Intelligent Medicine 3 (2023) 59–78

Contents lists available at ScienceDirect

Intelligent Medicine

journal homepage: www.elsevier.com/locate/imed

Review

Transformers in medical image analysis

Kelei He[1,2,#], Chen Gan[2,#], Zhuoyuan Li[1,2,#], Islem Rekik[3,4,#], Zihao Yin[2], Wen Ji[2], Yang Gao[2,5], Qian Wang[6,*], Junfeng Zhang[1,2,*], Dinggang Shen[6,7,8,*]

[1] Medical School of Nanjing University, Nanjing, Jiangsu 210093, China
[2] National Institute of Healthcare Data Science at Nanjing University, Nanjing, Jiangsu 210093, China
[3] BASIRA Laboratory, Faculty of Computer and Informatics Engineering, Istanbul Technical University, Istanbul, Turkey
[4] School of Science and Engineering, Computing, University of Dundee, UK
[5] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, Jiangsu 210093, China
[6] School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China
[7] Department of Research and Development, Shanghai United Imaging Intelligence Co., Ltd., Shanghai 200030, China
[8] Shanghai Clinical Research and Trial Center, Shanghai 201703, China

ARTICLE INFO

Keywords:
Transformer
Medical image analysis
Deep learning
Diagnosis
Registration
Segmentation
Image synthesis
Multi-task learning
Multi-modal learning
Weakly-supervised learning

ABSTRACT

Transformers have dominated the field of natural language processing and have recently made an impact in the area of computer vision. In the field of medical image analysis, transformers have also been successfully used in to full-stack clinical applications, including image synthesis/reconstruction, registration, segmentation, detection, and diagnosis. This paper aimed to promote awareness of the applications of transformers in medical image analysis. Specifically, we first provided an overview of the core concepts of the attention mechanism built into transformers and other basic components. Second, we reviewed various transformer architectures tailored for medical image applications and discuss their limitations. Within this review, we investigated key challenges including the use of transformers in different learning paradigms, improving model efficiency, and coupling with other techniques. We hope this review would provide a comprehensive picture of transformers to readers with an interest in medical image analysis.

1. Introduction

Transformers [1] have dominated the field of natural language processing (NLP), with applications in areas including speech recognition [2], synthesis [3], text to speech translation [4], and natural language generation [5]. As an instance of deep learning architectures, the first transformer was introduced to handle sequential inference tasks in NLP. Whereas recurrent neural networks [6] (*e.g.*, long short-term memory network [7]) explicitly use a sequence of inference processes, transformers capture long-term dependencies of sequential data with stacked self-attention layers. Thus, transformers are both efficient, as they solve a sequential learning problem in one-shot, and effective, owing to the stacking of very deep models. Several transformer architectures trained on large-scale architectures have become popular for solving NLP tasks; these include Bidirectional Encoder Representations from Transformers, BERT [8] and GPT-3 [9-10], to name just two.

Convolutional neural networks (CNNs) and their variants have achieved state-of-the-art (SOTA) performance in several computer vision (CV) tasks [11], partially owing to their progressively enlarged receptive fields that can learn hierarchies of structured image representations as semantics. Capturing vision semantics in images is usually regarded to be the core idea enabling building of successful networks in CV [12]. However, the long-term dependencies within images, such as the non-local correlation of objects in the image, are neglected in CNNs. Inspired by the aforementioned success of transformers in NLP, Dosovitskiy et al. [13] developed the vision transformer (ViT) by formulating image classification as a sequence prediction task for the image patch (region) sequence, thereby capturing long-term dependencies within the input image. ViT and its derived instances have achieved SOTA performance on several benchmark datasets. Transformers have become very popular across a wide spectrum of CV tasks, including image classification [13], detection [14], segmentation [15], generation [16], and captioning [17]. Furthermore, transformers have an important role in video-based applications [18].

Recently, transformers have also cross-pollinated the field of medical image analysis, where they are used for disease diagnosis [19–21] and

* Corresponding authors: Junfeng Zhang, Medical School of Nanjing University, Nanjing, Jiangsu 210093, China (Email: jfzhang@nju.edu.cn); Dinggang Shen, School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China (Email: Dinggang.Shen@gmail.com); Qian Wang, School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China (Email: wangqian2@shanghaitech.edu.cn).
# These authors contributed equally to this work.

https://doi.org/10.1016/j.imed.2022.07.002

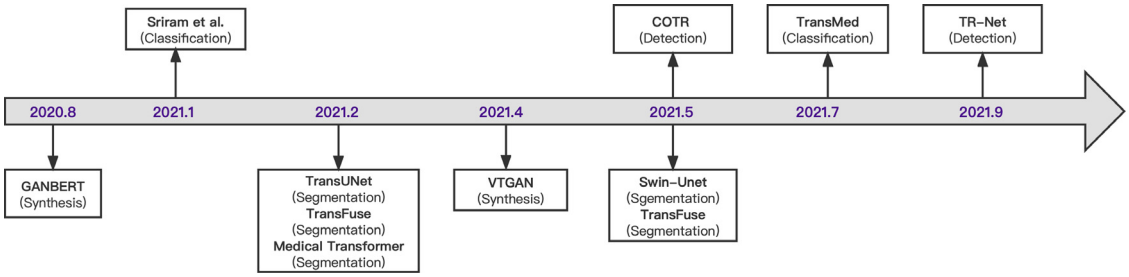

Figure 1. The development of transformers in medical image analysis. Selected methods are displayed relating to classification, detection, segmentation, and synthesis applications.

other clinical purposes. For instance, the work described in [22-23] used transformers to distinguish COVID-19 from other types of pneumonia using computed tomography (CT) or X-ray images, meeting the urgent need to treat COVID-19 patients fast and effectively. Transformers have also been successfully applied to image segmentation [24], detection [25], and synthesis [26], achieving SOTA results. Figure 1 displays the chronological adaptation of transformers to different medical image applications, which will be further discussed in Section 3.

Although many studies have been devoted to customizing transformers for medical image analysis tasks, this customization raised new challenges that remain unsolved. To encourage and facilitate the development of transformer-based applications in medical image analysis, we extensively review more than 170 existing transformer-based methods in the field, providing solutions for medical applications, and showing how transformers have been adopted in various clinical settings. Moreover, we present in-depth discussions on the design of transformer-based methods to solve complex real-world tasks, including weakly-supervised/multi-task/multi-modal learning paradigms. This paper also includes comparisons between transformers and CNNs and discusses new ways of improving the efficiency and interpretation of transformer networks.

The remainder of the paper is organized as follows. Section 2 introduces the preliminaries of transformers and their development in vision. Section 3 reviews recent applications of transformers in medical image analysis, and Section 4 discusses the potential future directions of transformers. Section 5 concludes the paper.

2. Transformers

2.1. Preliminaries

A typical transformer leverages the attention mechanism in neural networks. Hence, we start by introducing the core principle of the attention mechanism, followed by a detailed description of how the transformer works.

2.1.1. Attention mechanism

For information exploration, human beings usually leverage their "attention mechanism" to filter out irrelevant information while focusing on the meaningful parts of the data encountered in daily life. Inspired by this observation, researchers have designed attention mechanisms for deep learning that sift through homogeneous data while *paying attention* to the most significant components or elements.

Bahdanau attention. An attention mechanism was initially proposed in [27] for a language translation task, namely Bahdanau attention. This attention mechanism is calculated as the weighted sum of all annotations (*i.e.*, the results of each input generated by the encoder) and the previous decoder.

2.1.2. Attention mechanism in computer vision

Similar concepts have been developed in the field of CV. For example, Hu et al. [28] introduced a novel attention mechanism, *i.e., Squeeze-and-Excitation*, to execute *feature re-calibration*, in which informative features for a particular visual task are emphasized, and the remaining features are regarded as less important.

Self-attention. In [1], the attention mechanism was re-defined as a function working with queries, keys, and values derived from the input vectors of the module, in contrast to Bahdanau attention. The output is defined as a weighted sum of values, where the weight of each value is calculated as the attention between queries and keys.

The self-attention operation is usually performed in matrix form to accelerate calculation in parallel. To briefly illustrate the concept of self-attention, we first describe it in an element-wise form.

For each input $x_i \in \mathbb{R}^c$, $i = 1, .., n$, the corresponding query $q_i \in \mathbb{R}^d_q$, key $k_i \in \mathbb{R}^d_k$, and value $v_i \in \mathbb{R}^d_v$ vectors are generated through the parameters W^q, W^k, and W^v, respectively. d_q, d_k, d_v are the sizes of q_i, k_i, v_i and also the number of features that are learned from x_i.

$$\begin{aligned} q_i &= x_i \times W^q, \quad & W^q &\in \mathbb{R}^{c\times d_q}, \\ k_i &= x_i \times W^k, \quad & W^k &\in \mathbb{R}^{c\times d_k}, \\ v_i &= x_i \times W^v, \quad & W^v &\in \mathbb{R}^{c\times d_v}, \\ d_q &= d_k. \end{aligned} \tag{1}$$

The output is also a probability calculated as the weighted sum of the calculated weighting values:

$$\alpha_{ij} = \mathrm{Softmax}\left(\frac{\alpha'_{ij}}{\sqrt{d_k}}\right) = \frac{exp\left(\frac{\alpha'_{ij}}{\sqrt{d_k}}\right)}{\sum_j exp\left(\frac{\alpha'_{ij}}{\sqrt{d_k}}\right)}, \tag{2}$$

$$\alpha'_{ij} = q_i \times k_j^{\mathrm{T}}, \tag{3}$$

where α'_{ij} measures the contribution of the j^{th} element of the input to the i^{th} element of the output. Through this operation, α'_{ij} can be regarded as the attention assigned to the element v_i. Thereby the final output attentions can be computed as a weighted sum of all values as follows:

$$z_i = \sum_j \alpha_{ij} \times v_j. \tag{4}$$

The element-wise self-attention can be feasibly extended to matrices. In most cases, the query q_i, key k_i and value v_i for each input x_i are generated using parallel matrix computation. x_i, q_i, k_i, v_i can be stacked together to matrices, respectively. Let $X \in \mathbb{R}^{s\times c}$ denote the input matrix, Q denote the query matrix, K denote the key matrix, and V denote the value matrix, where s is the number of the samples and each matrix is consisted of the elements, *i.e.*, $X = [x_1; x_2; \cdots; x_s]^T$. Similarly, we compute the attention matrix A and output matrix Z as follows:

$$A = \mathrm{Softmax}\left(\frac{Q \times K^{\mathrm{T}}}{\sqrt{d_k}}\right) \in \mathbb{R}^{s\times s}, \tag{5}$$

$$Z = A \times V \in \mathbb{R}^{s\times d_v}. \tag{6}$$

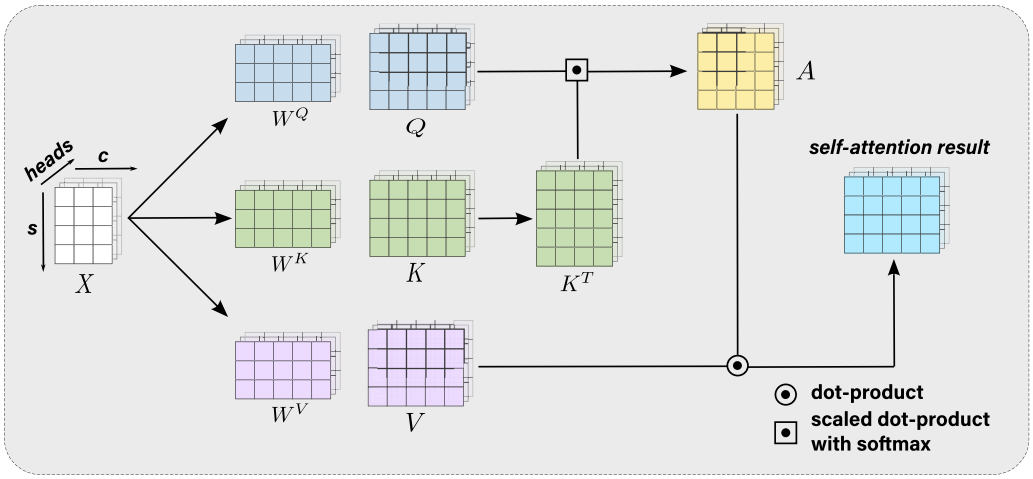

Figure 2. A brief illustration of a self-attention mechanism.

Multi-head self-attention. It was shown in [1] that applying multiple self-attentions to the same input could better capture hierarchical features. These self-attention layers work similarly to multiple kernels in convolution layers. Given h self-attentions (heads), the module outputs the final result by concatenating the calculated attentions:

$$Z_i = \text{Attention}(Q \times W_i^Q, K \times W_i^K, V \times W_i^V), \tag{7}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(Z_1, \cdots, Z_h)W^O, \tag{8}$$

where W_i^Q, W_i^J, W_i^V denote linear projection matrices that map matrices Q, K, V into different subspaces, respectively. W^O is an output projection matrix that concatenates self-attention outputs of all attention heads (Figure 2).

2.2. Architecture

In [1], the authors proposed a typical *transformer* network with an encoder–decoder structure. The encoder maps an input sequence $\{x_1, \ldots, x_n\}$ to an output sequence $\{z_1, \cdots, z_n\}$ of the same length. The decoder generates the output $\{y_1, \cdots, y_m\}$ from the encoded representation z in an element-wise manner and takes the previous output as an additional input. A typical transformer architecture is shown in Figure 3 and described below.

2.2.1. Encoder

The encoder in a typical transformer has $n = 6$ stacked blocks consisting of two types of layers, *i.e.*, the multi-head attention layer and the feed-forward layer. Residual connections and layer normalization layers are combined with the aforementioned layers. Concretely, in each block, the multi-head attention is first calculated, followed by a layer-wise normalization, calculating the sum of the input and output of the multi-head attention. This is followed by a feed-forward layer, then a layer-wise normalization of the sum of the feed-forward layer's input and output.

2.2.2. Decoder

The decoder also has $n = 6$ blocks, similar to the encoder, with some minor modifications. Specifically, an additional self-attention layer is inserted on top of the encoded output. Masking is employed in the first self-attention layer to block subsequent contributions to the state of the previous position, as the prediction is based on a known state. A linear layer and a Softmax layer are inserted after the output of the decoder to generate the final output.

2.3. Vision transformers

The success of transformers in NLP propagated to the CV research community, where several efforts have been made to adapt transformers to vision tasks. Transformer-based models in vision have been developed at an unprecedented pace; the most representative such models are detection transformer (DETR) [14], ViT [13], data-efficient image transformer (DeiT) [30], and Swin-Transformer [31].

DETR. DETR, proposed by Carion et al. [14], was the first application of transformers to a CV task, specifically the task of object detection. Unlike conventional object detection methods that involve hand-crafted processes, DETR is an end-to-end detection model that uses a transformer encoder to model the relation between image features extracted by a CNN backbone, a transformer decoder to generate object queries, and a feed-forward network to assign labels and bound the boxes around the objects.

ViT. Following DETR, Dosovitskiy et al. [13] proposed the ViT, as shown in Figure 4. ViT is an image classification model that adopts the basic architecture of the conventional transformer. In ViT, the input image is converted to a series of patches, each coupled with a positional encoding method that encodes the spatial positions of each patch to provide spatial information. The patches, along with a class token, are then fed into the transformer to calculate the MHSA and output the learned embeddings of patches. The state of the class token from the output of the ViT serves as the image representation. Last, a multi-layer perceptron (MLP) is used to classify the learned image representation. In addition to raw images, feature maps from CNNs can be fed into a ViT for relational mapping.

DeiT. In order to solve the problem of large-scale training data being required by ViT, Touvron et al. [30] proposed DeiT to ensure performance on small-scale data. They adopted a knowledge distillation framework with a teacher-student formulation and attached a distillation token (this is terminology for transformers) after the input sequence to learn from the output of the teacher model. In addition, they argued that using a CNN as the teacher model could facilitate training of the transformer as the student network to inherit inductive bias.

Swin-Transformer. To reduce the cost of calculating the attention of high-resolution images and deal with the varied patch sizes in scene-understanding tasks (*e.g.*, segmentation), Liu et al. [31] proposed the Swin-Transformer. They introduced a window self-attention to reduce the computational complexity and used the shifted window attention to model cross-window relationships. Moreover, they connected these attention blocks with patch merging blocks, which were used to merge neighboring patches to produce a hierarchical representation for handling variations in the scale of visual entities.

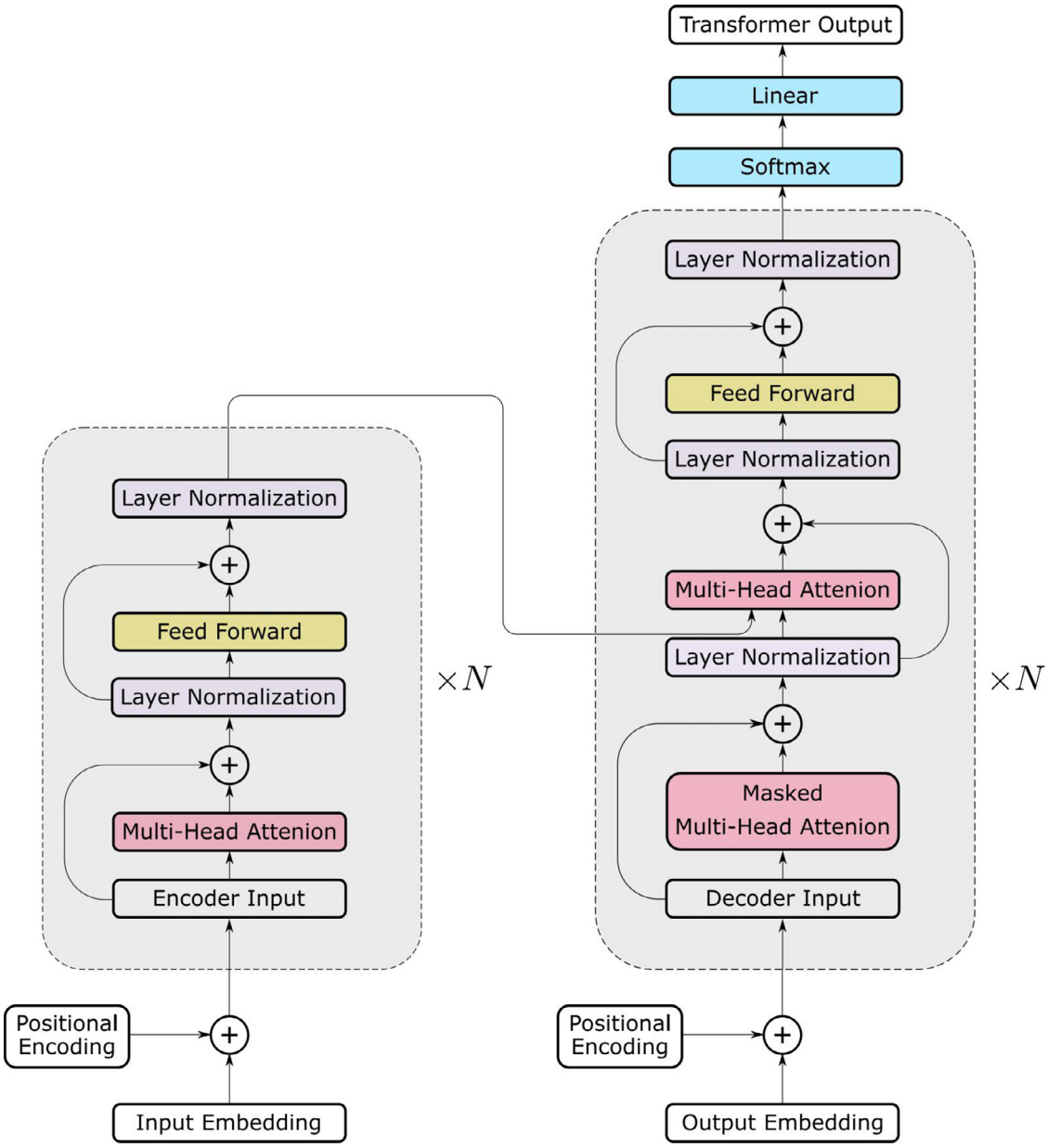

Figure 3. A brief illustration of a typical transformer architecture, as proposed in [29].

2.4. Other techniques

Recent studies have also validated MLP-based models and examined the effectiveness of attention mechanism, convolution, and other modules in CNNs or ViTs. Although CNNs and ViTs have been dominant for some time, the success of certain MLP-based models has had great repercussions. A representative example is MLP-Mixer, proposed by Tolstikhin et al. [32] in May 2021, which used a simple pure deep MLP architecture but showed competitive performance. MLP-Mixer uses per-patch flattening instead of the full flattening, and positional encoding and class token are not added to the patch sequence as in ViT. Following patch embedding learning, the Mixer MLP block is composed of a token-mixing MLP and a channel-mixing MLP, where the former is used to aggregate inter-patch features and the latter is used to integrate intra-patch features. The final class is predicted based on the features obtained following global average pooling.

Simultaneously with or following MLP-Mixer, many other MLP-based models have been proposed, *.e.g.*, gMLP [33], ResMLP [34], ASMLP [35], and CycleMLP [36]. MLP-Mixer not only inspired further exploration of MLP-based models but also led to further development of neural architectures in CV. As transformers, CNNs, and MLPs have shown competitive performance against each other, there is still no evidence as to which architecture is more suitable for particular CV learning tasks. In the case of medical image analysis, we provide a comparison of CNN and transformer models in part C of Section 4.

3. Transformers in medical image applications

Transformers have been widely used in full-stack clinical applications. In this section, we first introduce transformer-based medical image analysis applications, including classification, segmentation, image-to-image translation, detection, registration, and video-based applica-

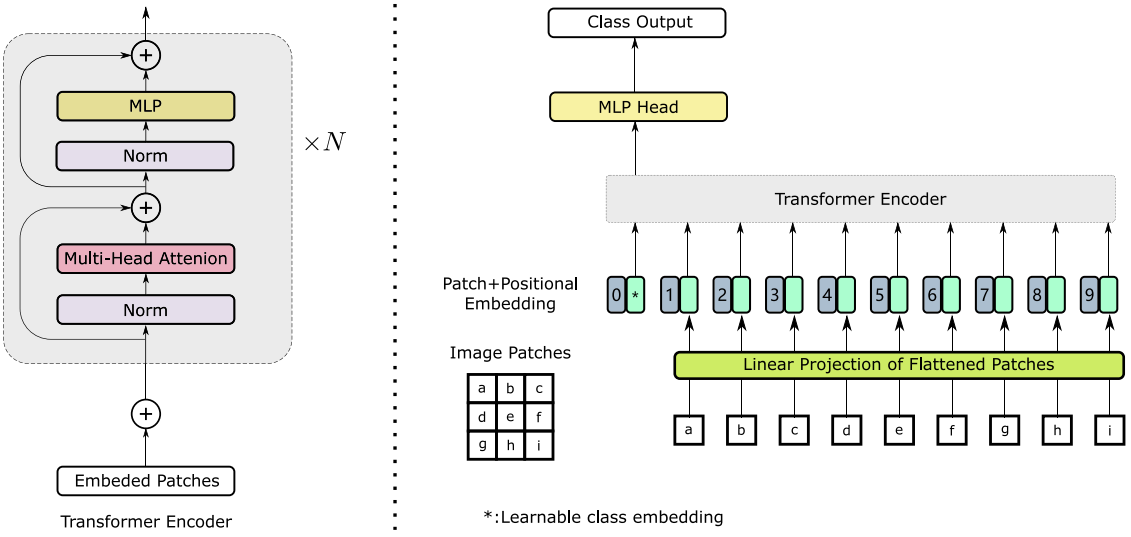

Figure 4. Architecture of ViT, as proposed in [13]. Sequential image patches are used as the input and processed with the transformer encoder, and the class prediction is output by an MLP head. The transformer encoder is constructed using N transformer blocks.

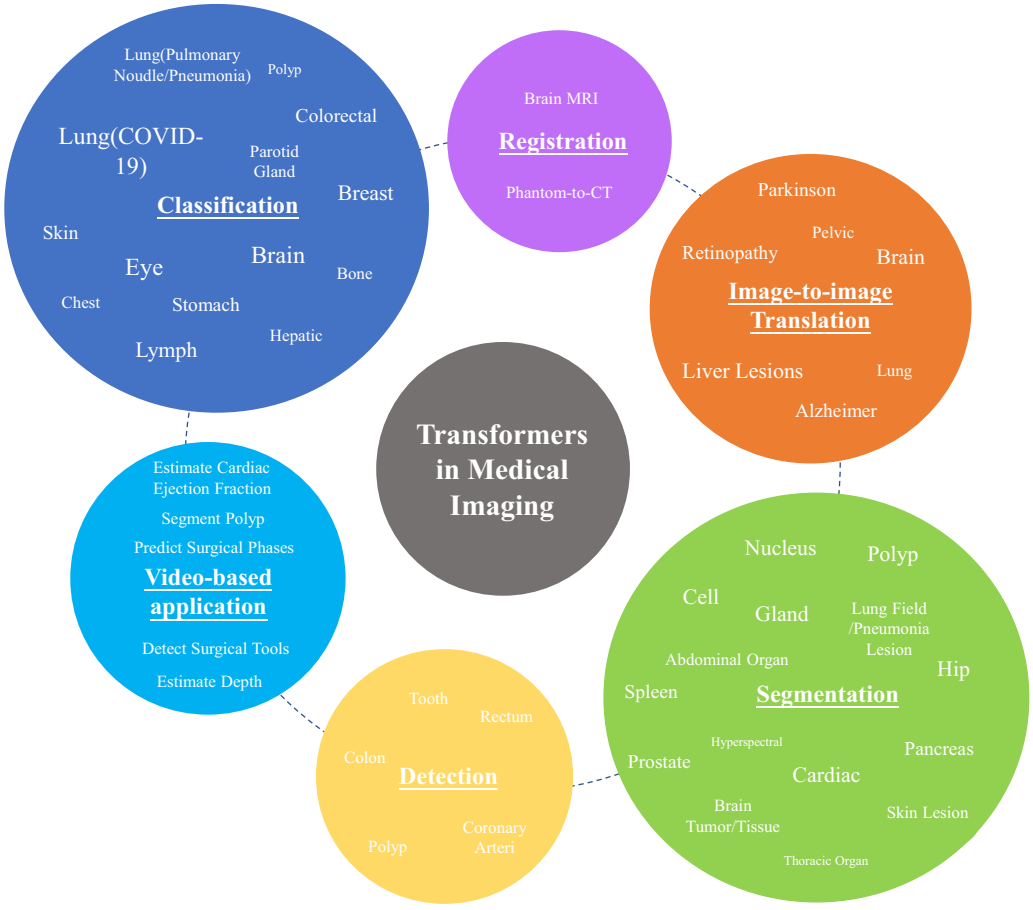

Figure 5. Applications of transformers in medical image analysis, as reviewed in this work.

tions. We categorize these applications according to their learning tasks as illustrated in Figure 5.

3.1. Classification

Methods using transformers for both disease diagnosis and prognosis are formulated as classification tasks, which can be divided into the following three categories:

(1) applying ViTs directly to medical images;

(2) combining ViTs with convolutions for more representative local feature learning;

(3) combining ViTs with graph representations to better handle complex data.

This section gives a comprehensive overview of the aforementioned three transformer categories used for classification tasks on medical images (Table 1).

Table 1 Transformers used in medical image classification tasks

References	Disease	Organ	Datasets	Highlight	Accuracy (%)
CT					
Costa et al. [22]	COVID-19	Lung	COVIDx	ViT with performer	91.0 91.0 96.0
COVID-VIT [19]	COVID-19	Lung	COV19-CT-DB	Use sub-volumes for 3D images	76.6
MIA-COV19D [20]	COVID-19	Lung	COV19-CT-DB	Segment lung first, use Swin-Trans	94.3
Liang et al. [37]	COVID-19	Lung	COV19-CT-DB	Feature aggregation by trans,CNN features, data resampling	-
Scopeformer [38]	Intracranial Hemorrhage	Brain	RSNA intracranial hemorrhage dataset	Multiple CNNs, GAN for domain alignment	98.0
Li et al. [39]	COVID-19	Lung	-	Teacher-student model for knowledge distillation	-
Than et al. [40]	COVID-19	Lung	COVID-CTset [41]	Research on patch size	95.4
Xia et al. [42]	Pancreatic Cancer	Pancreas	-	Anatomy-aware transformer with localization Unet	-
X-ray					
Park et al. [43]	COVID-19	Lung	-	Pretrained backbone on CXRs	-
Tanzi et al. [44]	Femur fracture	Bone	-	Unsupervised learning, compare CNNs with ViTs	77.0
Van et al. [23]	Mammography Chest X-ray	Breast Lung	CBIS-DDSM CheXpert	Trans combine multi-view info	–
Verenich et al. [45]	Chest X-ray	Lung	COVID-19 Radiology dataset [46–47]	Transformer × CNN	94.2 94.0
Liu et al. [48]	COVID-19	Lung	Cohen's dataset [49] COVID-19 database [47]	Outlooker attention	99.0 99.7
Shome et al. [50]	COVID-19	Lung	-	Grad-CAM-based visualization	98.0 92.0
Krishnan et al. [51]	COVID-19	Lung	COVID-19 X-ray database COVID19, Pneumonia and Normal Chest X-ray PA dataset	Large-scale COVID19 dataset; pretrained ViT-B/32 model	97.6
MRI					
He et al. [21]	*Brain Age*	Brain	BGSP, OASIS-3, NIH-PD, IXI ABIDE-I, DLBS, CMI, CoRR	Image-level and patch-level fusion with attention	-
Kim et al. [52]	*Gender classification Task decoding*	Brain Brain	HCP-Rest HCP-Task	Spatio-temporal attention for brain graph representation	88.2 87.0
mfTrans-Net [53]	Hepatocellular carcinoma	Hepatic	-	Trans combine multi-phase info; multi-level learning	-
3DMeT [54]	Knee cartilage defect	Knee	-	Generalize trans on 3D images	66.4 70.2
Histological Image					
Gao et al. [55]	Papillary renal cell carcinoma	Kidney	TCGA-KIRP	Instance-based patches; positions & grade encodings	89.2 93.0
Chen et al. [56]	Gastric histopatho-logical Image	Stomach	HE-GHI-DS	GIM and LIM modules; parallel structure	98.0
Zeid et al. [57]	CRC	Colorectal	Kather [58] colorectal cancer histology dataset	GIM and LIM modules; parallel structure	93.3 94.8
Ikromjanov et al. [59]	Prostate cancer	Prostate	Kaggle PANDA challenge dataset	Classify according to Gleason grading	-
Zhao et al. [60]	cervical cancer	Cell	a. Pap smear dataset b. SIPAKMeD c. Herlev	taming trans - T2T-ViT	-
Others					
POCFormer [61]	COVID-19	Lung	POCUS	Lightweight trans-based model	91.0 95.0 95.0
Gheflati et al. [62]	Breast Cancer	Breast	BUSI [63] Dataset B [64]	ViT on breast ultrasound images	85.7 86.0 86.7 85.0 86.4
Jiang et al. [65]	Acute lymphoblastic leukemia	Lymph	ISBI 2019 dataset	ViT and CNN ensemble	99.0
Xie et al. [66]	Melanoma	Skin	ISIC-2017 Skin dataset	SimAM with Swin-Trans	-
Li et al. [67]	Skin Lesion	Skin	HAM10000 DermNet	Trans on Out-of-Distribution Detection	-
Yu et al. [68]	Melanoma	Skin	ISIC 2020 dataset	Transformer × contrastive learning	-
Wu et al. [69]	Melanocytic lesions	Skin	MPATH-Dx	Encode multi-scale features with trans	60.0
TransEye et al. [70]	Fundus disease	Eye	OIA	Trans × CNN	84.1

3.1.1. Applications of pure transformers

We call ViTs that are similar to the originally proposed one [13] *p*ure transformers. These methods usually do not contain significant structural changes compared with the original method. We introduce the literature of pure transformers by image modality, *e.g.*, X-ray [44,48], CT [19-20], magnetic resonance imaging (MRI) [21], ultrasound [61], and optical coherence tomography (OCT) [71].

X-ray. X-ray is an inexpensive and convenient imaging technique that is widely used in screening and diagnosis of diseases including, breast cancer, pneumonia, and fracture. During the COVID-19 pandemic in particular, X-ray has played a very important part in disease screening and is thus a popular modality for AI researchers to use when designing transformer-based methods. Liu et al. [48] developed the vision outlooker (VOLO), a ViT model that replaced the original attention mechanism with the outlooker attention, as proposed in [72]. Their model achieved SOTA performance for the diagnosis of COVID-19 without pretraining on ImageNet. Shome et al. [50] proposed a ViT-based model for COVID-19 diagnosis that was trained on a self-collected large dataset of COVID-19 chest X-ray images. They also used Grad-CAM [73] to show the progression of COVID-19. Krishnan et al. [51] applied an ImageNet-pretrained ViT-B/32 network to distinguish COVID-19, using patches from chest X-ray images as inputs. Given the effectiveness of ViTs for COVID-19 diagnosis, Tanzi et al. [44] used a ViT model to classify femur fracture. Their work used clustering methods to validate the ability of

the ViT model to extract features and compared its performance against that of CNNs. The aforementioned models demonstrate the importance of large-scale datasets, which enhance the performance of transformers. Therefore, as the scale of the dataset for COVID-19-related tasks [48,50-51,73] was larger than that used for the femur fracture task [44], the performance on the COVID-19-related task was also higher.

Computed tomography. CT is based on the high contrast between gas and tissue and is commonly used for thoracic disease diagnosis. Thus, the applications of pure transformers to CT images have mainly focused on thoracic diseases. For example, Than et al. [40] studied the effect of patch size when using ViT for COVID-19 and diseased lung classification tasks. They found that the performance dropped with larger patch sizes, revealing a trade off between local and global information. The 32×32 patch resulted in the best accuracy. Costa et al. [22] used ViT and its variants to distinguish COVID-19 pneumonia and other pneumonia from normal cases. By comparing the performance of several models, they showed that pretrained models including DeiT [30] achieved competitive results. The conventional ViT and its variants using performer encoder also achieved good results, even without pretraining. Li et al. [39] designed a platform for COVID-19 diagnosis based on ViT. They converted CT images into a series of flattened patches to fit the input of ViT for diagnosis. They also adopted a teacher-student model to distill knowledge from a CNN pretrained on natural images. Gao et al. [19] applied ViT to both two-dimensional (2D) and 3D CT scans to diagnose COVID-19. They constructed an image sub-volume by extracting a fixed number of slices, thereby 'normalizing' imaging sequences with a varying number of slices. They also proved that the performance of ViT was better than that of DenseNet, which is a competitive CNN model. Zhang et al. [20] trained the popular Swin-Transformer on CT images. Specifically, the framework first segments the lung via a Unet and then feeds the lung region to the feature extractor. This strategy helped to reduce the computation burden of the transformer framework. The aforementioned works show the importance of pretraining for CT image classification tasks, as CT images are much harder to acquire than X-ray images. Also, methods that reduce computational complexity using attention mechanism are crucial to classification of CT images, owing to the large volume of the images.

Magnetic resonance imaging. MRI has a better imaging quality, particularly for subtle anatomical structures including vessels and nerves; however, acquisition of MRI images is time-consuming. As MRI is a powerful non-invasive imaging technology for soft tissues, it is commonly used in neuroimaging studies. For instance, He et al. [21] proposed a two-pathway network for brain age estimation. A global pathway was designed to capture the global contextual information from the brain MRI, whereas a local pathway was responsible for capturing fine-grained information from local patches. The local and global contextual representations were then fused by a global-local attention mechanism. Next, the concatenation of fused features and local patches was fed into a revised global-local transformer. MRI also has a wide spectrum of clinical applications, e.g., cancer diagnosis, which makes it a strong candidate modality for training ViTs.

Ultrasound. Ultrasound at point of care has expanded the range of applications of ultrasound, as specific positions are not necessary to acquire images. Perera et al. [61] developed a transformer-based architecture to diagnose COVID-19 based on ultrasound clips. To ensure memory and time efficiencies, they replaced the standard ViTs with Linformer, reducing the space time complexity from $O(n^2)$ for the conventional self-attention mechanism to $O(n)$. Moreover, ultrasound has became a prominent modality for imaging of breast cancer owing to its ease of use, low cost, and safety. Gheflati et al. [62] used ViTs to classify normal, malignant, and benign breast tissues based on ultrasound images. They also compared the performance of ViTs of various configurations against CNNs to demonstrate their efficiency.

Others. In addition to the above-mentioned imaging modalities, other imaging technologies have been adopted for the examination and diagnosis of specific diseases, *e.g.*, using dermoscopy images [66], fundus images [74], or histopathology images [59]. For instance, Xie et al. [66] aimed to detect melanoma using dermoscopy images. They combined the Swin-Transformer with a parameter-free attention module, SimAM, to learn better features for the target classification task. As the features fed into the classifier contained rich semantic information but lacked detailed information, they designed the output of the first three Swin-Transformer blocks as three SimAM blocks input separately; then, all SimAM block outputs including the final feature map were concatenated together to form the new final feature map, which served as the input to the final classification layer. Li et al. [67] evaluated the performance of transformers on out-of-distribution (OOD) detection tasks in medical image analysis. The original ViT and the DeiT with multi-head, soft distillation, and hard distillation are included in their work. The performance of these models on skin lesion datasets HAM10000 and DermNet showed the limited performance and safety critical problems of transformers on the OOD detection task. Ikromjanov et al. [59] used ViT to assist pathologists to grade prostate cancer according to the Gleason grading system on whole-slide histopathology images and reported promising results.

As shown in Table 1, despite the excellent performance of pure transformers in certain cases, *e.g.*, analysis of COVID-19 X-ray images, further development is necessary for other tasks.

3.1.2. Applications of hybrid transformers

Although pure ViTs can achieve promising results without much modification, there has been extensive exploration of the possibilities of combining ViTs with other learning components to better capture complex data distributions or achieve better performance. Typical cases include combinations of transformers with (1) convolutional layers and (2) graph representations. We next introduce both categories.

Transformers with convolutions. ViTs focus more on modeling the global relationship within the data, whereas conventional CNNs pay more attention to the local texture. These differences have inspired researchers to combine the advantages of ViTs and CNNs. In addition, the analysis of medical images involves not only the correlation of regions in the image but also subtle textures. Hence, many studies have explored CNN-ViT combinations.

Most applications have focused on the diagnosis of thoracic diseases, especially COVID-19 and related diseases. Benefiting from ViT's power of feature integration, Van et al. [23] used a transformer to conduct multi-view analysis of unregistered medical images in order to classify chest X-rays. They proposed a transformer-based approach that considered spatial information across different views at the feature-level by virtue of a trainable attention mechanism. They applied the transformer to intermediate feature maps produced by CNNs to retrieve features from one view and transfer them to another view. Thus, additional context was added to the original view without requiring pixel-wise correspondence. Their approach also contributed to a reduction in computational complexity by substituting a smaller number of visual tokens for the source pixels. Verenich et al. [45] introduced global spatial information from ViTs to CNNs for pulmonary disease classification, while preserving spatial invariance and equivariance. Liang et al. [37] used a CNN to mine effective features and a transformer for feature aggregation. In addition, an effective data sampling strategy can be used to reduce the size of the inputs while preserving sufficient information for diagnosis. Park et al. [43] designed a pretrained CNN backbone followed by a ViT for COVID-19 diagnosis. A large-scale public dataset for CXR classification was used for model pretraining. For the simple task of classifying thoracic diseases, existing methods are simple yet effective, with a CNN used to extract the features, followed by capture of high-level information with a transformer.

For applications other than COVID-19 diagnosis, Yassine et al. [38] combined several CNNs with a ViT by feeding extracted features into the ViT. They compared the number of CNNs as well as their pretraining configurations against the hybrid CNN-ViT model. Notably, they pretrained the CNN on images generated from the

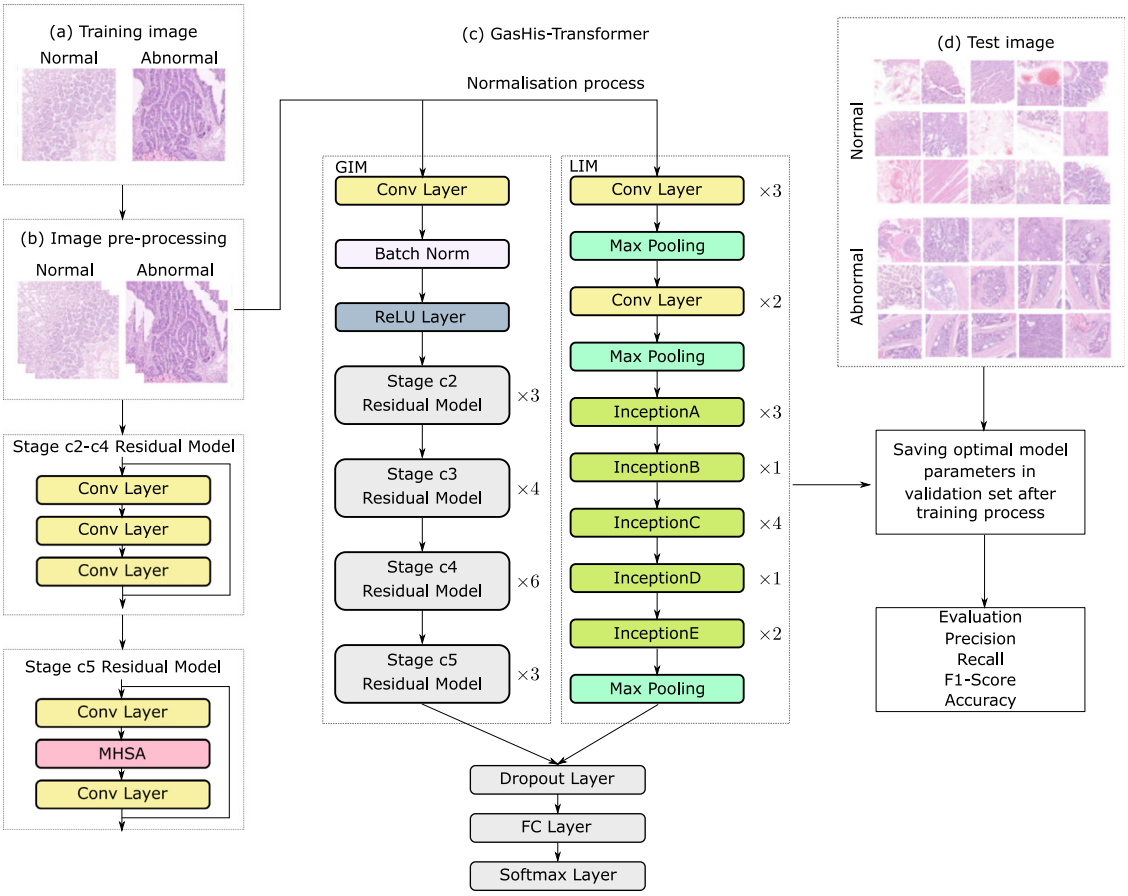

Figure 6. Structure of the GasHis transformer model [56].

ImageNet dataset [12] using a generative adversarial network (GAN) pretrained on brain CT images. They claimed that further pretraining on the generated images would lead to a better inductive bias for the target CT dataset as the dissimilarities of the two domains would be reduced. Zhao et al. [53] used a combination of CNNs and transformers to conduct multi-index quantification of hepatocellular carcinoma using multi-phase contrast-enhanced MRI (CEMRI). They proposed mrTrans-Net, which involves three parallel encoders, each followed by a non-local transformer that extracts features from the arterial phase, PV phase, and delay phase. Next, a phase-aware transformer is used to quantify the relevance of each phase for the target multi-phase CEMRI information fusion and selection. Quantification is conducted not only after the phase-aware transformer but also after the non-local transformers to form an enhanced loss function to constrain the quantification task. Jiang et al. [65] explored the effectiveness of ensemble learning by treating ViTs and CNNs as base learners to diagnose acute lymphoblastic leukemia based on microscopic images of B-lymphoid precursors and leukemic B-lymphoblast cells. They proposed an ensemble model based on the ViT and EfficientNet. As the two base models were complementary, the ensemble results showed some improvement. They also proposed a data enhancement method to handle the imbalance between normal and cancer cells in each image. Chen et al. [56] proposed the multi-scale ViT model shown in Figure 6, called GasHis-Transformer, for classification of gastric histopathological images. They designed a global information module (GIM) and local information module (LIM) (based on CNNs) to extract features. Moreover, they borrowed the parallel structure from Inception-V3 to learn multi-scale local representations. Their model was robust to ten different adversarial attacks or conventional noises and was generalizable to classification tasks of histopathological images of other cancers. Gao et al. [55] proposed the instance-based ViT (i-ViT) for papillary renal cell carcinoma subtyping. The i-ViT first extracts and selects instance features from instance-level patches, which include a nucleus with parts of the surrounding background and the nuclei grade. Next, it aggregates these features to further capture cell-level and cell-layer-level features. Last, the model encodes both types of fine-gained features into the final image-level representation, where grades and positions are embedded for subtyping. Wang et al. [54] proposed a 3D transformer that could outperform 3D CNNs. They used a 3D convolutional layer to extract features of 3D blocks and a teacher-student network to learn transformer weights from a CNN teacher. Xia et al. [42] proposed anatomy-aware transformers for pancreatic cancer screening, and showed to win the radiologists. Zeid et al. [57] validated ViTs and their variants compact convolutional transformers on a multi-class colorectal cancer (CRC) histology image classification task using a public CRC histology dataset. Zhao et al. [60] combined taming transformers with T2T-ViT to handle unbalanced samples with inconsistent image quality for a cervical cancer classification task. Yu et al. [68] adopted the transformer encoder to model dependency among features of skin lesions to detect the ugly duckling sign for melanoma identification. Yang et al. [70] proposed the transformer eye (TransEye) for fine-grained fundus disease image classification by combining CNN and transformer models. Wu et al. [69] proposed ScAT-Net to model inter-patch and inter-scale representations at multiple input scales to diagnose melanocytic lesions in biopsy images. These hybrid transformers for various applications contain rich innovations, including structural improvements, novel ViT modules, CNN modules, and learning strategies for pretraining and ensembling.

Transformers with graphs. Learning with graphs is a common practice in MIA. The core concept of graph learning is learning a compact representation of each sample (e.g., embeddings) while preserving the intrinsic inter-sample relationships via a data graph [75]. As an

attention-based network, transformer is suitable for operations on graph data, including aggregation of node features and calculation of node relationships.

In the field of network neuroscience, a brain network is modeled as a graph where each node denotes an anatomical region of interest (ROI) and the edge connecting two nodes encodes their interaction (e.g., neural firing). Brain graphs play an important part in advancing our understanding of the brain as a highly interconnected system in both health and disease [76-77]. Kim et al. [52] leveraged the dynamic characteristics of a functional connectivity network by incorporating dynamic features into a compact brain graph representation. Specifically, they proposed the spatio-temporal attention graph isomorphism network (STAGIN) for learning a dynamic graph representation of the brain connectome with spatio-temporal attention. In STAGIN, the GNN is used to extract graph-level representations for the functional brain connectome at each timestep, and a transformer encoder is used to obtain the final representation of a sequence of dynamic graphs. In detail, encoded timestamps are concatenated with node features to embed temporal information. The authors claim that the use of the transformer not only improved the classification performance of the model but also improved its spatial-temporal interpretability. Such methods have validated the power of transformers for mining both features and relationships of complex graphs, attracting more attention to this methodology.

We draw the following conclusions regarding the use of transformers for medical image classification tasks.

- Transformers have achieved performance comparable with or better than that of CNNs on most tasks.
- Transformers perform best on large-scale datasets, which somewhat limits their applicability, especially in the medical image analysis field. Pretraining could alleviate this problem.
- The computational burden of training transformers on large images is high. Hence, reducing model complexity and developing lightweight models are key factors to improve efficiency.
- Hybrid transformers have attracted increasing attention as they have the advantages of both conventional networks (*i.e.*, CNNs and GNNs) and transformers.

3.2. Segmentation

Transformer-based methods have also been applied to a variety of segmentation tasks, including abdominal multi-organ segmentation [25,78-80,82,86,93-94,96-97,106,113-114,119,121], thoracic multi-organ segmentation [114], cardiac segmentation [78,80,84,86-87,94,96-97,106,113,119,121,127], Pancreas segmentation [81,118], brain tumor/tissue segmentation [82,86,88,100,107-108,117-118,128-134], polyp segmentation [91,103,120,135-136], liver and hepatic lesion segmentation [86,93,117,137–140], kidney tumor segmentation [86,138], skin lesion segmentation [91,103,109,117,120,136,141], prostate segmentation [91,140], gland segmentation [24,95,100,109,120], nucleus segmentation [24,95,100,109,120], cell segmentation [103,142-143], spleen segmentation [107], lung field/COVID-19 pneumonia lesion segmentation [109], retinal vessel segmentation [144], and hyperspectral pathology image segmentation [145]. Several notable methods are listed and detailed in Table 2.

The U-shaped convolutional neural network architecture known as Unet has achieved tremendous success on most medical image segmentation tasks. However, owing to the use of convolution operations, Unet is also limited in its ability to model long-term dependencies. To overcome this limitation, researchers have designed robust hybrid transformers combined with the Unet architecture; these will be introduced in the first part of this section. Several methods also apply pure transformers to segmentation tasks; these will be introduced in the second part of this section.

3.2.1. Hybrid transformers

Most existing research on coupling transformers with the popular U-shaped architecture focuses on the following three aspects:

(1) inserting transformer layers at different levels of the U-shaped architecture;

(2) combining transformers and CNNs using different strategies;

(3) using multi-scale features or attention mechanisms.

We detail below each of these three categories.

3.2.1.1. Location of transformer in U-shaped architecture

An intuitive way to insert transformer layers into a U-shaped architecture is to insert a whole transformer between the encoder and decoder blocks to build long-term dependencies between high-level vision concepts. Based on this idea, Chen et al. [78] proposed TransUNet, shown in Figure 7, which extracts high-resolution spatial features using a CNN and then encodes the global context using a transformer. The self-attention features encoded by the transformer are then upsampled and combined with features at multiple scales extracted from the encoding path using skip connections for precise localization. TransUNet achieved superior performance compared with V-Net, AttnUNet, and ViT on multi-organ and cardiac segmentation tasks. Similarly, Yao et al. [79] combined a transformer network with a Claw Unet architecture; the resulting model outperformed TransUnet for synapse multi-organ segmentation. In another instance, Xu et al. [80] proposed LeViT-UNet, which integrates a LeViT Transformer into the Unet architecture. Sha et al. [81] designed a transformer-Unet by adding Transformer modules to Unet; the resulting model outperformed TransUnet.

In contrast to the above approaches, in which the transformer was inserted immediately after the encoder block, Li et al. [82] added an attention upsampling component to the decoder. They also proposed a window attention decoder and window attention upsampling, working on local windows, to reduce memory and computation costs. Gao et al. [84] presented a UTNet in which self-attention modules are applied to both encoder and decoder blocks to capture long-range dependencies at multiple scales with minimal overhead. They proposed an efficient self-attention mechanism along with relative position encoding, which reduced the complexity of the self-attention operation significantly from $O(n^2)$ to approximate $O(n)$. In an upgrade of their work, *i.e.*, UTNetV2 [87], they further proposed an efficient bidirectional attention. Fu et al. [86] proposed TF-Unet, which is built on the intertwined backbone of convolution and transformers at multiple scales. Several studies report improved strategies for feature concatenation [93,127].

3.2.1.2. Strategies for bridging transformers and CNNs

Unlike the aforementioned methods that combine transformers and U-shaped architectures within a single inference path, some studies have explored different transformer-CNN coupling strategies. Sun et al. [88] used Unet and transformer encoders to generate representations independently and then integrated their representations for subsequent decoding. Similarly, Li et al. [170] proposed X-Net, which used a CNN and a transformer to extract local and global features simultaneously. Zhang et al. [91] proposed TransFuse, which also combines transformers and Unet in a parallel style. In an improvement on the above-mentioned work, a novel fusion technique, *i.e.*, BiFusion module, was proposed to efficiently fuse multi-level features from both branches. Luo et al. [95] also used bidirectional cross-attention to fuse local information extracted by the convolution operations and global information learned by the self-attention mechanisms. Liu et al. [96] proposed PHTrans, which introduces a parallel hybrid module in deep stages, where convolution blocks and the modified 3D Swin-Transformer learn local features and global dependencies separately; then, a sequence-to-volume operation unifies the dimensions of the outputs to achieve feature aggregation (Figure 8).

Zhou et al. [97] claimed that most of the recently proposed transformer-based segmentation approaches simply treat transformers as assisted modules to help encode global context in convolutional

Table 2 Transformers for medical image segmentation tasks

References	Task	Dataset	Performance (%)	Highlight
TransUnet, Chen et al. [78]	a. ACT-MOS b. Cardiac segmentation	a. Synapse multi-organ CT b. ACDC	a. 77.48 b. 89.71	
TransClaw, Yao et al. [79]	ACT-MOS	Synapse multi-organ CT	78.09	Claw Unet
LeViT-Unet, Xu et al. [80]	a. ACT-MOS b. Cardiac segmentation	a. Synapse multi-organ CT b. ACDC	a. 78.53 b. 90.32	
LeViT Tunet, Sha et al. [81]	Pancreas segmentation	CT82 datasets	79.66	
Li et al. [82]	a. Brain tumor segmentation b. ACT-MOS	a. MSD-01 [83] b. Synapse multi-organ CT	a. 80.30 b. 74.75	
UTNet, Gao et al. [84]	Cardiac segmentation	M&Ms [85]	88.3	
TransBTSV2, Fu et al. [86]	Brain tumor segmentation	a. BraTS 2019 dataset b. BraTS 2020 c. LiTS2017 dataset (lesion, liver) d. KiTS2019 dataset (kidney, tumor)	a. 85.18 b. 84.90 c. 71.20, 96.20 d. 97.37, 83.69	
UTNetV2, Gao et al. [87]	Cardiac segmentation	ACDC	92.14	
HybridCTrm, Sun et al. [88]	Brain tissue segmentation	a. MRBrainS [89] b. iSeg-2017 [90]	a. 83.47 b. 87.16	Dual-Path Network
TransFuse, Zhang et al. [91]	a. Polyp segmentation b. Skin lesion segmentation c. Hip segmentation d. Prostate segmentation	a. KCCEE b. ISIC2017 [92] c. In-house dataset d. MSD dataset	a. 92.0, 94.2, 78.1, 89.4, 73.7 b. 87.2 c. - d. 76.4	Multi-level feature fusion
CA-GANformer,You et al. [93]	a. ACT-MOS b. Liver tumor segmentation	a. Synapse multi-organ CT b. LiTS dataset	a. 82.55 b. 73.82	GAN
ECT-NAS, Xu et al. [94]	a. ACT-MOS b. Cardiac segmentation	a. Synapse multi-organ CT b. ACDC	a. 78.97 b. 89.83	Searching strategy
HyLT, Luo et al. [95]	a. Gland segmentation b. Nuclear segmentation	a. GlaS dataset b. MoNuSeg dataset	a. 90.86 b. 80.25	Multi-level feature fusion
PHTrans, Liu et al. [96]	a. ACT-MOS b. Cardiac segmentation	a. BCV dataset b. ACDC	a. 88.55 b. 91.79	Multi-level feature fusion
nnFormer, Zhou et al. [97]	a. ACT-MOS b. Cardiac segmentation	a. Synapse multi-organ CT b. ACDC	a. 87.40 b. 91.78	
PMTrans, Zhang et al. [24]	a. Gland segmentation b. Nuclear segmentation	a. GlaS dataset [98] b. MoNuSeg dataset [99]	a. 81.48 b. 80.09	Multi-resolution images
MedT,Valanarasu et al. [100]	a. Brain anatomy segmentation b. Gland segmentation c. Nucleus segmentation	a. Brain US dataset [101] b. GlaS dataset c. MoNuSeg dataset	-	
Global + local CoTr, Xie et al. [25]	ACT-MOS	BCV dataset [102]	85.0	Multi-scale features
MCTrans, Ji et al. [103]	a. Cell segmentation b. Polyp segmentation c. Skin lesion segmentation	a. Pannuke [104] b. KCCEE c. ISIC2018 [105]	a. 68.40 b. 92.30, 86.58, 83.69, 86.20 c. 90.35	Multi-scale features
D-Former, Wu et al. [106]	a. ACT-MOS b. Cardiac segmentation	a. Synapse multi-organ CT b. ACDC	a. 88.83 b. 92.29	Dilated Transformer
TF-Unet, Fu et al. [86]	a. ACT-MOS b. Cardiac segmentation	a. Synapse multi-organ CT b. ACDC	a. 85.46 b. 91.72	Multi-scale features
UNETR, Hatamizadeh et al. [107]	a. Brain tumour segmentation b. Spleen CT segmentation	a. MSD-01 b. MSD dataset Task09	a. 71.81 b. 95.82	Multi-resolution
Swin UNETR, Hatamizadeh et al. [108]	Brain tumor segmentation	BraTS2021 dataset	91.3	Swin-Transformer
TransAttUnet, Chen et al. [109]	a. Skin lesion segmentation b. Lung field segmentation c. COPLE d. Nucleus segmentation e. Gland Segmentation	a. ISIC2018 dataset b. JSRT, Montgomery and NIH [110] c. Clean-CC-CCII dataset [111] d. Bowl dataset [112] e. GlaS dataset	a. - b. 98.88 c. 86.57 d. 91.62 e. 89.11	Feature fusion
MT-Unet, Wang et al. [113]	a. ACT-MOS b. Cardiac segmentation	a. Synapse multi-organ CT b. ACDC	a. 78.59 b. 90.43	
AFTer-Unet, Yan et al. [114]	a. ACT-MOS b. Thoracic segmentation	a. BCV dataset b. Thorax-85 [115], SegTHOR [116]	a. 81.02 b. 92.32, 92.10	
Axial fusion S^2WinTOUnet, Zhang et al. [117]	Skin lesion segmentation	ISIC2018 dataset	90.4	Star-shaped Window Self-attention
Karimi et al. [118]	a. Brain cortical plate segmentation b. Pancreas segmentation c. Hippocampus segmentation	-	a. 87.9 b. 82.6 c. 88.1	
Swin-Unet, Cao et al. [119]	a. ACT-MOS b. Cardiac segmentation	a. Synapse multi-organ CT b. ACDC	a. 79.13 b. 90.00	Swin-Transformer
DS-TransUNet, Lin et al. [120]	a. Polyp segmentation b. Skin lesion segmentation c. Gland segmentation d. Nucleus segmentation	a. Kvasir, CVC-ColonDB, EndoScene, ETIS, CVC-ClinicDB b. ISIC2018 c. GlaS Dataset d. Bowl dataset	a. 93.5, 79.8, 91.1, 77.2, 93.8 b. - c. 87.19 d. -	Swin-Transformer
MISSFormer, Huang et al. [121]	a. ACT-MOS b. Cardiac segmentation	a. Synapse multi-organ CT b. ACDC	a. 81.96 b. 87.90	Multi-scale feature

ACT-MOS: abdominal CT multi-organ segmentation; COPLE: COVID-19 pneumonia lesion segmentation; ACDC: automated cardiac diagnosis challenge; MSD-01: Medical Segmentation Decathlon dataset Task01; KCCEE: Kvasir [122], CVC-ClinicDB [123], CVC-ColonDB [124], EndoScene [125], and ETIS [126].

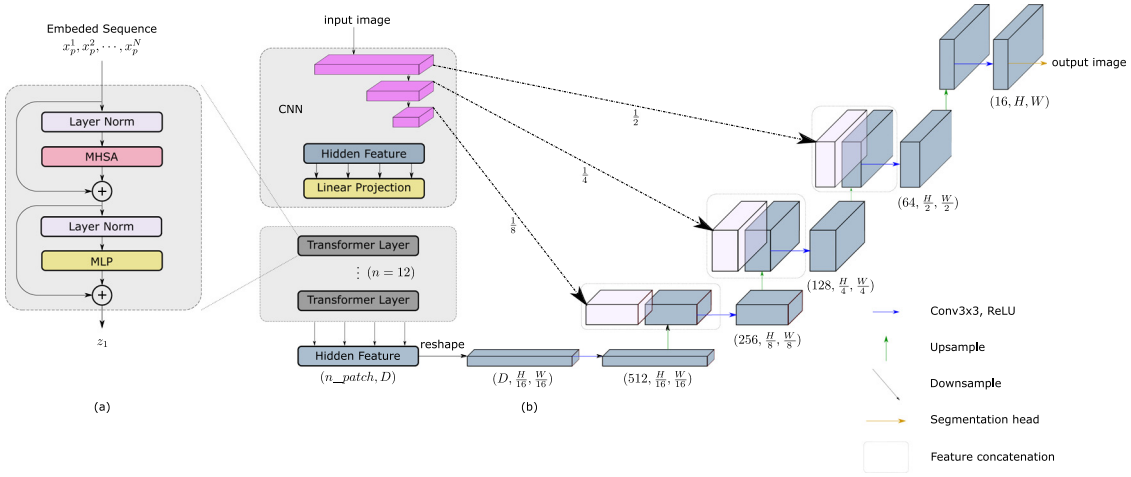

Figure 7. Overview of TransUNet for medical image segmentation. (a) Schematic design of the Transformer layer; (b) Architecture of TransUNet. [78] Transformer layers are inserted into the encoder of the Unet.

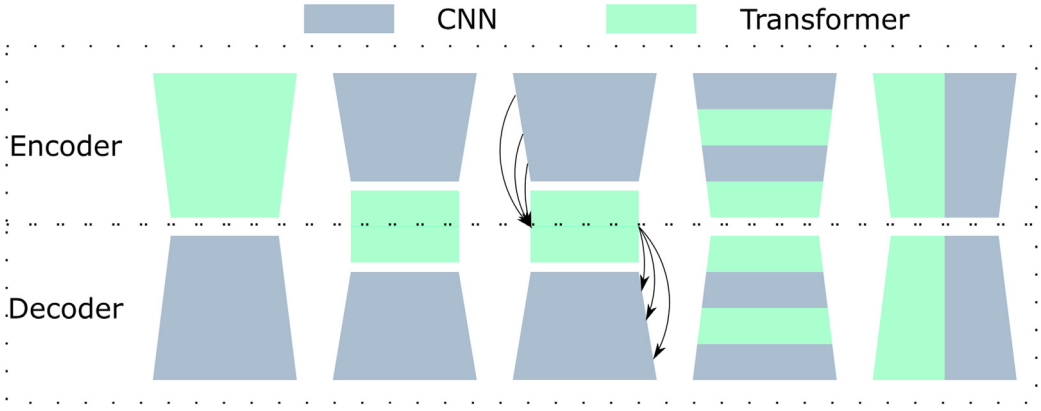

Figure 8. Comparison of several hybrid architectures of Transformer and CNN [96].

representations, without investigating how to optimally combine self-attention with convolution. To address this issue, they introduced the nnFormer, which has an interleaved architecture based on empirical combination of self-attention and convolution. Xu et al. [94] proposed the ECT-NAS method to search for efficient CNN-transformer architectures for medical image segmentation based on a multi-scale space search.

3.2.1.3. Multi-scaling

The multi-scale strategy for transformers in MIA uses features in a multi-scale manner or takes multi-scale images as inputs.

(1) Multi-resolution images. Zhang et al. [24] proposed a pyramidal network architecture, namely pyramid medical transformer (PMTrans), which captures multi-range relations by working on multi-resolution images. Valanarasu et al. [100] added gated axial transformer layers in the encoder, which contains the basic building block of both height- and width-gated multi-head attention blocks. The whole image and patches were used to learn global and local features, respectively, and a local-global training strategy was proposed to further boost the overall performance.

(2) Multi-scale features. In contrast to TransUNet, which only uses a transformer to process the low-resolution feature maps learned from the previous layer, Xie et al. [25] proposed a deformable transformer to process multi-scale and high-resolution feature maps. Ji et al. [103] proposed a multi-compound transformer (MCTrans), which embeds multi-scale convolutional features as a sequence of tokens and performs intra- and inter-scale self-attention. In contrast to models that use CNNs to extract features, Hatamizadeh et al. [107] introduced Unet transformers (UNETR), which use a pure transformer as an encoder to learn sequence representations of the input volume. The transformer encoder is directly connected to a decoder via skip connections at different resolutions to compute the final semantic segmentation output. Zhang et al. [117] proposed S^2WinTOUnet, which uses a star-shaped window self-attention to obtain fine-grained details and coarse-grained semantic information.

(3) Multi-level attention. Chen et al. [109] proposed TransAttUnet, in which a multi-level guided attention and multi-scale skip connection are jointly designed to effectively enhance traditional U-shaped architectures. Both transformer self attention and global spatial attention are incorporated into TransAttUnet to effectively learn non-local interactions between encoded features. Wang et al. [113] proposed the mixed transformer module, which calculates self-affinities through well-designed local-global Gaussian-weighted self-attention and then mines interconnections between data samples through external attention. Wu et al. [106] proposed the dilated transformer, which conducts self-attention for pairwise patch relations that are captured alternately in local and global scopes.

(4) Multi-axial fusion. Yan et al. [114] applied an axial fusion transformer to fuse inter-slice and intra-slice information, which reduced the computational complexity of calculating self-attention in 3D space.

To conclude, the aforementioned methods all leverage additional features learned using a feature fusion strategy for more effective learning.

Table 3 Transformers for image-to-image translation tasks in medical images

References	Application	DataSet	Metrics	Task
GIT, Watanabe et al. [26]	Parkinson	the Parkinson's Progression Marker Initiative database [146]	-	Image synthesis (SPECT)
VTGAN, kamran et al. [147]	Retinopathy	Fundus & FA [148]	FID, KID	Image synthesis (Fundus → FA)
GANBERT, Shin et al. [149]	Alzheimer	ADNI[1]	PSNR, SSIM, RMSE	Image synthesis (MRI → PET)
Hu et al. [150]	Brain	IXI[2]	PSNR, SSIM	Image synthesis (MRI → T1/T2)
SLATER, Korkmaz et al. [151]	Brain	IXI[3]; fastMRI [152]	PSNR, SSIM	Zero-shot MRI Reconstruction
CyTran, Ristea et al. [153]	Lung	Coltea-Lung-CT-100W [153]	MAE, SSIM, RMSE	CT Translation (Non-Contrast → Contrast)
ResVit, Dalmaz et al. [154]	Brain; Pelvic	IXI; BRATS [155–157]; pelvic MRI-CT database [158]	PSNR, SSIM	Multi-model Image synthesis
T^2Net, Feng et al. [159]	Brain	IXI; Clinical dataset	PSNR, SSIM, NMSE	MRI Reconstruction & Super-resolution
PTNet, Zhang et al. [160]	Infant brain	dHCP [161]	PSNR, SSIM	MRI synthesis & Super-resolution
TED-net, Wang et al. [162]	Liver lesions	2016 NIH-AAPMMayo Clinic LDCT Grand Challenge dataset [163]	SSIM, RMSE	Low-dose CT Denoising
Eformer, Luthra et al. [164]	Liver lesions	2016 NIH-AAPMMayo Clinic LDCT Grand Challenge dataset [163]	PSNR, SSIM, RMSE	Low-dose CT Denoising

FID: frechet inception distance; KID: kernel inception distance; PSNR: peak signal-to-noise ratio; SSIM: structural similarity index; RMSE: root mean square error; MAE: mean absolute error; NMSE: normalized mean square error.

[1] http://adni.loni.usc.edu/.

[2] http://brain-development.org/ixidataset/.

[3] http://brain-development.org/ixidataset/.

3.2.2. Pure transformers

In addition to the aforementioned variants of the Unet architecture that combine a transformer with convolutions, Karimi et al. [118] used simple self-attention between adjacent image patches without convolution operations. A 3D image is divided into n^3 3D patches ($n = 3$ or 5), and a 1D embedding is learned for each patch. Through the self-attention between patch embeddings, the network outputs the segmentation result of the center patch. Methods using this assumption can be easily recognized as pure transformers.

Cao et al. [119] developed an Unet-like pure transformer for medical image segmentation by feeding tokenized image patches into the a transformer-like U-shaped encoder-decoder architecture with skip connections for semantic feature learning in a local-global manner. Lin et al. [120] went a step further and proposed DS-TransUNet, which first adopts dual-scale encoder subnetworks based on Swin-Transformer to extract coarse- and fine-grained feature representations on different semantic scales. A well-designed transformer interactive fusion module was also proposed to effectively establish global dependencies between features of different scales through the self-attention mechanism. To better leverage the natural multi-scale feature hierarchies of transformers, Huang et al. [121] proposed MISSFormer, which has two appealing design features: (1) an enhanced transformer block as a feed-forward network with better feature consistency, long-range dependencies, and local context; and (2) an enhanced transformer context bridge to model long-range dependencies and local context of multi-scale features generated by the hierarchical transformer encoder.

3.3. Image-to-image translation

Transformer models also have been shown to have strong learning ability in many image-to-image translation applications including image synthesis [16], reconstruction [171], and super-resolution [172]. However, in the field of medical image analysis, studies (*e.g.*, [26,147]) on image-to-image translation have recently started to emerge. We list existing transformer-based image-to-image translation methods in Table 3, as well as the corresponding evaluation metrics.

3.3.1. Image synthesis

In the medical field, image synthesis remains very challenging owing to inter-subject variability and the fact that anatomical hallucinations (e.g., hallucinating a white spot in a brain MRI) might be detrimental to diagnostic tasks. In recent years, generative adversarial learning has been widely used to tackle image synthesis tasks. Therefore, transformers have been combined with a generative adversarial learning paradigm for image synthesis. For example, Hu et al. [150] introduced a double-scale discriminator GAN for cross-modal medical image synthesis, consisting of a transformer-based global discriminator and a CNN-based local discriminator. Watanabe et al. [26] proposed a generative model architecture based on a transformer decoder block, owing to its powerful ability in modeling time series. During data processing, they normalized the pixel values of single photon emission CT (SPECT) images by the specific/nonspecific binding ratio. During the training process, they used a transformer decoder to construct an auto-regression model and trained the model on [^{123}I]FP-CIT SPECT images from the Parkinson's Progressive Marker Initiative database in an unpaired manner. The trained model could generate SPECT images that had characteristics of Parkinson's disease patients. Kamran et al. [147] proposed a transformer-based conditional GAN, shown in Figure 9, that could simultaneously perform semi-supervised image synthesis from fundus photographs to fluorescein angiography (FA) for diagnosis of retinal disease.

To tackle the problem of the intensity range of positron emission tomography (PET) often being wide and dense and even heavily biased toward zero, Shin et al. [149] built a GAN utilizing BERT, namely GANBERT, to generate PET images from MRI images. Luo et al. [173] proposed a 3D transformer GAN to reconstruct high-quality PET image at a low dose. In order to overcome the limitation of scarce access to large medical datasets, Korkmaz et al. [151] introduced an unsupervised reconstruction method based on zero-Shot Learned Adversarial TransformERs (SLATER) to perform MRI synthesis. SLATER is an unconditional adversarial architecture consisting of a synthesizer, a discriminator, and a mapper. The synthesizer uses cross-attention transformer blocks to capture long-range relationships, and the mapper maps noise and latent variables onto MR images. Ristea et al. [153] proposed an architecture named CyTran, which is based on generative adversarial convolutional transformers and integrates the cycle-consistency loss for translation of unpaired CT images between contrast and non-contrast CT scans.

In addition to their applications to image synthesis between two modalities, transformer models have been used successfully in multi-modal medical image synthesis. For example, Dalmaz et al. [154] proposed a generative adversarial approach, ResViT, for multi-modal medical image synthesis. The generator in ResViT is based on encoder-decoder architecture, with a central bottleneck that comprises aggregated residual transformer blocks capable of synergistically preserving local and global contexts.

3.3.2. Image super-resolution

Super-resolution imaging comprises a class of techniques that enhance the resolution of an imaging system. It is also a popular sub-

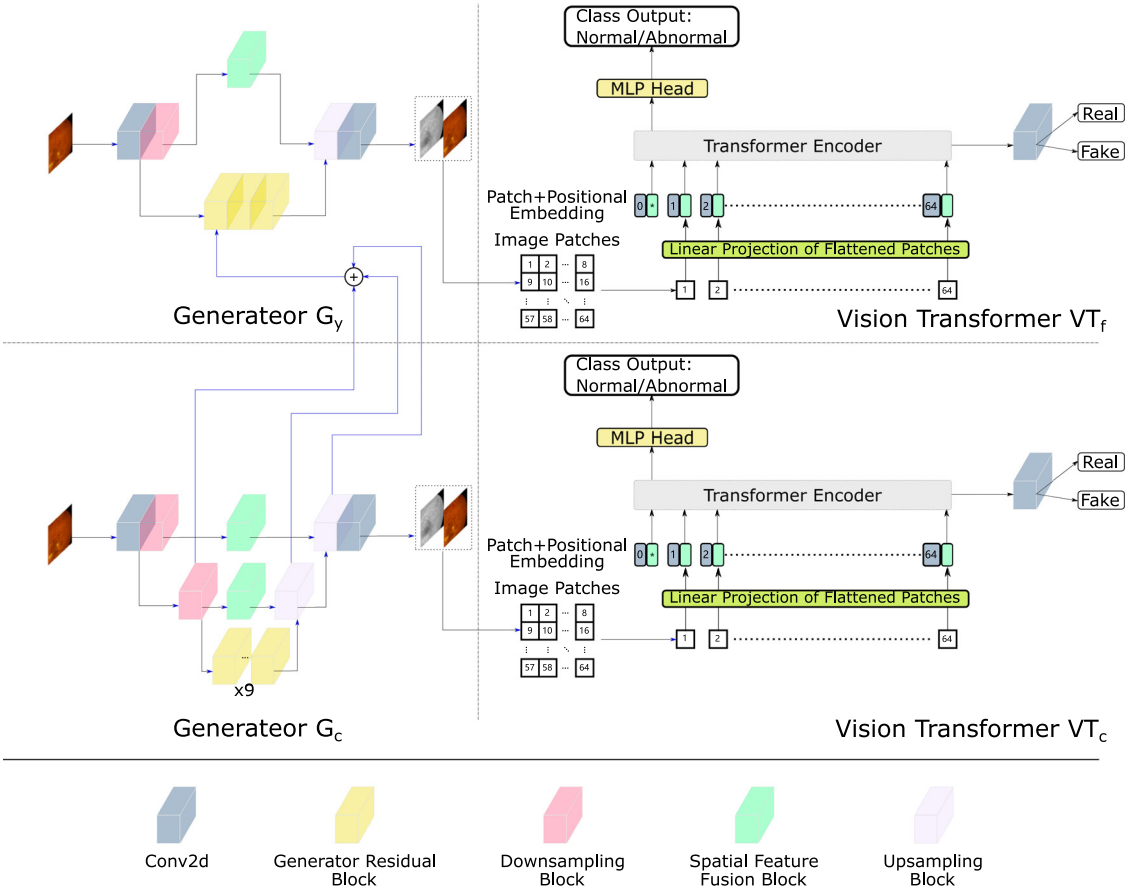

Figure 9. Overview of the architecture of VTGAN, which uses coarse and fine generators G_f and G_c, and ViTs VT_f, VT_c as discriminators [147].

field of image synthesis. Outstanding contributions have been made by transformer models on super-resolution tasks in medical image analysis. For instance, Feng et al. [159] introduced a task transformer network (T^2Net) to jointly learn image reconstruction and super-resolution tasks in MRI. This multi-task framework included a super-resolution branch and a common resolution branch, and the authors designed the transformer module to embed the similarity and align the gap between the two branches. Zhang et al. [160] proposed a high-resolution synthesizer based on pyramid transformer (PTNet) and used it for MRI synthesis of images of infant brains. PTNet consists of a performer encoder, a performer decoder, and a transformer bottleneck that inherits U-structures as well as multi-resolution pyramid structures.

3.3.3. *Image denoising*

Image denoising is the task of removing noise from an image. It is a fundamental step in several clinical applications. For example, Wang et al. [162] used a transformer for low-dose CT (LDCT) denoising for the first time. They developed an encoder-decoder dilation network based on token-to-token (T2T) ViT, namely TED-net. TED-net is a U-structure model that uses the dilation in the T2T stage to enlarge the receptive field. Luthra et al. [164] proposed an edge-enhancement-based transformer (Eformer) that uses transformer blocks to construct an encoder-decoder architecture for medical image denoising. Transformer models and their applications in the task of LDCT denoising remain scarce.

3.4. *Detection*

The meaning and terminology of 'detection' varies across technical and clinical fields. In technical areas, it often refers to checking for the existence of diseases or lesions, whereas in clinical practice it often means diagnosis or disease classification, as discussed above. In computer vision, detection aims to identify the location of objects in an input image and predict their categories/classes. In this section, detection refers to object detection.

Transformers dealing with detection tasks using medical images are often combined with CNN blocks, where a CNN is used to extract features from medical images, and the transformer architecture is used to enhance the extracted features for downstream detection. Shen et al. [166] proposed a DETR-based model, namely COTR, for the detection of polyps in the colon. DETR [14] is a primer method for object detection in computer vision. COTR is composed of a CNN for feature extraction, transformer encoder layers interleaved with convolutional layers for feature encoding and recalibration, transformer decoder layers for object querying, and a feed-forward network for detection prediction. They inserted convolutional layers into the transformer encoder for high-level image feature reconstruction and convergence acceleration. Ma et al. [167] proposed a TR-Net that combines CNN and transformer nets to detect significant stenosis in multiplanar reformatted images. Their model employs a shallow 3D-CNN to extract local semantic features of coronary regions while ensuring the model's efficiency. Next, transformer encoders are used to learn correlations between different regions of the local stenosis at each position of a coronary artery. Thus, TR-Net can accurately detect stenosis after aggregating information from local semantic features and global semantic features. Jiang et al. [165] constructed a YOLOv5s-based transformer for the detection of caries, called RDFNet. The model uses the FReLU activation function to activate complex visual-spatial information of images for efficiency boosting. Kong et al. [168] proposed CT-CAD, a context-aware hybrid

Table 4 Transformers for detection tasks

References	Disease	Organ	Dataset	Highlight
RDFNet [165]	Caries	Tooth	-	Transformer for feature extraction
COTR [166]	Polyp Lesion	Colon& rectum	ETIS-LARIB CVC-ColonDB	Convolutions × Transformer
TR-Net [167]	Coronary arteries significant stenosis	Coronary arteries	-	
CT-CAD [168]	Chest abnormality detection	Chest	Vinbig chest Chest Det 10	Context-aware feature extractor
Tao et al. [169]	Vertebrae detection	Spine	VerSe 2019 challenge MICCAI-CSI 2014 challenge	Inscribed sphere-based object detector

transformer for end-to-end detection of chest abnormalities on on X-ray images. Tao et al. [169] designed a spine-transformer to address automatic detection and localization of vertebrae in arbitrary field-of-view spine CT (Table 4). They formulated the detection as an one-to-one set prediction problem.

3.5. Registration

Transformers have several advantages in image registration tasks owing to their self-attention mechanism, which enables precise spatial mapping between moving and fixed images. Chen et al.pioneered the use of transformers for image registration. Inspired by the architecture of TransUnet [78], they proposed ViT-V-Net [29], which combines ViT and V-Net by simply altering the network architecture of VoxelMorph (a conventional registration network) [174]. ViT-V-Net produced superior performance against benchmark methods. In an extension of their work, they developed TransMorph [175] for volumetric medical image registration. In this method, the Swin-Transformer [31] was used as the encoder network to capture the spatial correspondence between input moving and fixed images, and a ConvNet decoder was used to map the information provided by the transformer encoder onto a dense displacement field. Long skip connections were deployed to maintain the flow of local information between the encoder and decoder stages. Transformers-based registration methods remain rare and need further exploration and research in the future.

3.6. Video-based applications

Because of a limited receptive field, CNNs cannot fully utilize the global temporal and spatial information in continuous video frames; however, transformers can overcome this defect. Ji et al. [176] proposed PNS-Net (Progressively Normalized Self-attention Network) for accurate polyp segmentation from colonoscopy videos. Kondo et al. [177] proposed LapFormer to detect surgical tools in laparoscopic surgery videos. Czempiel et al. [178] introduced OperA to predict surgical phases from long video sequences. Reynaud et al. [179] adopted a transformer architecture, which contained a residual autoencoder Network and a BERT model, to analyze videos of arbitrary length. Long et al. [180] applied transformers to estimate surgical scene depth.

4. Discussion

Transformers have been successfully applied to many applications in almost all fields of medical image analysis. However, the deployment of machine learning methods in real clinical applications can lead to poor performance owing to several challenges. Among them, the most urgent challenge is label scarcity, especially in scene-understanding tasks, e.g., segmentation and detection, which usually need pixel-wise precise labeling. Learning from noisy labels presents a bigger challenge. In addition, building advanced computer aided diagnosis (CADx) methods requires the use of multi-modality clinical data in a multi-task manner – a versatile learning approach that is difficult in design.

4.1. Transformers under different learning scenarios

4.1.1. Multi-task learning

Building models with multiple tasks helps to improve their generalizability, for which there is high demand in the field of medical image analysis. A frequently used framework unifies classification and segmentation in one model [187-188]. For instance, Chen et al. [188] proposed Multi-Task TransUNet (MT-TransUNet) to jointly learn segmentation and classification of skin lesions. With local details (e.g., skin color, texture) and long-range context (e.g., skin lesion shape, physical size) extracted by CNNs and ViTs, the method achieved SOTA performance and efficiency improvements in model parameters and inference speed. Sui et al. [189] combined detection with segmentation tasks to develop a novel transfer learning method, CST, with a transformer-based framework for joint CRC region detection and tumor segmentation. For detection, the generated region proposals of the input images, as well as the position features obtained by the encoder-decoder module, were used as the input to a DETR network. For segmentation, the model used image patches as inputs, which were projected into a sequence of embeddings.

4.1.2. Multi-modal learning

Using multiple modality data provides complementary evidence for diagnosis. For example, researchers have explored the use of combinations of OCT and visual field (VF) testing to aid in the diagnosis of eye diseases. Song et al. [71] used transformers for glaucoma diagnosis. Their model used an attention mechanism to model the pairwise relations between OCT features and VF features. Next, the attention mechanism was applied again to calculate the regional relations of features between the VF areas and the quadrants of the retinal nerve fiber layer. The complementary information was passed from one modality to another by a transformer model.

Monajatipoor et al. [184] developed a transformer-based vision-and-language model that combined the efficient PixelHop++ model with the BERT model. Specifically, the BERT model was pretrained using in-domain knowledge. The model was proved to be effective when trained on small-scale datasets. The extracted vision features and the word embeddings were fed into the transformer for final diagnosis. Although the model decreased the need for massive annotations of medical images, the pretraining of the language model still needed a large amount of clinical report data. Jacenków et al. [186] combined text with CXR for disease classification. They observed that the interpretation and reporting of an image was affected by the scan request text, which served as the indication field in the radiology report. Zheng et al. [182] focused on feature fusion of multi-modal information, considering the latent inter-modal correlation. They proposed a transformer-like modal-attentional feature fusion approach (MaFF) to extract rich information from each modality while mining the inter-modal relationships. Next, an adaptive graph learning mechanism was utilized to construct latent robust graphs for downstream tasks based on the fused features. The method achieved significant improvements in the prediction of AD and autism. Dai et al. [185] proposed TransMed for the diagnosis of parotid gland tumor. TransMed combines the advantages of CNN and transformer networks to capture both low-level textures and cross-modality high-level relationships. The model first processes multi-modal images as sequences by chaining and sending them to a CNN for feature extraction. The feature sequences are then fed into the transformers to

Table 5 Transformers for multi-modal learning

References	Disease	Organ	Dataset	Highlight	Task
CLIMAT [181]	Alzheimer's Disease	Brain	ADNI	Use multi-trans to mimic radiologist and general practitioner interactions	Prognosis
Zheng et al. [182]	Alzheimer's disease	Brain	TADPOLE	Proposes a modal-attentional multi-modal fusion	Prediction
Qiu et al. [183]	Alzheimer's disease	Brain	ADNI	Build a graph based on rich multi-modal features; proposes graph trans to classify	Classification
	Autism spectrum	Brain	ABIDE		Prediction
BERTHop [184]	Thoracic Disease	Chest	OpenI	Incorporates PixelHop+ into a trans-based model; adopts in-domain pretrained BERT	Diagnosis
DRT [71]	Glaucoma	Eye	ZOC-OCT&VF	Use two relation module to extract inter-modal relation; use trans to fuse features	Diagnosis
TransMed [185]	Parotid gland tumor	Parotid gland	*	Use trans to capture cross-modality mutual information and fuse features	Prediction
Jacenków [186]	Thoracic diseases	thorax	MIMIC-CXR	Use text to assist image classification	Classification

Table 6 Transformers for weakly-supervised learning

References	Disease	Organ	Dataset	Highlight	Task
Weakly Supervised					
Li et al. [192]	Diabetic Retinopathy	Eye	Messidor	Induced Self-Attention to model relation of instances within a bag	Classification
Rymarczyk et al. [193]	Diabetic Retinopathy	Eye	Messidor	self-attention with Attention-based MIL Pooling	Classification
Yang et al. [194]	Multiple Nodule Malignancy	Lung	LIDC-IDRI	inter-solitary-nodule relationships	Classification
	Lung Nodule	Lung	LUNA16, Tianchi Val		Detection
MIL-VT [195]	Diabetic Retinopathy	Eye	APTOS2019	MIL-head to provide complementary information of patches to the class token	Classification
	Retinal Fundus Disease	Eye	RFMiD2020		Diagnosis
TransMIL [196]	Breast Cancer Metastasis	Breast	CAMELYON16	explore both morphological and spatial information between different instances	Detection
	Lung Cancer	Lung	TCGA-NSCLC		Classification
	Kidney Cancer	Kidney	TCGA-RCC		Classification
Self-Supervised					
Park et al. [43]	COVID-19	Lung	*	Pretrain model on large scale data; evaluate necessity of self-supervised pretraining	Diagnosis
Jun et al. [129]	Alzheimer's disease	Brain	ADNI	Pretrain trans with masked encoding vector prediction as SSL proxy task	Diagnosis
	Brain Age	Brain	ADNI		Prediction
TransPath [197]	CRC	Colorectal	NCT-CRC-HE	Collect approximately 2.7 million images for self-supervised pretraining	Classification
	Breast Cancer	Breast	PatchCamelyon		Classification
	Colorectal polyps	Colorectal	MHIST		Classification
Truong et al. [198]	Axillary lymph node cancer	Lymph	PatchCam	Validate ViT-based self-supervised method by comparison	Classification
	Diabetic retinopathy	Eye	APTOS		Classification
	Pneumonia	Chest	Pneumonia chest X-ray		Classification
	Thorax Disease	Chest	NIH chest X-ray		Classification
Sriram [199]	COVID-19	Lung	NYU COVID	Adopt momentum contrastive learning for SSL pretraining	Prognosis
Chen [199]	Tissue phenotyping	Tissue	TCGABRCA cohort CRC-100K BreastPathQ	DINO-based knowledge distillation applied on ViT	Classification

learn the relationships between sequences as well as conducting feature fusion. Their work leveraged transformers to capture mutual information from images of different modalities, resulting in better performance and efficiency. Nguyen et al. [181] attempted to mimic the interaction between a radiologist and a general practitioner in the diagnosis of knee osteoarthritis and prediction of prognosis. They proposed a clinically-inspired multi-agent transformers (CLIMAT) framework with a tri-transformer architecture (Table 5). In this framework, first, a feature extractor with a combination of transformer and CNN is used to predict the current state of a disease. Next, the non-image auxiliary information is fed into another transformer to extract context embedding. Finally, an additional transformer-based general practitioner module forecasts disease trajectory based on the current state and context embedding.

To conclude, transformers are regarded as a promising approach to bridge CV and NLP tasks [190]. Under this assumption, Radford et al. [191] built a multi-modal transformer, CLIP, that provided zero-shot ability for recognizing images from text descriptions without image labeling. These strength of this approach also indicates a potential way of building more robust and accurate CADx methods for real clinical applications, where multiple data types, *e.g.*, clinical, laboratory, and imaging data, can be used as diverse source of information.

4.1.3. Weakly supervised learning

One of the weakly supervised conditions in medical images is that the ROI for a certain disease is relatively small in the image, and only image-level labeling is available. Multiple instance learning (MIL) was adopted as a solution to this problem. In MIL, the training samples include sets of instances, called bags. The supervision is provided only for bags, and individual labels of the instances contained in the bags are not provided [200].

Although many existing MIL methods assume that positive and negative instances are sampled independently from a positive and a nega-

tive distribution [200], instances in a bag are relational, especially in medical image analysis. The learning scenario of MIL does not follow the independent and identically distributed assumption, as the relationships between instances are not neglected. In such situations, ViTs can be leveraged to build correlations between instances to achieve better high-level representations. Li et al. [192] proposed a transformer-based MIL framework with an induced attention block, which calculates the attention while bypassing the quadratic computational complexity caused by the pairwise dot product. The feature aggregator of the framework is also based on multi-head attentions. It merges the previously mentioned features into bag representations. Yang et al. [194] treated multiple pulmonary nodules of a patient as a bag and each nodule as an instance. Unlike conventional MIL methods that use a pooling operation to get bag-level representations, they used a 3D DenseNet to learn solitary-nodule-level representations at the voxel level. Next, the generated representations were fed into the transformer to learn the nodule relationships from the same patient. To reduce the computational burden, they applied in-group scaled-dot-production attention, extracted from split channel features. Shao et al. [196] focused on the correlations between different instances as opposed to simply assuming that instances are independent and identically distributed. To this end, they proposed a transformer-based MIL framework to deal with the whole-slide image classification problem. Their framework used transformer layers to aggregate morphological information and a pyramid position encoding generator to extract spatial information. They also used the Nystrom method to calculate approximated self-attentions, which reduced computational complexity from $O(n^2)$ to $O(n)$. Rymarczyk et al. [193] focused on the attention mechanism and revised attention-based MIL pooling (AbMILP), which aggregates information from a varying number of instances. They developed self-attention AbMILP (SA-AbMILP) to model the dependencies between different instances within a bag. They also extended the calculation of attentions by introducing different kernels, which played the same part as the dot product. They evaluated their work on histological, microbiological, and retinal datasets. Yu et al. [195] explored the applicability of ViTs to retinal disease classification in fundus images (Table 6). They developed a MIL-enhanced ViT (MIL-VT) by adding a plug-and-play MIL learning head to the ViT to exploit the features extracted from individual patches.

Another weakly supervised example is semi-supervised learning, which requires only a small amount of labeled data to exploit knowledge from a large amount of unlabeled data. Luo et al. [201] first combined a CNN and transformer for semi-supervised medical image segmentation. They introduced cross-teaching between the CNN and the transformer, with the prediction of each network used as a pseudo label to supervise the other network. Zhao et al. [202] proposed a context-aware network called CA-Net for semi-supervised LA segmentation from 3D MRI. CA-Net contains two main modules, a trans-V module that combines a transformer and V-net to learn contextual information, and a discriminator to calculate an adversarial loss for learning the unlabeled data. Xiao et al. [203] used a dual teacher structure involving a CNN and a transformer to guide a student segmentation model.

4.1.4. Self-supervised learning (SSL)

Successful training of a transformer model relies on *large-scale* annotated data, which are rarely available in real clinical facilities. The SSL paradigm was created to handle such issue. SSL aims to improve the performance of downstream tasks (e.g., classification, detection, and segmentation) by transferring knowledge from a related unsupervised upstream task (*i.e.*, learning of vision concepts), and pretrains the model using self-contained information in the unlabeled data [204]. In practice, training of SLL ViTs generally involves pretraining the model on ImageNet, followed by a fine-tuning step on the target medical image dataset. This can boost the performance of ViTs in comparison with CNNs and enable SOTA accuracy to be achieved [205–208].

Truong et al. [198] evaluated the transferability of self-supervised features in medical images. They pretrained features using DINO, a self-supervised ViT. They used the ViT as a backbone and demonstrated that it could outperform SimCLR and SwAV. Park et al. [43] used a public large-scale CXR classification dataset to pretrain the backbone network. The features extracted by the pretrained backbone model were then fed into a ViT to diagnose COVID-19. Jun et al. [129] proposed a self-supervised transfer learning framework that could better represent the spatial relationships in 3D volumetric images to facilitate downstream tasks. They converted 3D volumetric images into sequences of 2D image slices from three views and fed them into the pretrained backbone network, which consisted of a convolutional encoder and a transformer. The pretraining of the transformer was implemented using masked encoding vectors, which served as a proxy task for SSL. The downstream tasks included brain disease diagnosis, brain age prediction, and brain tumor segmentation, using 3D volumetric images. They also explored a parameter-efficient transfer learning framework for 3D medical images. Wang et al. [197] collected a large public histopathological image dataset to pretrain their proposed hybrid CNN-transformer framework. Moreover, they designed a token-aggregating and excitation module to further enhance global weight attention by taking all tokens into consideration. Sriram et al. [199] explored the applications of transformers for COVID-19 prognosis. They proposed a multiple image prediction model that could take a sequence of images along with the corresponding scanning time as input. To deal with missing COVID-19 images, they used momentum contrast learning, a self-supervised method, to pretrain the feature extractor network. In addition to the features extracted from X-rays, they used continuous positional embedding to add information based on the time-step. The concatenation of features and continuous positional embeddings was fed into the transformer to predict the possibility of an adverse event. Chen et al. [209] showed that ViT using DINO-based knowledge distillation could learn data-efficient and interpretable features in histology images by training various self-supervised models with validation on different weakly supervised tissue phenotyping tasks. Notably, they achieved excellent performance on different attention heads in the ViT while learning distinct morphological phenotypes.

4.2. Model improvement: quantification, acceleration, and interpretation

Several studies have focused on model efficiency within the medical imaging field. A natural idea is to simplify the attention mechanism, which demands the largest workload in transformers. Gao et al. [84] proposed an efficient self-attention mechanism and position encoding, which significantly reduced the complexity of the self-attention operation from $O(n^2)$ to approximately $O(n)$. This circumvented the hurdle of transformers requiring huge amounts of data to learn vision inductive bias. The hybrid-layer design could initialize transformers as convolutional networks without the need for pretraining. The aforementioned VOLO proposed by Liu et al. [48] replaced the standard ViTs with Linformer, which performs an internal self-attention mechanism, reducing the original space time complexity of $O(n^2)$ to a smaller complexity of $O(n)$. Li et al. [210] redesigned the transformer block in their TransBTSV2 model, resulting in a shallower but wider architecture compared with conventional transformer-based methods. Inspired by dilated convolution kernels, Wu et al. [106] conducted global self-attention in a dilated manner, enlarging the receptive fields without increasing the patches and thus reducing computational costs. Xu et al. [94] built a multi-scale searching space composed of a multi-branch parallel searching block, which connected a CNN and transformer in parallel. They also proposed an efficient resource-constrained search strategy to simultaneously optimize accuracy and costs (e.g., params. and FLOPs) of the model.

There have been fewer studies attempting to solve the model efficiency problem in MIA rather than in CV. However, as medical images generally come in large sizes and small quantities, there is an urgent

need to solve this problem. Thus, we would like to see more work in this specific research direction.

4.3. Comparison with CNNs

CNNs were dominant in CV prior to the emergence of ViTs, including in the field of medical image analysis. Much effort has been invested in improving the performance of CNN-based classifiers for both natural and medical images. Several studies have investigated whether CNN-based methods could work on ViTs. Moreover, as ViTs have ranked top among several benchmarks, many studies have focused on performance comparisons between ViTs and CNNs.

Large-scale datasets are required for to obtain desirable performance with transformers. However, in the medical image analysis field, available images and annotations are limited. To alleviate this problem, many methods have adopted convolutional layers in ViTs to boost performance with limited medical images and have also leveraged the power of transfer learning and SSL. Matsoukas et al. [205] explored whether transfer learning and SSL regimes could benefit ViTs. They conducted several experiments to compare the performance of a CNN (*i.e.*, ResNet50) and a ViT (*i.e.*, DEIT-S) using different initialization strategies: (1) randomly initialized weights, (2) transfer learning using ImageNet pretrained weights, and (3) self-supervised pretraining on the target dataset with the same initialization as in (2). They evaluated these methods on the APTOS 2019, ISIC 2019, and CBIS-DDSM datasets. It can be concluded that standard procedures, e.g., initialization using ImageNet pretrained weights and leveraging SSL, can bridge the performance gap between CNN and ViT. Krishnamurthy et al. [211] adopted a transfer learning scheme in both CNNs and ViTs for Pneumonitis diagnosis. They first pretrained their models on ImageNet and fine-tuned the classifier on their private dataset. However, their comparison was based on fine-tuning with frozen backbone layers, which limited the performance of feature extraction when adapted to the target domain. Truong et al. [198] assessed the transferability of self-supervised features in medical imaging tasks. They chose ResNet-50 as the backbone and pretrained it using three self-supervised methods: SimCLR, SwAV, and DINO. DINO used ViT as the backbone and consistently outperformed other self-supervised techniques as well as the supervised baseline by a large margin. They proposed a model-agnostic technique, *i.e.*, dynamic visual meta-embeddings, to combine pretrained features from multiple SSL methods with self-attention.

For the task of multi-scale cell image classification, Liu et al. [212] developed an experimental platform to compare multiple deep learning methods, including CNNs and ViTs. They validated the performance of deep learning models on standard and scaled data by changing the cell aspect ratios of the images. The results suggested that deep learning models, including ViTs, are robust to changes in the cell aspect ratio in cervical cytopathological images. For shoulder implant X-ray manufacturer classification, Zhou et al. [213] compared the performance of various models, including traditional machine learning methods, CNN-based deep learning methods, and ViTs. The results showed that ViTs achieved the best performance in these tasks, and that transfer learning improved ViT by a large margin. Altay et al. [214] aimed to achieve early pre-clinical prediction of AD using MRI. They compared transformers against a baseline 3D CNN model and 3D recurrent visual attention model, and showed that transformers achieved the best accuracy and F1 scores. Adjei-Mensah et al. [215] showed that CNNs outperformed ViTs on low-resolution medical image recognition. Galdran et al. [216] also showed that CNNs outperformed ViTs on diabetic foot ulcer classification in a few-data regime.

In summary, existing studies have not shown that ViTs can outperform CNNs in all scenarios, particularly in both few-shot and low-resolution medical image analysis. Thus, similar to the case for CV methods, most recent studies have built hybrid models with convolutions.

5. Conclusion

Transformers are now transforming the field of computer vision. Also, research using transformers is undergoing rapid growth in the field of medical image analysis. However, most of the current transformer-based methods can be naturally and simply applied to medical imaging problems without drastic changes. Thus, advanced methodologies, *e.g.*, weakly supervised learning, multi-modal learning, multi-task learning, and model improvement, are rarely explored. Also, only a few studies have focused on general problems of the model, *e.g.*, parallelization, interpretability, quantification, and safety. These indicate future directions of for medical transformer research.

Conflicts of interest statement

The authors declare no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 62106101), the Natural Science Foundation of Jiangsu Province (Grant No. BK20210180).

Author contributions

Kelei He, Junfeng Zhang, and Dinggang Shen led the project. Kelei He, Chen Gan, Zhuoyuan Li, Islem Rekik, Zihao Yin, and Wen Ji wrote the paper. Dinggang Shen, Qian Wang, Yang Gao, and Junfeng Zhang revised the paper.

References

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need 2017. doi:10.48550/ARXIV.1706.03762.
[2] Dong L, Xu S, Xu B. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2018. p. 5884–8.
[3] Li N, Liu S, Liu Y, et al. Neural speech synthesis with transformer network. Proceedings of the AAAI Conference on Artificial Intelligence, vol 33; 2019. p. 6706–13.
[4] Vila LC, Escolano C, Fonollosa JA, et al. End-to-end speech translation with the transformer. In: IberSPEECH; 2018. p. 60–3.
[5] Topal MO, Bas A, van Heerden I. Exploring transformers in natural language generation: Gpt, bert, and xlnet. arXiv 2021:210208036.
[6] Graves A, Mohamed Ar, Hinton G. Speech recognition with deep recurrent neural networks. 2013 IEEE international conference on acoustics, speech and signal processing. IEEE; 2013. p. 6645–9.
[7] Sak H, Senior A, Beaufays F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. arXiv 2014:14021128.
[8] Devlin J, Chang MW, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv 2018:181004805.
[9] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training; 2018. Available from https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf, 2018.
[10] Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. arXiv 2020:200514165.
[11] Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017;19:221–48. doi:10.1146/annurev-bioeng-071516-044442.
[12] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Commun ACM 2012;25:1097–105. doi:10.1145/3065386.
[13] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2021.
[14] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers. European Conference on Computer Vision. Springer; 2020. p. 213–29.
[15] Zheng S, Lu J, Zhao H, et al. Rethinking semantic segmentation from a sequence–to-sequence perspective with transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 6881–90.
[16] Parmar N, Vaswani A, Uszkoreit J, et al. Image transformer. International Conference on Machine Learning. PMLR; 2018. p. 4055–64.
[17] Li G, Zhu L, Liu P, et al. Entangled transformer for image captioning. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019. p. 8928–37.
[18] Zhou L, Zhou Y, Corso JJ, et al. End-to-end dense video captioning with masked transformer. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 8739–48.
[19] Gao X, Qian Y, Gao A. Covid-vit: Classification of covid-19 from ct chest images based on vision transformer models. 2021.arXiv. doi:10.48550/arXiv.2107.01682.

[20] Zhang L, Wen Y. Mia-cov19d: A transformer-based framework for covid19 classification in chest cts. arXiv 2021.
[21] He S, Grant PE, Ou Y. Global-Local Transformer for Brain Age Estimation 2021. doi:10.48550/ARXIV.2109.01663.
[22] Costa GSS, Paiva AC, Junior GB, et al. Covid-19 automatic diagnosis with ct images using the novel transformer architecture. Anais do XXI Simpósio Brasileiro de Computação Aplicada à Saúde. SBC; 2021. p. 293–301.
[23] van Tulder G, Tong Y, Marchiori E. Multi-view analysis of unregistered medical images using cross-view transformers. arXiv 2021:210311390.
[24] Zhang Z, Sun B, Zhang W. Pyramid medical transformer for medical image segmentation. arXiv 2021:210414702.
[25] Xie Y, Zhang J, Shen C, et al. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. arXiv 2021:210303024.
[26] Watanabe S, Ueno T, Kimura Y, et al. Generative image transformer (git): unsupervised continuous image generative and transformable model for [123i] fp-cit spect images. Ann Nucl Med 2021;35(11):1203–13. doi:10.1007/s12149-021-01714-4.
[27] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate 2014. doi:10.48550/ARXIV.1409.0473.
[28] Hu J, Shen L, Albanie S, et al. Squeeze-and-Excitation Networks. IEEE Trans Pattern Anal Mach Intell. 2020;42(8):2011–23. doi:10.1109/tpami.2019.2913372.
[29] Chen J, He Y, Frey EC, et al. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. arXiv 2021:210406468.
[30] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention. International Conference on Machine Learning. PMLR 2021.
[31] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. arXiv 2021:210314030.
[32] Tolstikhin I, Houlsby N, Kolesnikov A, et al. Mlp-mixer: An all-mlp architecture for vision. arXiv 2021:210501601.
[33] Liu H, Dai Z, So DR, et al. Pay attention to MLPS. Pay attention to MLPS; 2021.
[34] Touvron H, Bojanowski P, Caron M, et al. Resmlp: Feedforward networks for image classification with data-efficient training. arXiv 2021:210503404.
[35] Lian D, Yu Z, Sun X, et al. As-mlp: An axial shifted mlp architecture for vision. arXiv 2021:210708391.
[36] Chen S, Xie E, Ge C, et al. Cyclemlp: A mlp-like architecture for dense prediction. arXiv 2021:210710224.
[37] Liang S. A hybrid deep learning framework for covid-19 detection via 3d chest ct images. arXiv 2021:210703904.
[38] Barhoumi Y, Ghulam R. Scopeformer: n-cnn-vit hybrid model for intracranial hemorrhage classification. 2021. doi:10.48550/ARXIV.2107.04575.
[39] Li J, Yang Z, Yu Y. A medical ai diagnosis platform based on vision transformer for coronavirus. 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI). IEEE; 2021. p. 246–52.
[40] Than JC, Thon PL, Rijal OM, et al. Preliminary study on patch sizes in vision transformers (vit) for covid-19 and diseased lungs classification. 2021 IEEE National Biomedical Engineering Conference (NBEC). IEEE; 2021. p. 146–150.
[41] Rahimzadeh M, Attar A, Sakhaei SM. A fully automated deep learning-based network for detecting covid-19 from a new and large lung ct scan dataset. Biomed Signal Process Control 2021;68:102588. doi:10.20944/preprints202006.0031.v3.
[42] Xia Y, Yao J, Lu L, et al. Effective pancreatic cancer screening on non-contrast ct scans via anatomy-aware transformers. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 259–69.
[43] Park S, Kim G, Oh Y, et al. Vision transformer for covid-19 cxr diagnosis using chest x-ray feature corpus. arXiv 2021:210307055.
[44] Tanzi L, Audisio A, Cirrincione G, et al. Vision Transformer for femur fracture classification. 2021. doi:10.48550/ARXIV.2108.03414.
[45] Verenich E, Martin T, Velasquez A. Pulmonary disease classification using globally correlated maximum likelihood: an auxiliary attention mechanism for convolutional neural networks. arXiv 2021:210900573.
[46] Cohen JP, Morrison P, Dao L. Covid-19 image data collection. 2020. Available from https://arxiv.org/abs/2003.11597. doi:10.48550/ARXIV.2003.11597.
[47] Chowdhury MEH, Rahman T, Khandakar A, et al. Can ai help in screening viral and covid-19 pneumonia? IEEE Access 2020;8:132665–76. doi:10.1109/ACCESS.2020.3010287.
[48] Liu C, Yin Q. Automatic diagnosis of covid-19 using a tailored transformer-like network. Conference Series, 2010. IOP Publishing; 2021. p. 012175.
[49] Cohen JP, Morrison P, Dao L, et al. Covid-19 image data collection: Prospective predictions are the future. 2020.
[50] Shome D, Kar T, Mohanty SN, et al. Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare. Int J Environ Res Public Health 2021;18(21):11086.
[51] Krishnan KS, Krishnan KS. Vision transformer based covid-19 detection using chest x-rays. 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC). IEEE; 2021. p. 644–8. doi:10.48550/arXiv.2110.04458.
[52] Kim BH, Ye JC, Kim J. Learning dynamic graph representation of brain connectome with spatio-temporal attention. 2021. doi:10.48550/arXiv.2105.13495.
[53] Zhao J, Xiao X, Li D, et al. mftrans-net: Quantitative measurement of hepatocellular carcinoma via multi-function transformer regression network. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 75–84.
[54] Wang S, Zhuang Z, Xuan K, et al. 3dmet: 3d medical image transformer for knee cartilage defect assessment. International Workshop on Machine Learning in Medical Imaging. Springer; 2021. p. 347–55.
[55] Gao Z, Hong B, Zhang X, et al. Instance-based vision transformer for subtyping of papillary renal cell carcinoma in histopathological image. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 299–308.
[56] Chen H, Li C, Li X, et al. Gashis-transformer: A multi-scale visual transformer approach for gastric histopathology image classification. arXiv 2021:210414528.
[57] Zeid MAE, El-Bahnasy K, Abo-Youssef S. Multiclass colorectal cancer histology images classification using vision transformers. 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS). IEEE; 2021. p. 224–30.
[58] Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. J Pathol Inform 2016;7.
[59] Ikromjanov K, Bhattacharjee S, Hwang YB, et al. Whole slide image analysis and detection of prostate cancer using vision transformers. 2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). IEEE; 2022. p. 399–402.
[60] Zhao C, Shuai R, Ma L, et al. Improving cervical cancer classification with imbalanced datasets combining taming transformers with t2t-vit. Multimed Tools Appl 2022:1–36.
[61] Perera S, Adhikari S, Yilmaz A. Pocformer: A lightweight transformer architecture for detection of covid-19 using point of care ultrasound. arXiv 2021:210509913.
[62] Gheflati B, Rivaz H. Vision transformer for classification of breast ultrasound images. arXiv 2021:211014731.
[63] Al-Dhabyani W, Gomaa M, Khaled H, et al. Dataset of breast ultrasound images. Data Brief 2020;28:104863. doi:10.1016/j.dib.2019.104863.
[64] Yap MH, Pons G, Martí J, et al. Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J Biomed Health Inform 2018;22(4):1218–26. doi:10.1109/JBHI.2017.2731873.
[65] Jiang Z, Dong Z, Wang L, et al. Method for diagnosis of acute lymphoblastic leukemia based on vit-cnn ensemble model. Comput Intell Neurosci 2021;2021.
[66] Xie J, Wu Z, Zhu R, et al. Melanoma detection based on swin transformer and simam. 2021 IEEE 5th Information Technology,Networking,Electronic and Automation Control Conference (ITNEC), vol 5; 2021. p. 1517–21.
[67] Li X, Desrosiers C, Liu X. Out-of-distribution detection using vision transformers; 2021.
[68] Yu Z, Mar V, Eriksson A, et al. End-to-end ugly duckling sign detection for melanoma identification with transformers. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 176–84.
[69] Wu W, Mehta S, Nofallah S, et al. Scale-aware transformers for diagnosing melanocytic lesions. IEEE Access 2021;9:163526–41.
[70] Yang H, Chen J, Xu M. Fundus disease image classification based on improved transformer. 2021 International Conference on Neuromorphic Computing (ICNC). IEEE; 2021. p. 207–14.
[71] Song D, Fu B, Li F, et al. Deep relation transformer for diagnosing glaucoma with optical coherence tomography and visual field function. IEEE Trans Med Imaging 2021.
[72] Yuan L, Hou Q, Jiang Z, et al. Volo: Vision outlooker for visual recognition. arXiv 2021:210613112.
[73] Selvaraju RR, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision; 2017. p. 618–26.
[74] AlDahoul N, Karim HA, Tan MJT, et al. Encoding retina image to words using ensemble of vision transformers for diabetic retinopathy grading. F1000research 2021;10(948):948.
[75] Chen F, Wang YC, Wang B, et al. Graph representation learning: a survey. APSIPA Transactions on Signal and Information Processing 2020;9. doi:10.1017/atsip.2020.13.
[76] Liu J, Li M, Pan Y, et al. Complex brain network analysis and its applications to brain disorders: a survey. Complexity 2017;2017.
[77] Bessadok A, Mahjoub MA, Rekik I. Graph neural networks in network neuroscience. arXiv 2021:210603535.
[78] Chen J, Lu Y, Yu Q, et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv 2021:210204306.
[79] Chang Y, Menghan H, Guangtao Z, et al. Transclaw u-net: Claw u-net with transformers for medical image segmentation. arXiv 2021:210705188.
[80] Xu G, Wu X, Zhang X, et al. Levit-unet: Make faster encoders with transformer for medical image segmentation. arXiv 2021:210708623.
[81] Sha Y, Zhang Y, Ji X, et al. Transformer-unet: Raw image processing with unet. arXiv 2021:210908417.
[82] Li Y, Cai W, Gao Y, et al. More than encoder: Introducing transformer decoder to upsample. arXiv 2021:210610637.
[83] Simpson AL, Antonelli M, Bakas S, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv 2019:190209063.
[84] Gao Y, Zhou M, Metaxas DN. Utnet: a hybrid transformer architecture for medical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 61–71.
[85] Campello VM, Gkontra P, Izquierdo C, et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. IEEE Trans Med Imaging 2021;40(12):3543–54. doi:10.1109/TMI.2021.3090082.
[86] Fu Z, Zhang J, Luo R, et al. Tf-unet: An automatic cardiac mri image segmentation method. Math Biosci Eng 2022;19(5):5207–22.
[87] Gao Y, Zhou M, Liu D, et al. A multi-scale transformer for medical image segmentation: Architectures, model efficiency, and benchmarks. arXiv 2022:220300131.
[88] Sun Q, Fang N, Liu Z, et al. Hybridctrm: Bridging cnn and transformer for multimodal brain image segmentation. J Healthc Eng 2021;2021. doi:10.1155/2021/7467261.

[89] Mendrik AM, Vincken KL, Kuijf HJ, et al. Mrbrains challenge: online evaluation framework for brain image segmentation in 3t mri scans. Comput Intell Neurosci 2015;2015.
[90] Wang L, Nie D, Li G, et al. Benchmark on automatic six-month-old infant brain segmentation algorithms: the iseg-2017 challenge. IEEE Trans Med Imaging 2019;38(9):2219–30. doi:10.1109/TMI.2019.2901712.
[91] Zhang Y, Liu H, Hu Q. Transfuse: Fusing transformers and cnns for medical image segmentation. arXiv 2021:210208005.
[92] Codella NC, Gutman D, Celebi ME, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). IEEE; 2018. p. 168–72.
[93] You C, Zhao R, Liu F, et al. Class-aware generative adversarial transformers for medical image segmentation. arXiv 2022:220110737.
[94] Xu S, Quan H. Ect-nas: Searching efficient cnn-transformers architecture for medical image segmentation. 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2021. p. 1601–4.
[95] Luo H, Changdong Y, Selvan R. Hybrid ladder transformers with efficient parallel-cross attention for medical image segmentation; 2021.
[96] Liu W, Tian T, Xu W, et al. Phtrans: Parallelly aggregating global and local representations for medical image segmentation. arXiv 2022:220304568.
[97] Zhou HY, Guo J, Zhang Y, et al. nnformer: Interleaved transformer for volumetric segmentation. arXiv 2021:210903201.
[98] Sirinukunwattana K, Pluim JP, Chen H, et al. Gland segmentation in colon histology images: The glas challenge contest. Med Image Anal 2017;35:489–502. doi:10.1016/j.media.2016.08.008.
[99] Kumar N, Verma R, Sharma S, et al. A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Trans Med Imaging 2017;36(7):1550–60. doi:10.1109/TMI.2017.2677499.
[100] Valanarasu JMJ, Oza P, Hacihaliloglu I, et al. Medical transformer: Gated axial-attention for medical image segmentation. arXiv 2021:210210662.
[101] Valanarasu JMJ, Yasarla R, Wang P. Learning to segment brain anatomy from 2d ultrasound with less data. IEEE J STSP 2020;14(6):1221–34. doi:10.1109/JSTSP.2020.3001513.
[102] Landman B, Xu Z, Igelsias J, et al. 2015 miccai multi-atlas labeling beyond the cranial vault workshop and challenge; 2015.
[103] Ji Y, Zhang R, Wang H, et al. Multi-compound transformer for accurate biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 326–36.
[104] Gamper J, Koohbanani NA, Benet K, et al. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. European Congress on Digital Pathology. Springer; 2019. p. 11–19.
[105] Codella N, Rotemberg V, Tschandl P, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv 2019:190203368.
[106] Wu Y, Liao K, Chen J, et al. D-former: A u-shaped dilated transformer for 3d medical image segmentation. arXiv 2022:220100462.
[107] Hatamizadeh A, Yang D, Roth H, et al. Unetr: Transformers for 3d medical image segmentation. arXiv 2021:210310504.
[108] Hatamizadeh A, Nath V, Tang Y, et al. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. arXiv 2022:220101266.
[109] Chen B, Liu Y, Zhang Z, et al. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. arXiv 2021:210705274.
[110] Tang YB, Tang YX, Xiao J, et al. Xlsor: A robust and accurate lung segmentor on chest x-rays using criss-cross attention and customized radiorealistic abnormalities generation. International Conference on Medical Imaging with Deep Learning. PMLR; 2019. p. 457–67.
[111] He X, Wang S, Shi S, et al. Benchmarking deep learning models and automated model design for covid-19 detection with chest ct scans. medRxiv 2020.
[112] Caicedo JC, Goodman A, Karhohs KW, et al. Nucleus segmentation across imaging experiments: the 2018 data science bowl. Nat Methods 2019;16(12):1247–53. doi:10.1038/s41592-019-0612-7.
[113] Wang H, Xie S, Lin L, et al. Mixed transformer u-net for medical image segmentation. arXiv 2021:211104734.
[114] Yan X, Tang H, Sun S, et al. After-unet: Axial fusion transformer unet for medical image segmentation. arXiv 2021:211010403.
[115] Chen X, Sun S, Bai N, et al. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. Radiother Oncol 2021;160:175–84. doi:10.1016/j.radonc.2021.04.019.
[116] Lambert Z, Petitjean C, Dubray B, et al. Segthor: Segmentation of thoracic organs at risk in ct images. 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA). IEEE; 2020. p. 1–6.
[117] Zhang J, Liu Y, Wu Q, et al. Swin-Unet: Star-shaped window transformer onion u-net for medical image segmentation.
[118] Karimi D, Vasylechko S, Gholipour A. Convolution-free medical image segmentation using transformers. arXiv 2021:210213645.
[119] Cao H, Wang Y, Chen J, et al. Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv 2021:210505537.
[120] Lin A, Chen B, Xu J, et al. Ds-transunet: Dual swin transformer u-net for medical image segmentation. arXiv 2021:210606716.
[121] Huang X, Deng Z, Li D, et al. Missformer: An effective medical image segmentation transformer. arXiv 2021:210907162.
[122] Jha D, Smedsrud PH, Riegler MA, et al. Kvasir-seg: A segmented polyp dataset. International Conference on Multimedia Modeling. Springer; 2020. p. 451–462.
[123] Bernal J, Sánchez FJ, Fernández-Esparrach G, et al. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Comput Med Imaging Graph 2015;43:99–111. doi:10.1016/j.compmedimag.2015.02.007.
[124] Tajbakhsh N, Gurudu SR, Liang J. Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans Med Imaging 2015;35(2):630–44. doi:10.1109/TMI.2015.2487997.
[125] Vázquez D, Bernal J, Sánchez FJ, et al. A benchmark for endoluminal scene segmentation of colonoscopy images. J Healthc Eng 2017;2017. doi:10.1155/2017/4037190.
[126] Silva J, Histace A, Romain O, et al. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. Int J Comput Assist Radiol Surg 2014;9(2):283–93. doi:10.1007/s11548-013-0926-3.
[127] Ning Y, Zhang S, Xi X, et al. Cac-emvt: Efficient coronary artery calcium segmentation with multi-scale vision transformers. 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2021. p. 1462–1467.
[128] Wang W, Chen C, Ding M, et al. Transbts: Multimodal brain tumor segmentation using transformer. arXiv 2021:210304430.
[129] Jun E, Jeong S, Heo DW, et al. Medical transformer: Universal brain encoder for 3d mri analysis. arXiv 2021:210413633.
[130] Ranem A, González C, Mukhopadhyay A. Continual hippocampus segmentation with transformers. arXiv 2022:220408043.
[131] Laiton-Bonadiez C, Sanchez-Torres G, Branch-Bedoya J. Deep 3d neural network for brain structures segmentation using self-attention modules in mri images. Sensors 2022;22(7):2559. doi:10.2174/2210327910999200728145536.
[132] Rao VM, Wan Z, Ma DJ, et al. Improving across-dataset brain tissue segmentation using transformer. arXiv 2022:220108741.
[133] Liang J, Yang C, Zeng M, et al. Transconver: transformer and convolution parallel network for developing automatic brain tumor segmentation in mri images. Quant Imaging Med Surg 2022;12(4):2397.
[134] Hatamizadeh A, Xu Z, Yang D, et al. Unetformer: A unified vision transformer model and pre-training framework for 3d medical image segmentation. arXiv 2022:220400631.
[135] Wang Q, Li L, Ni B, et al. Medical image segmentation using transformer. Artificial Intelligence in China. Springer; 2022. p. 92–9.
[136] Huang Z, Liao J, Wei J, et al. Transde: A transformer and double encoder network for medical image segmentation. 2021 11th International Conference on Information Technology in Medicine and Education (ITME). IEEE; 2021. p. 374–378.
[137] Hille G, Agrawal S, Wybranski C, et al. Joint liver and hepatic lesion segmentation using a hybrid cnn with transformer layers. arXiv 2022:220110981.
[138] Wang L, Wang X, Zhang B, et al. Multi-scale hierarchical transformer structure for 3d medical image segmentation. 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2021. p. 1542–5.
[139] Li L, Ma H. RDCTrans U-Net: A Hybrid Variable Architecture for Liver CT Image Segmentation. Sensors 2022;22(7):2452. doi:10.3390/s22072452.
[140] Wang B, Wang F, Dong P, et al. Multiscale transunet + +: dense hybrid u-net with transformer for medical image segmentation. Signal Image Video Process 2022:1–8. doi:10.1007/s11760-021-02115-w.
[141] Wang J, Wei L, Wang L, et al. Boundary-aware transformers for skin lesion segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 206–16.
[142] Zhang Y, Higashita R, Fu H, et al. A multi-branch hybrid transformer networkfor corneal endothelial cell segmentation. arXiv 2021:210607557.
[143] Prangemeier T, Reich C, Koeppl H. Attention-based transformers for instance segmentation of cells in microstructures. 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2020. p. 700–7.
[144] Chen D, Yang W, Wang L, et al. Pcat-unet: Unet-like network fused convolution and transformer for retinal vessel segmentation. PLoS One 2022;17(1):e0262689. doi:10.1371/journal.pone.0262689.
[145] Yun B, Wang Y, Chen J, et al. Spectr: Spectral transformer for hyperspectral pathology image segmentation. arXiv 2021:210303604.
[146] Marek K, Jennings D, Lasch S, et al. The parkinson progression marker initiative (ppmi). Prog Neurobiol 2011;95(4):629–35. doi:10.1016/j.pneurobio.2011.09.005.
[147] Kamran SA, Hossain KF, Tavakkoli A, et al. Vtgan: Semi-supervised retinal image synthesis and disease prediction using vision transformers. arXiv 2021:210406757.
[148] Hajeb Mohammad Alipour S, Rabbani H, Akhlaghi MR. Diabetic retinopathy grading by digital curvelet transform. Comput Math Methods Med 2012;2012. doi:10.1155/2012/761901.
[149] Shin HC, Ihsani A, Mandava S, et al. Ganbert: Generative adversarial networks with bidirectional encoder representations from transformers for mri to pet synthesis. arXiv 2020:200804393.
[150] Hu Z, Liu H, Li Z, et al. Data-enabled intelligence in complex industrial systems cross-model transformer method for medical image synthesis. Complexity 2021;2021.
[151] Korkmaz Y, Dar SU, Yurt M, et al. Unsupervised mri reconstruction via zero-shot learned adversarial transformers. arXiv 2021:210508059.
[152] Knoll F, Zbontar J, Sriram A, et al. fastMRI: A publicly available raw k-space and dicom dataset of knee images for accelerated mr image reconstruction using machine learning. Radiol Artif Intell 2020;2(1):e190007. doi:10.1148/ryai.2020190007.
[153] Ristea NC, Miron AI, Savencu O, et al. Cytran: Cycle-consistent transformers for non-contrast to contrast ct translation. arXiv 2021:211006400.
[154] Dalmaz O, Yurt M, Çukur T. Resvit: Residual vision transformers for multi-modal medical image synthesis. arXiv 2021:210616031.

[155] Menze BH, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging 2014;34(10):1993–2024. doi:10.1109/TMI.2014.2377694.
[156] Bakas S, Akbari H, Sotiras A, et al. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Sci Data 2017;4(1):1–13. doi:10.1038/sdata.2017.117.
[157] Bakas S, Reyes M, Jakab A, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv 2018:181102629.
[158] Nyholm T, Svensson S, Andersson S, et al. Mr and ct data with multiobserver delineations of organs in the pelvic area-part of the gold atlas project. Med Phys 2018;45(3):1295–300. doi:10.1002/mp.12748.
[159] Feng CM, Yan Y, Fu H, et al. Task transformer network for joint mri reconstruction and super-resolution. arXiv 2021:210606742.
[160] Zhang X, He X, Guo J, et al. Ptnet: A high-resolution infant mri synthesizer based on transformer. arXiv 2021:210513993.
[161] Makropoulos A, Robinson EC, Schuh A, et al. The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction. Neuroimage 2018;173:88–112. doi:10.1016/j.neuroimage.2018.01.054.
[162] Wang D, Wu Z, Yu H. Ted-net: Convolution-free t2t vision transformer-based encoder-decoder dilation network for low-dose ct denoising. arXiv 2021:210604650.
[163] McCollough C. Tu-fg-207a-04: Overview of the low dose ct grand challenge. Medical physics 2016;43(6Part35):3759–60.
[164] Luthra A, Sulakhe H, Mittal T, et al. Eformer: Edge enhancement based transformer for medical image denoising. arXiv 2021:210908044.
[165] Jiang H, Zhang P, Che C, et al. RDFNet: A Fast Caries Detection Method Incorporating Transformer Mechanism. Comput Math Methods Med 2021;2021. doi:10.1155/2021/9773917.
[166] Shen Z, Lin C, Zheng S. Cotr: Convolution in transformer network for end to end polyp detection. arXiv 2021:210510925.
[167] Ma X, Luo G, Wang W, et al. Transformer network for significant stenosis detection in ccta of coronary arteries. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 516–25.
[168] Kong Q, Wu Y, Yuan C, et al. Ct-cad: Context-aware transformers for end-to-end chest abnormality detection on x-rays. 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2021. p. 1385–8.
[169] Tao R, Zheng G. Spine-transformers: Vertebra detection and localization in arbitrary field-of-view spine ct with transformers. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 93–103.
[170] Li Y, Wang Z, Yin L, et al. X-net: a dual encoding–decoding method in medical image segmentation. The Visual Computer 2021:1–11.
[171] Liang J, Cao J, Sun G, et al. Swinir: Image restoration using swin transformer. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 1833–44.
[172] Yang F, Yang H, Fu J, et al. Learning texture transformer network for image super-resolution. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 5791–800.
[173] Luo Y, Wang Y, Zu C, et al. 3d transformer-gan for high-quality pet reconstruction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 276–85.
[174] Balakrishnan G, Zhao A, Sabuncu MR, et al. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. IEEE Trans Med Imaging 2019;38(8):1788–800. doi:10.1109/TMI.2019.2897538.
[175] Chen J, Du Y, He Y, et al. Transmorph: Transformer for unsupervised medical image registration. arXiv 2021:211110480.
[176] Ji GP, Chou YC, Fan DP, et al. Progressively normalized self-attention network for video polyp segmentation. arXiv 2021:210508468.
[177] Kondo S. Lapformer: surgical tool detection in laparoscopic surgical video using transformer architecture. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 2021;9(3):302–7. doi:10.1080/21681163.2020.1835550.
[178] Czempiel T, Paschali M, Ostler D, et al. Opera: Attention-regularized transformers for surgical phase recognition. arXiv 2021:210303873.
[179] Reynaud H, Vlontzos A, Hou B, et al. Ultrasound video transformers for cardiac ejection fraction estimation. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 495–505.
[180] Long Y, Li Z, Yee CH, et al. E-dssr: efficient dynamic surgical scene reconstruction with transformer-based stereoscopic depth perception. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 415–25.
[181] Nguyen HH, Saarakkala S, Blaschko MB, et al. Climat: Clinically-inspired multi-agent transformers for disease trajectory forecasting from multi-modal data. arXiv 2021:210403642.
[182] Zheng S, Zhu Z, Liu Z, et al. Multi-modal graph learning for disease prediction. 2021.
[183] Qiu Y, Yu S, Zhou Y, et al. Multi-channel sparse graph transformer network for early alzheimer's disease identification. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI); 2021. p. 1794–7.
[184] Monajatipoor M, Rouhsedaghat M, Li LH, et al. Berthop: An effective vision-and-language model for chest x-ray disease diagnosis. arXiv 2021:210804938.
[185] Dai Y, Gao Y, Liu F. Transmed: Transformers advance multi-modal medical image classification. Diagnostics 2021;11(8):1384. doi:10.3390/diagnostics11081384.
[186] Jacenków G, O'Neil AQ, Tsaftaris SA. Indication as prior knowledge for multimodal disease classification in chest radiographs with transformers. 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI); 2022. p. 1–5.
[187] Azzuni H, Ridzuan M, Xu M, et al. Color space-based hover-net for nuclei instance segmentation and classification. arXiv: 2022:220301940.
[188] Chen J, Chen J, Zhou Z, et al. Mt-transunet: Mediating multi-task tokens in transformers for skin lesion segmentation and classification. 2021. Available from https://arxiv.org/abs/2112.01767. doi:10.48550/ARXIV.2112.01767.
[189] Sui D, Zhang K, Liu W, et al. Cst: A multitask learning framework for colorectal cancer region mining based on transformer. Biomed Res Int 2021;2021. doi:10.1155/2021/6207964.
[190] Han K, Wang Y, Chen H, et al. A survey on visual transformer. arXiv 2020:201212556.
[191] Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. 2021.
[192] Li Z, Yuan L, Xu H, et al. Deep multi-instance learning with induced self-attention for medical image classification. 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2020. p. 446–50.
[193] Rymarczyk D, Borowa A, Tabor J, et al. Kernel self-attention for weakly-supervised image classification using deep multiple instance learning. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2021. p. 1721–30.
[194] Yang J, Deng H, Huang X, et al. Relational learning between multiple pulmonary nodules via deep set attention transformers. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE; 2020. p. 1875–8.
[195] Yu S, Ma K, Bi Q, et al. Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 45–54.
[196] Shao Z, Bian H, Chen Y, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classication. arXiv 2021:210600908.
[197] Wang X, Yang S, Zhang J, et al. Transpath: Transformer-based self-supervised learning for histopathological image classification. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 186–95.
[198] Truong T, Mohammadi S, Lenga M. How transferable are self-supervised features in medical image classification tasks? arXiv 2021:210810048.
[199] Sriram A, Muckley M, Sinha K, et al. Covid-19 prognosis via self-supervised representation learning and multi-image prediction. arXiv 2021:210104909.
[200] Carbonneau MA, Cheplygina V, Granger E, et al. Multiple instance learning: A survey of problem characteristics and applications. Pattern Recognit 2018;77:329–53. doi:10.1016/j.patcog.2017.10.009.
[201] Luo X, Hu M, Song T, et al. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. arXiv 2021:211204894.
[202] Zhao C, Xiang S, Cai Z, et al. Context-aware network for semi-supervised segmentation of 3d left atrium.
[203] Xiao Z, Su Y, Deng Z, et al. Efficient combination of cnn and transformer for dual-teacher uncertainty-aware guided semi-supervised medical image segmentation.
[204] Zhuang F, Qi Z, Duan K, et al. A comprehensive survey on transfer learning. Proceedings of the IEEE 2020;109(1):43–76.
[205] Matsoukas C, Haslum JF, Söderberg M, et al. Is it time to replace cnns with transformers for medical images? arXiv 2021:210809038.
[206] Luo Y, Chen Z, Gao X. Self-distillation augmented masked autoencoders for histopathological image classification. arXiv 2022:220316983.
[207] Malkiel I, Rosenman G, Wolf L, et al. Pre-training and fine-tuning transformers for fmri prediction tasks. arXiv 2021:211205761.
[208] Xie Y, Zhang J, Xia Y, et al. Unified 2d and 3d pre-training for medical image classification and segmentation. arXiv 2021:211209356.
[209] Chen RJ, Krishnan RG. Self-supervised vision transformers learn visual concepts in histopathology. arXiv 2022:220300585.
[210] Li J, Wang W, Chen C, et al. Transbtsv2: Wider instead of deeper transformer for medical image segmentation. arXiv 2022:220112785.
[211] Krishnamurthy S, Srinivasan K, Qaisar SM, et al. Evaluating deep neural network architectures with transfer learning for pneumonitis diagnosis. Comput Math Methods Med 2021;2021. doi:10.1155/2021/8036304.
[212] Liu W, Li C, Rahamana MM, et al. Is aspect ratio of cells important in deep learning? a robust comparison of deep learning methods for multi-scale cytopathology cell image classification: from convolutional neural networks to visual transformers. 2021.
[213] Zhou M, Mo S. Shoulder implant x-ray manufacturer classification: Exploring with vision transformer. 2021.
[214] Altay F, Sanchez GR, James Y, et al. Preclinical stage alzheimer's disease detection using magnetic resonance image scans. 2020.
[215] Adjei-Mensah I, Zhang X, Baffour AA, et al. Investigating vision transformer models for low-resolution medical image recognition. 2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP); 2021. p. 179–83.
[216] Galdran A, Carneiro G, Ballester MAG. Convolutional nets versus vision transformers for diabetic foot ulcer classification. Diabetic Foot Ulcers Grand Challenge. Springer; 2021. p. 21–9.