformer-unet: Raw image processing with unet*. arXiv preprint arXiv:2109.08417 https://arxiv.org/abs/2109.08417.

Silva, J., et al. (2014). Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International Journal of Computer Assisted Radiology and Surgery, 9*, 283–293. https://link.springer.com/article/10.1007/s11548-013-0926-3.

Szegedy, C., et al. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9). https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Szegedy_Going_Deeper_With_2015_CVPR_paper.html.

Tang, Y., et al. (2022). Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20730–20740). https://openaccess.thecvf.com/content/CVPR2022/html/Tang_Self-Supervised_Pre-Training_of_Swin_Transformers_for_3D_Medical_Image_Analysis_CVPR_2022_paper.html.

Tomczak, K., et al. (2015). Review the cancer genome atlas (TCGA): An immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia, 2015*(1), 68–77. https://doi.org/10.5114/wo.2014.47136

Valanarasu, J., J, M., et al. (2021). Medical transformer: Gated axial-attention for medical image segmentation. In *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference* (pp. 36–46). Springer International Publishing. Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24 https://link.springer.com/chapter/10.1007/978-3-030-87193-2_4.

Vaswani, A., et al. (2017). Attention is all you need. In *Advances in neural information processing systems* (Vol. 30). https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Veeling, B. S., et al. (2018). Rotation equivariant CNNs for digital pathology. In *Medical image computing and computer assisted intervention–MICCAI 2018: 21st international conference* (Vol. 11, pp. 210–218). Granada, Spain: Springer International Publishing. September 16-20, 2018, Proceedings, Part II https://link.springer.com/chapter/10.1007/978-3-030-00934-2_24.

Wang, Y., et al. (2019). Flickr1024: A large-scale dataset for stereo image super-resolution, 0-0. In *Proceedingsof the IEEE/CVF international conference on computer vision workshops*. openaccess.thecvf.com/content_ICCVW_2019/html/LCI/Wang_Flickr1024_A_Large-Scale_Dataset_for_Stereo_Image_Super-Resolution_ICCVW_2019_paper.html.

Wang, W., et al. (2021a). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 568–578). https://openaccess.thecvf.com/content/ICCV2021/html/Wang_Pyramid_Vision_Transformer_A_Versatile_Backbone_for_Dense_Prediction_Without_ICCV_2021_paper.html.

Wang, W., et al. (2021b). Transbts: Multimodal brain tumor segmentation using transformer. In *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference* (pp. 109–119). Strasbourg, France: Springer International Publishing. September 27–October 1, 2021, Proceedings, Part I 24 https://link.springer.com/chapter/10.1007/978-3-030-87193-2_11.

Wang, X., et al. (2021c). Transpath: Transformer-based self-supervised learning for histopathological image classification. In *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference* (pp. 186–195). Strasbourg, France: Springer International Publishing. September 27–October 1, 2021, Proceedings, Part VIII 24 https://link.springer.com/chapter/10.1007/978-3-030-87237-3_18.

Wang, B., et al. (2022a). Multiscale transunet++: Dense hybrid u-net with transformer for medical image segmentation. *Signal, Image and Video Processing, 16*(6), 1607–1614. https://link.springer.com/article/10.1007/s11760-021-02115-w.

Wang, L., et al. (2022b). NTIRE 2022 challenge on stereo image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF conference on computer vision and*

pattern recognition (pp. 906–919). https://openaccess.thecvf.com/content/CVPR2022W/NTIRE/html/Wang_NTIRE_2022_Challenge_on_Stereo_Image_Super-Resolution_Methods_and_Results_CVPRW_2022_paper.html.

Wang, Z., et al. (2022c). *Medclip: Contrastive learning from unpaired medical images and text*. https://doi.org/10.48550/arXiv.2210.10163. arXiv preprint arXiv:2210.10163.

Wang, D., et al. (2023a). CTformer: Convolution-free Token2Token dilated vision transformer for low-dose CT denoising. *Physics in Medicine and Biology, 68*(6), Article 065012. https://doi.org/10.1088/1361-6560/acc000

Wang, S., et al. (2023b). *ChatCAD: Interactive computer-aided diagnosis on medical image using large language models*. https://doi.org/10.48550/arXiv.2302.07257. arXiv preprint arXiv:2302.07257.

Wang, Z., et al. (2003). Multiscale structural similarity for image quality assessment. In *The thrity-seventh asilomar conference on signals, systems & computers* (Vol. 2, pp. 1398–1402). Ieee, 2003 https://utw10503.utweb.utexas.edu/publications/2003/zw_asil2003_msssim.pdf.

Wei, J., et al. (2021). A petri dish for histopathology image analysis. In , *Virtual eventArtificial intelligence in Medicine19th international conference on artificial intelligence in medicine, AIME 2021* (pp. 11–24). Springer International Publishing. June 15–18, 2021, Proceedings https://link.springer.com/chapter/10.1007/978-3-030-77211-6_2.

Xie, & C, Y. (2022). *Overview of transformer research[J]. Computer knowledge and technology*. https://doi.org/10.14004/j.cnki.ckt.2022.0194

Xie, Y., et al. (2021). In *Cotr: Efficiently bridging cnn and transformer for 3D medical image segmentation.In medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference* (pp. 171–180). Springer International Publishing. Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24 https://link.springer.com/chapter/10.1007/978-3-030-87199-4_16.

Xiong, Y., et al. (2019). Reinforced transformer for medical image captioning. In *Machine learning in medical imaging: 10th international workshop, MLMI 2019, held in conjunction with MICCAI 2019, shenzhen, China* (Vol. 10, pp. 673–680). Springer International Publishing. October 13, 2019, Proceedings https://link.springer.com/chapter/10.1007/978-3-030-32692-0_77.

Xu, G., et al. (2021a). *Levit-unet: Make faster encoders with transformer for medical image segmentation*. https://doi.org/10.48550/arXiv.2107.08623. arXiv preprint arXiv:2107.08623.

Xu, G., et al. (2021b). *Levit-unet: Make faster encoders with transformer for medical image segmentation*. https://doi.org/10.48550/arXiv.2107.08623. arXiv preprint arXiv:2107.08623.

Yan, X., et al. (2022). After-unet: Axial fusion transformer unet for medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3971–3981). https://openaccess.thecvf.com/content/WACV2022/html/Yan_AFTer-UNet_Axial_Fusion_Transformer_UNet_for_Medical_Image_Segmentation_WACV_2022_paper.html.

Yang, L., et al. (2022). Low-dose ct denoising via sinogram inner-structure transformer. *IEEE Transactions on Medical Imaging, 42*(4), 910–921. https://doi.org/10.1109/TMI.2022.3219856

Yang, H., et al. (2023). CoT-TransUNet: Lightweight context Transformer medical image segmentation network. *Journal of Computer Engineering & Applications, 59*(3). https://doi.org/10.3778/j.issn.1002-8331.2205-0046

Yanping, L., et al. (2023). Digital subtraction angiography image segmentation based on multiscale Hessian matrix applied to medical diagnosis and clinical nursing of coronary stenting patients. *Journal of Radiation Research and Applied Sciences, 16*(3), 1687–8507. https://doi.org/10.1002/mp.12345

Yan, C., Shi, G., & Wu, Z. (2021). Smir: A transformer-based model for MRI super-resolution reconstruction. In *2021 IEEE international conference on medical imaging physics and engineering (ICMIPE)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICMIPE53131.2021.9698880.

Ye, M., et al. (2022). Forest fire detection algorithm based on an improved SwinTransformer. *Journal of Central South University of Forestry & Technology, 42*, 101–110. https://doi.org/10.14067/j.cnki.1673-923x.2022.08.010, 08.

Yu, J., et al. (2016). Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia* (pp. 516–520). https://dl.acm.org/doi/abs/10.1145/2964284.2967274.

Zhai, H., et al. (2022). Extraction of liver capsule and assessment of liver cirrhosis based on attention gate and hollow space pyamid pooling UNet model. *Chinese Medical Imaging Technology, 38*, 1385–1390. https://doi.org/10.13929/j.issn.1003-3289.2022.09.023, 09.

Zhang, J., et al. (2020). Attention gate resU-Net for automatic MRI brain tumor segmentation. *IEEE Access, 8*, 58533–58545. https://doi.org/10.1109/ACCESS.2020.2983075

Zhang, W., et al. (2022). 3D cross-scale feature transformer network for brain mr image super-resolution. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 1356–1360). IEEE. https://doi.org/10.1109/ICASSP43922.2022.9746092.

Zhang, H., et al. (2023). *HuatuoGPT, towards taming language model to Be a doctor*. https://doi.org/10.48550/arXiv.2305.15075. arXiv preprint arXiv:2305.15075.

Zhou, H. Y., et al. (2021). *nnformer: Interleaved transformer for volumetric segmentation*. https://doi.org/10.48550/arXiv.2109.03201. arXiv preprint arXiv:2109.03201.

Zhu, D., et al. (2022a). Dual attention mechanism network for lung cancer images super-resolution. *Computer Methods and Programs in Biomedicine, 226*, Article 107101. https://doi.org/10.1016/j.cmpb.2022.107101

Zhu, D., et al. (2022b). Feedback attention network for cardiac magnetic resonance imaging super-resolution. *Computer Methods and Programs in Biomedicine, 231*(14), Article 107313. https://doi.org/10.1016/j.cmpb.2022.107313

Zhu, D., & Qiu, D. (2021). Residual dense network for medical magnetic resonance images super-resolution. *Computer Methods and Programs in Biomedicine, 209*, Article 106330. https://doi.org/10.1016/j.cmpb.2021.106330