BioMedical Engineering OnLine

REVIEW Open Access

A survey of Transformer applications for histopathological image analysis: New developments and future directions

Chukwuemeka Clinton Atabansi[1], Jing Nie[1*], Haijun Liu[1], Qianqian Song[1], Lingfeng Yan[1] and Xichuan Zhou[1*]

*Correspondence: jingnie@cqu.edu.cn; zxc@cqu.edu.cn

[1] School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China

Abstract

Transformers have been widely used in many computer vision challenges and have shown the capability of producing better results than convolutional neural networks (CNNs). Taking advantage of capturing long-range contextual information and learning more complex relations in the image data, Transformers have been used and applied to histopathological image processing tasks. In this survey, we make an effort to present a thorough analysis of the uses of Transformers in histopathological image analysis, covering several topics, from the newly built Transformer models to unresolved challenges. To be more precise, we first begin by outlining the fundamental principles of the attention mechanism included in Transformer models and other key frameworks. Second, we analyze Transformer-based applications in the histopathological imaging domain and provide a thorough evaluation of more than 100 research publications across different downstream tasks to cover the most recent innovations, including survival analysis and prediction, segmentation, classification, detection, and representation. Within this survey work, we also compare the performance of CNN-based techniques to Transformers based on recently published papers, highlight major challenges, and provide interesting future research directions. Despite the outstanding performance of the Transformer-based architectures in a number of papers reviewed in this survey, we anticipate that further improvements and exploration of Transformers in the histopathological imaging domain are still required in the future. We hope that this survey paper will give readers in this field of study a thorough understanding of Transformer-based techniques in histopathological image analysis, and an up-to-date paper list summary will be provided at https://github.com/S-domain/Survey-Paper.

Keywords: Transformer, Histopathological imaging, CNN, Whole slide image, Survival analysis, Digital pathology

Introduction

Histopathological imaging has been regarded as a technique for identifying nearly all types of cancers since it provides a more thorough understanding of the diseases [1, 2]. They are a very important source of primary information in clinical domains, which assists pathologists in performing cancer diagnosis. Histopathological images are mostly

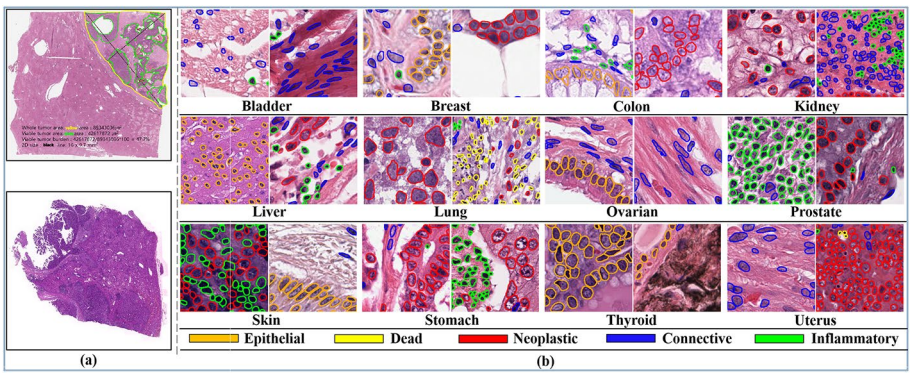

Fig. 1 Some samples of histopathological images. **a** Whole slide images (WSIs). **b** Annotated PanNuke dataset from different tissue types for nuclei instance classification and segmentation

used for cancer grading and offer more detailed information for diagnosis when compared to other medical imaging techniques, including magnetic resonance imaging (MRI), computerized tomography (CT), transrectal ultrasound (TRUS), mammography, and many others, and diseases are also examined by identifying the cells and tissue present in lesions [1, 3]. For various cancer types, pathologists may choose treatment plans based on histopathological images coupled with genomic records. With the recent development and deployment of digital slide scanners in different clinical areas, the digitization of histopathological slides (i.e., whole slide images (WSIs)) into gigapixel images is becoming more prevalent. In computational pathology, histopathological slides (WSIs) display a hierarchical formation of visual tokens across different resolutions and can have a pixel size up to 160,000 $\times$ 160,000 pixels at 20$\times$magnification. Figure 1 shows some samples of histopathological slides and some annotated patches extracted from the slides that contain different tissue types.

The technique of digitizing histopathological images, known as digital pathology, creates a new approach to collecting image data for artificial intelligence technologies. In recent years, artificial intelligence techniques that process and analyze histopathological images have become more common in both scientific research and clinical settings. This is primarily due to the rise of deep learning, especially convolutional neural networks (CNNs), which have achieved outstanding results in many computer vision tasks [4–6]. Recently, an alternative CAD system that is capable of modeling long-range pixel information, such as transformers, has been developed. Transformers [7] have emerged as one of the most recent technological developments in deep learning for achieving robust results in many computer vision tasks. It was first built as a robust example of using deep learning techniques to tackle sequential inference tasks in natural language processing (NLP). Dosovitskiy [8] et al. introduced a vision transformer (ViT)-based architecture for image classification tasks, demonstrating that relying on CNNs for image classifications is unnecessary and that a pure transformer applied straight to sequences of image patches can get excellent results. Other than images and NLP tasks, transformers have also been adopted and applied to other deep learning domains, including autonomous driving [9], video classification [10], security [11], general audio representations [12], audio–video synchronization [13], mobile devices [14] and so on. Motivated by this innovation, several studies have adopted a

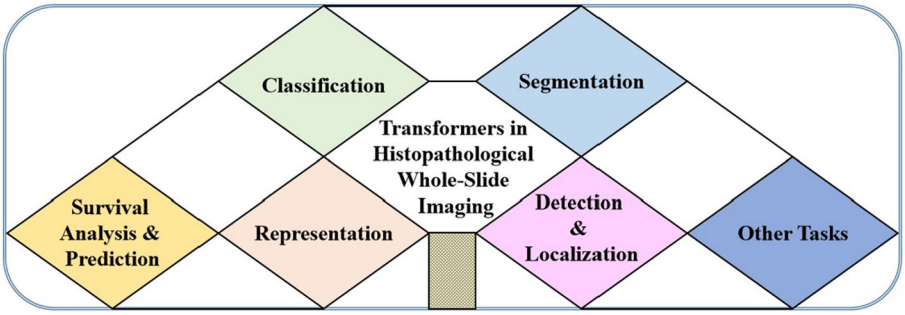

Fig. 2 Current transformer applications in histopathological image analysis, as surveyed in this research work

variety of approaches to solve different deep learning challenges, including CNN- and transformer-based approaches, but it still remains unclear whether ViT architectures can produce better results than CNNs for histopathological image analysis. Transformers, like any other deep learning or machine learning technique, have pros and cons. Besides, transformers, unlike CNN-based approaches, are devoid of convolution-induced biases, which enables them to capture long-range contextual information and learn more complex relations in the image data. This is advantageous in histopathological imaging, where it is critical to consider not just the region of interest, but also the neighboring tissues when diagnosing a particular disease. Transformers, on the other hand, are data-demanding and require greater computing effort. This can be a difficult problem, especially in the field of histopathological imaging, where resources may be inadequate due to concerns about patient privacy. At present, many studies have been conducted in the field of histopathological imaging using transformer-based approaches, including image segmentation [2, 15], classification [16, 17], detection [18, 19], representation [20–22], cross-modal retrieval [23], image generation [24], survival analysis [25] and survival prediction [26]. Figure 2 displays current transformer applications in histopathological image analysis, as surveyed in this research work, which will be further explored in "Current progress" Sect.

However, based on the recently published studies, it has been shown that transformer architectures have the capacity to achieve higher performance on various histopathological imaging tasks than the previous models.

Moreover, the primary aim of the paper is to provide a thorough review of transformer applications in the histopathological imaging field and demonstrate how transformers are applied to a variety of tasks. In particular, it provides readers in this field of study with a thorough understanding of transformer-based techniques in histopathological image analysis and also establishes the foundation for future innovation to improve the performance of transformer architectures in this domain. To this end, our key contributions include: (1) this work provides a thorough evaluation of more than 100 research publications across the histopathological imaging field to cover the most recent innovations; (2) it provides a thorough overview of the entire domain by classifying the research papers according to how they apply to histopathological imaging, as shown in Fig. 2; (3) it classified each of these applications, pointed out task-specific challenges, and highlighted the approaches used to address them

based on the proposed work, as demonstrated in Tables 1, 2, 3, 4, and 5 in "Current progress" , "Discussion" Section it provides a thorough analysis of designing transformer-based approaches for handling more difficult real-world challenges and also compares transformers with CNN-based models based on recently published works. The remainder of this survey paper is structured as follows: "Background" Sect. Provides a brief background on the study and basic components of transformers. In Current progressSect. Current applications of transformers in histopathological image analysis are investigated. The discussions and conclusion are covered in Discussion, "Conclusion "Sect. Respectively.

Background

Over the years, histopathological imaging computer-aided diagnosis (CAD) systems have witnessed a lot of technological advancement following the advent of transformer architectures. However, in this part, we will give a quick overview of CNN-based approaches and outline the basic operating principles together with their main advantages and drawbacks in the field of histopathological imaging. In addition, we will also discuss the fundamental ideas that underlie the success of the transformer-based techniques and then provide further information in subsequent sections. Finally, we compare the CNN methods versus the transformer methods.

CNN applications in histopathological image analysis

For some years now, CNNs have proven to be good at analyzing image data and are the most widely used deep learning networks for many medical and clinical challenges, especially histopathological imaging. This is as a result of the strong prior that the convolution operations impose on the weights, forcing the identical weights to be shared across each and every pixel [27]. The major advantage of CNN-based approaches compared to previous architectures is their ability to automatically identify important features in an image without any form of human oversight. The process of building any CNN architecture for histopathological image analysis is a collaborative effort between researchers and medical professionals. These innovations are primarily driven by a lot of architectural advancements, improved loss functions, the accessibility of specialized hardware devices, and publicly accessible libraries created for specific purposes. Therefore, we direct readers who are interested in this research direction to some previously published survey papers on CNN applications in the histopathological imaging field [4–6]. Although CNN-based techniques have experienced a lot of architectural improvements over the years, their ability to be applied to the full range of histopathological image tasks is also constrained by their dependency on huge amounts of labeled datasets. The study of histopathological imaging for different clinical tasks has also been cross-pollinated by the CNN models [28–30], and they sometimes function as black box solutions and are typically more difficult to explain. However, the success of CNN-based methods is primarily due to their capacity to extract useful information from input images, doing away with the necessity for conventional manual image processing techniques. Despite increasing the receptive field, they still face a lot of challenges in modeling long-range information as well as spatial dependencies due to their weight sharing and inductive bias locality. The local nature of the convolutional operations in CNNs is

the major challenge associated with CNN-based techniques, as it prevents them from capturing long-range semantic dependencies from the given input images. Thus, an alternative CAD system that is capable of modeling long-range pixel information, such as transformers, is required to achieve more robust results than the previous models.

Transformers

Basics

Transformer-based architectures are the most advanced technique for handling sequences. They make use of attention mechanisms due to their capacity to model long-range semantic information. Besides, they also make use of an encoder–decoder design strategy that produces an output without relying on recurrence and convolutions. As a result, we first begin by giving a brief introduction to the basic ideas behind the attention mechanism, followed by a comprehensive explanation of how the transformer operates.

Attention mechanism

The attention mechanism evolved naturally from sequence-related challenges. Nowadays, it is often used to extract unimportant information from the data while concentrating on the relevant portions of the data, and it can be used for a number of deep learning architectures across different clinical domains and downstream tasks. An attention mechanism was initially developed to boost machine translation encoder–decoder performance. It was initially introduced by Bahdanau et al. [31] for the language translation task to tackle the bottleneck that results from the use of a fixed-length encoding vector, where the decoder would have minimal access to the information delivered by the input. This is viewed as being especially troublesome for long or sophisticated sequences because the representation's dimensionality would be limited to match that of unsophisticated or shorter sequences.

(i) Attention mechanisms in computer vision tasks:

The concept of emulating human attention emerged in the computer vision domain in an attempt to minimize the computational problem of image processing while increasing accuracy by adding a model that only focused on certain portions of images rather than the whole image. However, the attention mechanisms we employ today in our various models originated in the field of NLP. Several studies have been proposed in the past to incorporate attention mechanisms into their architecture. For example, the work in [32] instead focuses on the interaction between channels and develops a new attention mechanism framework known as squeeze-and-excitation that explicitly models the interdependencies between channels and adaptively recalibrates channel-wise feature responses. In contrast to Bahdanau attention, the attention mechanism, as proposed in [7], has been reconstructed as a function that uses values, keys, and queries that are attained from the module's input vectors. In practice, the values and keys are constructed together into matrices V and K, while the attention function is computed simultaneously on a set of queries and arranged together into a matrix Q. Then, the output function is determined as a weighted sum of values, where each value of the weight is computed as the attention between queries and keys, respectively. In addition, the operation of self-attention, as illustrated in Fig. 3, is typically performed in matrix formation in order to speed up the parallel calculation. Additionally,

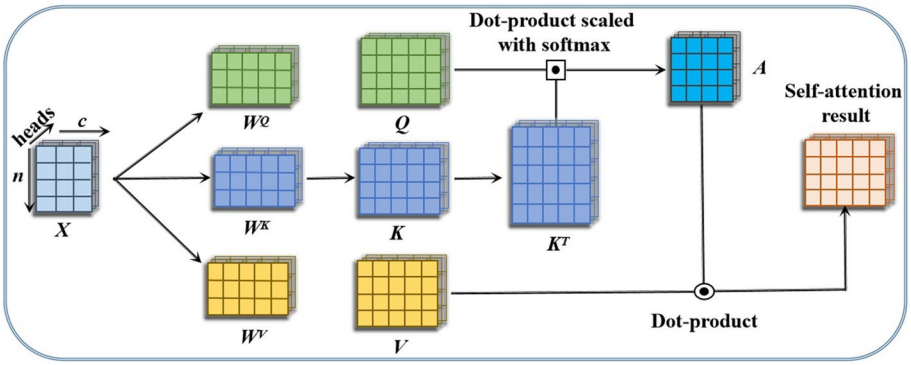

Fig. 3 A schematic demonstration of the self-attention mechanism

in order to quickly demonstrate a clear picture of the self-attention mechanism, we begin by defining it in an element-wise manner. Let $x_i \in \mathbb{R}^c, i = 1, .., m$, be the input image, and the corresponding vectors generated by the parameters (i.e., W^q, W^k, *and* W^v) be query $q_i \in \mathbb{R}_q^g$, key $k_i \in \mathbb{R}_k^g$, and value $v_i \in \mathbb{R}_v^g$, respectively. Again, g_q, g_k, and g_v represent the number of features learned from x_i and also the sizes of q_i, k_i, and v_i, respectively.

$$\begin{cases} q_i = x_i \times W^q, where\ W^q \in \mathbb{R}^{c \times g_q}, \\ k_i = x_i \times W^k, where\ W^k \in \mathbb{R}^{c \times g_k}, \\ v_i = x_i \times W^v, where\ W^v \in \mathbb{R}^{c \times g_v}, \\ g_q = g_k. \end{cases} \tag{1}$$

A softmax function is used to calculate the weights β_{ij} and is represented by the following equation:

$$\beta_{ij} = \Gamma\left(\frac{\beta'_{ij}}{\sqrt{g_k}}\right) = \frac{exp\left(\frac{\beta'_{ij}}{\sqrt{g_k}}\right)}{\sum_j exp\left(\frac{\beta'_{ij}}{\sqrt{g_k}}\right)}. \tag{2}$$

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

where Γ represents the softmax function and β'_{ij} computes the contribution of the *jth* input element to the *ith* output element. Throughout this process, β'_{ij} is considered to be the attention attributed to the factor v_i. As a result, the final resultant attention can be calculated as a weighted total of each and every value, as shown below:

$$y_i = \sum_j \beta_{ij} \times v_j. \tag{4}$$

In addition, it is reasonable to extend element-wise self-attention into matrices. However, for each input x_i, parallel matrix computation is commonly used to produce and create the query q_i, key k_i, and value v_i, respectively. Matrices can be formed by stacking up the input x_i, value v_i, query q_i, and key k_i, accordingly. Let $X \in \mathbb{R}^{n \times c}$ be the input matrix, and the value, query, and key matrices be V, Q, and K, respectively. The number

of samples is represented by n, and each individual matrix is made up of the components (i.e., $X = [x_1; x_2; x_3; \cdots ; x_n]^T$). Therefore, the attention matrix A and resultant matrix Y are now computed as shown below:

$$A = \Gamma\left(\frac{Q \times K^T}{\sqrt{g_k}}\right) \in \mathbb{R}^{n \times n}, \tag{5}$$

$$Y = A \times V \in \mathbb{R}^{n \times g_v}, \tag{6}$$

The authors [7] created a different form of attention mechanism known as multi-head self-attention (MHSA). They demonstrated that applying several self-attentions to the same input allows for a more efficient acquisition of hierarchical information. However, in the mechanism, h (*i.e.*, $h = 8$) distinct attention heads were generated, each with a unique set of weight matrices ($W(Q)$, $W(K)$, and $W(V)$). The key, value, and query matrices are then created for each attention head by multiplying the input matrix by each of the weight matrices (W^Q, W^K, and W^V). Again, these query, key, and value matrices are subjected to attention mechanisms in order to produce an output matrix from each attention head. In addition, the output of the MHSA layer is produced by concatenating the output matrix acquired from each attention head (h) and the dot product with the weight (W^O). Finally, given self-attentions (heads) denoted as h, the system produces the desired output result by integrating the computed attentions as illustrated in the equation below:

$$Y_i = A(Q \times W_i^Q, K \times W_i^K, V \times W_i^V), \tag{7}$$

$$M_H(Q, K, V) = f_C(Y_1, Y_2, Y_3, \cdots, Y_h)W^O, \tag{8}$$

where M_H denotes the multi-head self-attention operator and f_C is the concatenating function. The linear projection matrices W_i^Q, W_i^K, and W_i^V map the Q, K, and V matrices into the appropriate subspaces.

Transformer architecture

Transformers are generally designed to handle sequence-related tasks while also dealing with long-term dependencies. In the paper titled "Attention Is All You Need" [7], the authors introduced a standard transformer architecture that employs an encoder–decoder formation, as shown in Fig. 4, which will be discussed further in the subsequent sections. In the architecture, the encoder framework converts an input sequence $(x_1, x_2, x_3, ..., x_n)$ into a series of continuous representations (i.e., an output sequence) z $= (z_1, z_2, z_3, ..., z_n)$. The decoder then produces the resultant sequence $(y_1, y_2, y_3, ..., y_m)$ one component at a time from the encoded representation z, using the previous output as additional input when generating the next. The transformer follows this general architectural framework, which employs different layers in both the encoder and decoder modules, as demonstrated on the left and right sides of Fig. 4.

(i) Transformer encoder

Transformer architectures, as shown in Fig. 4 mainly consist of both encoder and decoder blocks. The encoder is composed of $N = 6$ identical layers built on top of one

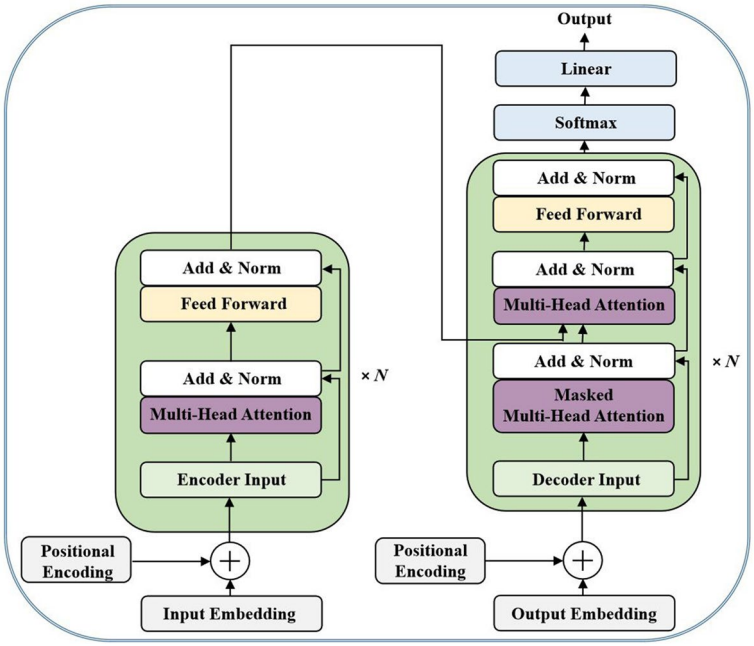

Fig. 4 A schematic demonstration of a standard transformer architecture

another that extract features from the input sequence. Each layer is made up of two sub-layers known as the feed-forward network layer (FFNL) and the multi-head self-attention mechanism (MHSA). Again, residual connections were employed across each of the sub-layers, followed by layer normalization. First, the multi-head attention is computed in each block, followed by a layer-wise normalization block. The sum of the multi-head attention input and output is computed primarily using layer-wise normalization. After applying a feed-forward layer, the input and output of the feed-forward layer are summed together using layer-wise normalization.

(ii) Transformer decoder

The transformer decoder shown on the right-hand side of Fig. 4 uses the extracted features to generate the output sequence. It consists of $N = 6$ identical layers with a few modifications. An additional sub-layer block is added on top of the encoded output, which carries out multi-head attention over the encoder stack output. Since the prediction is based on a known state, masking was utilized in the first self-attention block to prevent further contributions to the state of the preceding position. In addition, after the decoder's output layer, a linear and a softmax layer are added to produce the final result.

Vision transformer (ViT)

Transformers were initially introduced in NLP tasks where the objective was to understand the text and draw relevant and useful conclusions. Transformer architectures have accomplished significant results and have become a de facto standard in the field of NLP because of their generalization abilities and simplicity. Following their success in NLP tasks, researchers in this domain have made numerous attempts to adapt transformer architectures to various vision challenges. Among the most common transformer-based

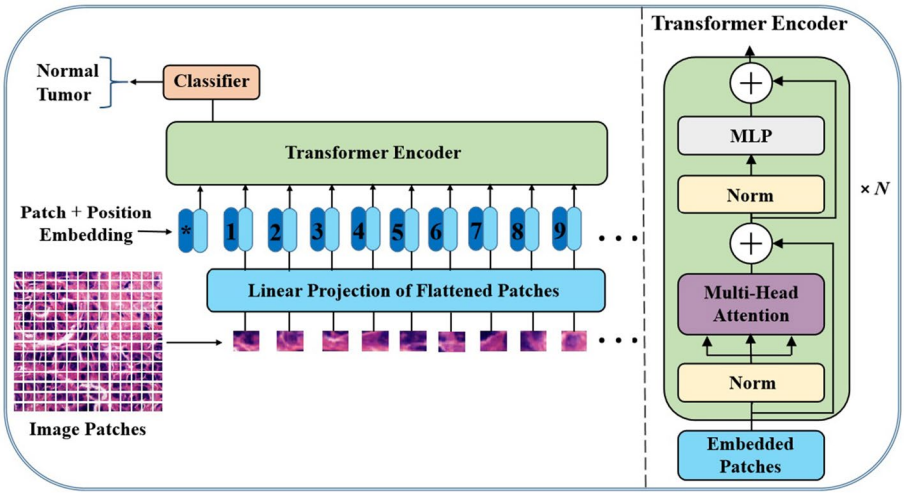

Fig. 5 A schematic diagram of a standard ViT model. Sequential image patches are used as the input, which is then processed with a transformer encoder and uses an MLP head module to generate a class prediction

architectures in vision that have been established are the DETR [33], Swin-transformer [34], ViT [8], DeiT [35], and BEiT [36]. In [33], the authors were the first to make use of transformers in computer vision for object detection tasks. The proposed architecture, known as DETR, focuses on a transformer encoder–decoder architecture and a set-based global loss that forces unique predictions through bipartite matching. Unlike other traditional object detection approaches that rely heavily on handcrafted techniques, the DEtection transformer does not need any special layers, which makes it easy to replicate in any model that has common transformer and CNN classes. On the other hand, it is simple to generalize and create unified panoptic segmentation. In 2021, Dosovitskiy et al. [8] introduced a vision-based transformer known colloquially as ViT, stating that CNNs were no longer required and that a pure transformer architecture applied instantly to sequences of image patches can produce robust results, particularly on image classification problems. The input image, as presented in Fig. 5, is split into a number of patches, each of which is encoded spatially to provide spatial information using a positional encoding technique. The ViTs have produced better or even higher results, outperforming state-of-the art (SOTA) CNNs for many downstream tasks, especially when pre-trained on huge datasets. To this end, transformer architectures require more training data to obtain comparable results or even higher than CNNs, and more details will be provided in subsequent sections.

Pros and cons of a transformer architecture

Transformers have been widely used in many computer vision challenges and have shown the capability of producing better results than other deep learning techniques. Some of the advantages of transformers in computer vision tasks include efficient parallel processing, adaptability with variable-length sequences, effective handling of global dependencies, higher network capacity, and so on. Due to the attention mechanisms incorporated into the networks, they can process sequences in parallel and also handle global dependencies, making them more efficient and faster than standard sequential

networks such as recurrent neural networks. In addition, transformer architectures also produce robust results on NLP tasks due to higher network capacity and the ability to capture complicated relationships in sequential data. Despite the fact that transformers can enable higher network capacity and learn more complex relations in the image data, they also have some drawbacks. Some of the disadvantages of transformers in computer vision tasks include high computing costs, overfitting vulnerability, data inefficiencies, and so on. Transformers are more resource-intensive than any other deep learning technique due to the self-attention mechanism built into the networks, which necessitates a lot of computation as well as training time. Furthermore, insufficient data to train the model effectively is another notable disadvantage of transformers, which can pose a lot of problems in NLP tasks where there is a limited amount of labeled data.

Transformer methods versus CNN methods

Over the years, CNNs have shown outstanding performances for histopathological image analysis, while transformers such as ViTs have produced better or even higher results, outperforming SOTA CNNs for many downstream tasks, especially when pretrained on huge datasets. CNN architectures are more mature and make use of pixel arrays, so they are easier to implement, study, and train when compared to transformer architectures. During training, as the depth of the networks increases, the receptive field of CNNs significantly widens; therefore, the features mined at lower stages differ significantly from those at later stages. Besides, CNNs make use of convolution, a "local" technique limited to a tiny area of an image, which makes them more advantageous in capturing local semantic structures. The feature maps created by the CNNs through the convolution process using these trainable convolutional filters, which are hidden representations of the true image, only affect a tiny portion of the image at a time. Additionally, CNNs are also limited in capturing long-distance correlations between image regions due to their small receptive field. On the other hand, transformer architectures make use of a self-attention mechanism, a "global" technique since it gathers relevant information from the entire image. This enables them to effectively capture more distant and important information in an image. The representation in transformer architectures is similar in every layer and can gather global information early owing to self-attention. Again, the MHSA in particular provides a global receptive field, which results in identical representations in distinct numbers of layers. Moreover, all attention outputs are linearly concatenated to the appropriate dimensions by the MHSA layer, and the block of each layer of the MHSA has the capacity of aggregating features globally to produce accurate knowledge of long-distance interactions. To this end, transformer architectures require more training data to obtain comparable results or even higher than CNNs, and more details will be provided in subsequent sections.

Current progress

Vision transformers (ViTs) have been generally used for a variety of clinical purposes. However, in this section, we will first discuss the searching procedures used to obtain all the papers reviewed in this survey (see "Article searching and selection procedures" Sect.). Then, we will present and discuss different ways of employing transformers for histopathological imaging in "Different ways of employing Transformers for

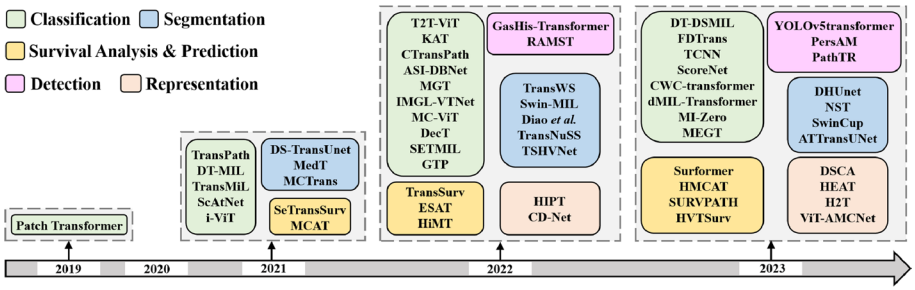

Fig. 6 Transformer-based architectures for histopathological image analysis. The figure shows some of the existing approaches for different downstream tasks, including segmentation, survival analysis and prediction, representation, detection, and classification

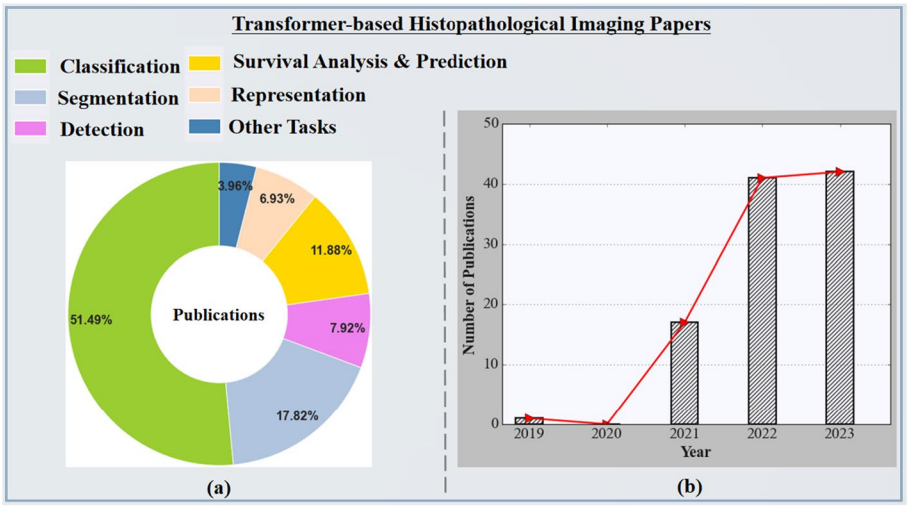

Fig. 7 The chart **a** displays the statistics of the papers presented in this survey according to histopathological imaging problem settings. The rightmost figure **b** demonstrates consistent growth in recent development (from 2019 to July 2023)

histopathological imaging" Sect. Finally, the current transformer applications in histopathological image analysis, as shown in Fig. 2, are discussed in " Current transformer applications in histopathological imaging" Sect.

Article searching and selection procedures

This section presents a brief discussion of the methods used in searching for and selecting the research papers. The newly built architectures, as shown in Fig. 6, are classified based on their learning tasks.

As demonstrated in Fig. 7(b), the histopathological imaging domain has been slightly impacted by transformer-based architectures since the inception of the first ViT architecture. Figure 7(a) displays the statistics of the papers presented in this survey according to histopathological imaging problem settings. In particular, we explore publications from Science Direct, Springer, Xplore, PubMed, IEEE, and conference proceeding papers, especially those from conferences on medical imaging like SPIE, RSNA, IPMI, MICCAI, ISBI, and so on. In addition, we use Google Scholar to search for paper references and manuscripts. As a result of our search queries using various keywords such as

vision transformers, transformers in medical imaging, transformers in histopathological imaging, transformers in image classification and segmentation, and so on, we found more than a thousand papers about the transformer, some of which are from the fields of natural imaging or language studies. Again, we construct the concepts of our survey from the self-attention and the ViT published papers, which are major milestones for the investigation of transformers in histopathological image analysis. Finally, we limited the survey research to exclusively cover transformer applications in the histopathological imaging domain. As presented in Fig. 6, we show the categorization of some recently developed models based on the learning tasks in the histopathological imaging field. Then, in Fig. 7, we show the percentage of the papers presented in this survey according to histopathological imaging problem settings and consistent growth in recent development, which will be further discussed in the following subsections.

Different ways of employing transformers for histopathological imaging

Recently, numerous studies have been conducted on how to apply transformers for histopathological image analysis. Some studies attempted to use only pure transformers (i.e., transformers without convolution blocks (see Fig. 5), while others tried to integrate the benefits of transformers (e.g., DETR [33], ViT [8], DeiT [35], BEiT [36], Swin-transformer [34], and so on) and CNNs (e.g., EfficientNet [37], Unet [38], ResNet [39], and so on) for different downstream tasks. However, in this section, we will classify them into three distinct types, which will be further discussed in the following subsections.

(i). Pure transformers: Pure transformers, as shown in Fig. 5, are described as those ViT-based architectures that resemble the ones originally proposed by Dosovitskiy et al. [8] which typically do not include major structural adjustments. They outperform conventional CNN models in terms of scalability and efficiency at both small and large computational sizes. TransWS [40], MCAT [26], HIPT [22], PyT2T-ViT [41], and ViT-WSI [17] are some examples of pure transformer models developed for different histopathological imaging tasks.

(ii). Graph-based transformer methods: These are the types of transformer networks that introduce graphs into traditional vision transformers (see Fig. 9(a) GTP). Moreover, graphs are a common type of data structure, and there are several areas of application in which datasets can be characterized as graphs, such as biological networks, social networks, and several other types of multimedia domain-specific data. However, using graph-based learning methods is a normal practice in both histopathological and other medical image analysis. As a result, analyzing graph data can reveal important information about node classification, and the basic idea behind graph learning is to use the data graph to learn a dense representation of each and every sample, such as embeddings, while maintaining the intrinsic inter-sample relationships. Transformer, as an attention-based model, is capable of processing graph data, including aggregating node information and determining the relationship between the nodes. Dwivedi et al. [42] developed a graph transformer network (GTN) that supports the use of specific domain information as edge features and provides interpretability via self-attention modules that locate the key regions of the graphs for prediction. AMIGO [43], LA-MIL [44], Wang et al. [45], and GTP [46], MEGT [47] are some examples of graph-based transformer models that have been proposed for different histopathological image classification tasks.

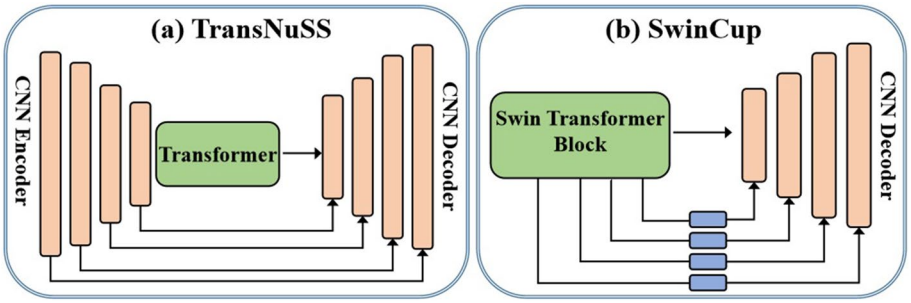

Fig. 8 Some typical transformer U-shaped architectures

(iii). Hybrid transformer–CNN: In histopathological image analysis, there are many ways in which transformers can be combined with CNN to form a hybrid model. The simplest method is to use both in an effort to capitalize on both of their advantages. These hybrid networks either use transformer to replace some parts of the network or incorporate transformer into the entire network by using CNN as the backbone of the network. However, we find out that current research in histopathological image segmentation focuses mainly on the following three issues to develop transformers combined with the widely used U-shaped framework: firstly, transformer blocks are inserted at various positions in the U-shaped structure, as shown in Fig. 8. Secondly, employing several techniques to combine CNN and transformer networks. Finally, making use of attention mechanisms or employing multi-scale features. SeTranSurv [25], ATTransU-Net [2], TCNN [1], SwinCup [48], and DHUnet [49], TransNuSS [50] are some examples of hybrid transformer–CNN models that have been developed for different histopathological imaging tasks.

Current transformer applications in histopathological imaging

This section presents the current applications of transformers in histopathological image analysis, such as classification, segmentation, survival analysis and prediction, representation, detection and localization, and other tasks. These applications, as demonstrated in Fig. 2, are classified based on their learning tasks.

Histopathological image classification

Vision transformer [8] has demonstrated remarkable performance in several natural image classification tasks since its inception. From previous and past studies, transformer-based techniques for cancer investigation and prediction are often referred to as classification tasks and can be classified into three distinct classes. Firstly, the direct application of transformer architectures to histopathological images. Secondly, making use of transformer architectures in conjunction with convolutions to learn more representative local features. Finally, making use of transformer architectures in conjunction with graph representations will help better manage data with complex sizes. This section, as demonstrated in Table. 1, will provide a thorough overview of current transformer applications for histopathological image classification. Figure 9 shows some examples of SOTA transformer architectures developed for histopathological image classification.

Table 1 Transformer applications in histopathological image classification tasks

Method	Tissue	Dataset	Challenge	Highlight	ACC / F1 / AUC (%)
ScoreNet [16]	Breast	BRACS, BACH and CAMELYON16	The huge size of WSIs and the cost of exhaustive localized annotations	Efficient transformer-based architecture local and global attention mechanism	–/ 81.10 /–
BreaST-Net [51]	Breast	BreakHis	Differentiating subtypes of benign and malignant cancers	Ensemble of Swin transformers	99.60 / 99.50 / 99.40
HATNet [52]	Breast	Custom	Diagnostic variability and misdiagnosis of breast cancer	End-to-end ViTs with self-attention mechanism	71.00 / 70.00 /–
dMIL-transformer *et al.* [53]	Breast (LNM)	CAMELYON16 and 17 and the SLN-Breast	Taking into account the morphology and spatial distribution of cancerous regions	Two-stage double max–min MIL transformer architecture	89.23 / 84.83 / 91.67
ASI-DBNet [54]	Brain	UHP	Lack of precision and accuracy in grading brain tumor	An adaptive sparse interactive ResNet ViT dual network	95.24 / 95.23 / 96.83
Ding *et al.* [55]	Brain	NCT-CRC-HE, BreaKHis and LDCH	Aliasing phenomena caused by downsampling operations and smoothing discontinuous	ViT-based network with wavelet position embedding	99.01 /–/–
DT-DSMIL [56]	Colorectal	Custom	Data annotations	Weakly supervised ViT-based MIL	93.50 / 94.37 / 97.69
IMGL-VTNet [57]	Gastric	IMGL	The problem of identifying IM glands	Multi-scale deformable transformer	–/ 94.00 /–
tRNAsformer [58]	Kidney	TCGA	Gather the information needed to learn WSI representations	Transformer-based learning to predict RNA sequence expressions	96.25 / 96.25 /–
i-ViT [59]	Kidney	TCGA-KIRP	Capturing cellular and cell-layer level patterns	Instance-based Vision Transformer network	93.01 / 93.60 /–
GTP [46]	Lung	CPTAC, TCGA and NLST	Label noise	Graph-transformer with vision transformer	91.20 /–/ 97.70
FDTrans [60]	Lung	TCGA-NSCLC	Large intra-class differences and a lack of annotated datasets	Frequency domain transformer-based architecture	92.33 / 94.64 / 93.16
Yacob *et al.* [45]	Skin	Custom	Time-consuming and inter-pathologist variability	Weakly supervised approach using graph-transformer	93.50 /–/–
KAT [61]	Stomach	Gastric-2K, Endometrial-2K	Over-smoothing and High computational complexity	Kernel attention transformer	94.9 /–/ 98.30
DT-MIL [62]	Lung and breast	CPTAC-LUAD and BREAST-LNM	The problem of learning an effective WSI representation	Deformable transformer model for MIL	–/ 96.92 / 99.06

Table 1 (continued)

Method	Tissue	Dataset	Challenge	Highlight	ACC / F1 / AUC (%)
TCNN [1]	Breast, Lung, etc.	MDD and RWD	Artifacts in WSIs	Transformer with CNN	96.90 / 97.40 / 98.50
CWC-transformer [63]	Breast and Lung	CAMELYON16, TCGA-LUNG and MSK	Loss of spatial information and problems associated with feature extraction in WSI	Combination of transformer and CNN	92.59 /–/ 94.88
TransPath [64]	Breast, Lung, etc.	TCGA, PAIP, PatchCam, etc.	Data annotation	Self-supervised learning transformer-based network	95.85 / 95.82 / 97.79
TransMIL [65]	Breast, Lung and Kidney	CAMELYON16, TCGA (NSCLC and RCC)	Correlation among different instances, Huge size and the lack of pixel-level annotations	Transformer-based multiple-instance learning (MIL)	94.66 /–/ 98.82
DecT [66]	Breast, Endometrium	BreakHis, BACH, and UC	Not taking into account the staining properties of histopathological images	Color deconvolution with transformer architecture	93.02 / 93.89 /–
LA-MIL [44]	Colorectal and stomach	TCGA-CRC and TCGA-STAD	Quadratic complexity of transformer architectures with respect to the sequence length	MIL local attention graph-based transformer model	–
Prompt-MIL [67]	Breast and colorectal	TCGA(BRCA and CRC and BRIGHT	Overfitting problems and a lack of annotated data	Prompt Tuning MIL transformer	93.47 /–/–
HAG-MIL [68]	Breast, Gastric, Lung, etc.	CAMELYON16, IMGC, TCGA-RCC and NSCLC	The difficulties in locating the most discriminative patches	Hierarchical attention-guided MIL transformer framework	91.40 / 89.40 / 98.20
MI-Zero [69]	Breast, cell, and lung	TCGA (BRCA, NSCLC and RCC), etc.	Computational issues and a scarcity of large-scale publicly available datasets	Transformer-based visual language pre-trained MI zero-shot transfer	70.20 /–/ –
HAG-MIL [69]	Breast, cell, and lung	TCGA (BRCA, NSCLC and RCC), etc.	Computational issues and a scarcity of large-scale publicly available datasets	Transformer-based visual language pre-trained MI zero-shot transfer	70.20 /–/ –
MEGT [47]	Kidney and breast	TCGA-RCC and CAMELYON16	The problem of learning multi-scale image representation from large images like gigapixel WSIs	Multi-scale efficient graph transformer-based network	96.91 / 96.26 / 97.30
MSPT [70]	Breast, and lung	TCGA-NSCLC and CAMELYON16	The problem of uneven representation between the negative and positive instances in bags	Multi-scale prototypical transformer-based network	95.36 /–/ 98.69
GLAMIL [71]	Breast, lung, and kidney	TCGA(RCC and NSCLC) and CAMELYON16	Overfitting, WSI-level feature aggregation, and imbalanced data challenges	Local-to-global spatial learning	95.01 /–/ 99.26

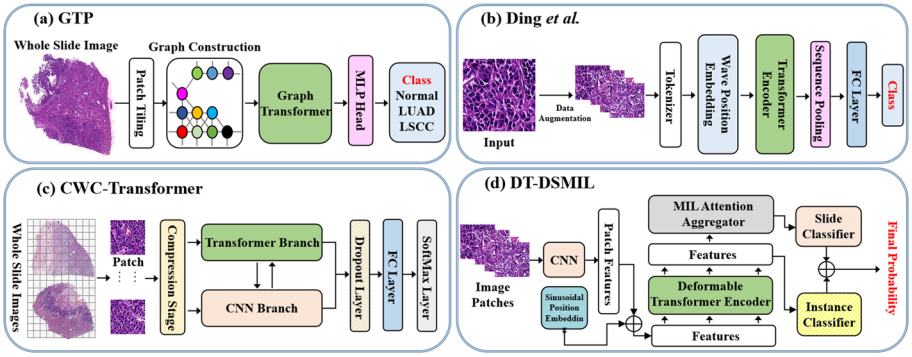

Fig. 9 Some examples of SOTA transformer architectures for histopathological image classification

For breast cancer histopathological image classification, DCET-Net [72] proposed a dual-stream convolution-expanded transformer architecture; Breast-Net [51] explores the ability of ensemble learning techniques using four Swin transformer architectures; HATNet [52] uses end-to-end vision transformers with a self-attention mechanism; ScoreNet [16] developed an efficient transformer-based architecture that integrates a coarse-grained global attention framework with a fine-grained local attention mechanism framework; LGVIT [73] built a local–global ViT model by introducing a new local–global MHSA mechanism and a ghost geed-forward network block into the network; dMIL-transformer [53] developed a two-stage double max–min multiple-instance learning (MIL) transformer architecture that combines both the spatial and morphological information of the cancer regions. Other than breast cancer classification, transformers have also been applied to other histopathological image cancer classification tasks, such as bone cancer classification (NRCA-FCFL [74]), brain cancer classification (ViT-WSI [17], ASI-DBNet [54], Ding et al. [55]), colorectal cancer classification (MIST [75], DT-DSMIL [56]), gastric cancer classification (IMGL-VTNet [57]), kidney subtype classification (i-ViT [59], tRNAsformer [58]), thymoma or thymic carcinoma classification (MC-ViT [76]), lung cancer classification (GTP [46], FDTrans [60]), skin cancer classification (Wang et al. [45]), and thyroid cancer classification (Wang et al. [77], PyT2T-ViT [41], Wang et al. [78]) using different transformer-based architectures. Furthermore, other transformer models such as Transmil [65], KAT [61], ViT-based unsupervised contrastive learning architecture [79], DecT [66], StoHisNet [80], CWC-transformer [63], LA-MIL [44], SETMIL [81], Prompt-MIL [67], GLAMIL [67], MaskHIT [82], HAG-MIL [68], MEGT [47], MSPT [70], and HistPathGPT [69] have also been evaluated on more than one tissue type, such as liver, prostate, breast, brain, gastric, kidney, lung, colorectal, and so on, for histopathological image classification using different transformer approaches. As shown in Fig. 9, GTP [46] introduced a graph-based transformer architecture that combines a vision transformer for processing histopathological images and a graph-based representation of a WSI for disease grade prediction. Ding et al. [55] built an improved ViT-based architecture by introducing a wavelet position embedding framework into the network to reduce the aliasing phenomenon in histopathological features brought about by smooth discontinuous feature information and downsampling operations. CWC-transformer [63] presents a two-stage network module that successfully addresses the feature extraction and spatial information loss problems in classifying

WSIs. DT-DSMIL [56] proposed a weakly supervised transformer architecture that is based on MIL to do away with the time-consuming and labor-intensive manual annotations and also to handle gigapixel images at once.

To this end, structural improvements, newly built transformer architectures, CNN backbones, pre-training, multiple-instance learning, and ensembling learning techniques are among the numerous innovations included in these transformer architectures for a wide range of tasks. As listed in Table 1, even though pure transformers, transformers with graphs, and hybrid transformers perform exceptionally well in a number of papers surveyed, such as breast and lung cancer classification, further improvement is still required in future research. On the whole, we therefore summarize the transformer applications for histopathological image classification as follows: firstly, transformer architectures have obtained equal or superior results in many classification tasks in comparison with CNN-based models. Secondly, transformer architectures are somewhat limited in their application, particularly in the field of histopathological imaging, because of their desire for extensive annotated datasets. However, an alternative approach to resolving this challenge could be pre-training. Thirdly, it is computationally expensive to train transformer models using gigapixel images. Therefore, in order to boost their performance, it is crucial to lower the computational cost of the model and create lightweight architectures. Fourthly, most of the current transformer-based architectures focus on 2D histopathological imaging. With the increasing application of transformers in histopathological image classification and prediction, we believe that more work will be put towards building 3D transformer models. Finally, the increasing popularity of hybrid transformers has recently gathered so much attention, as they have gained from both sides of transformers and conventional networks such as CNN and GNN.

Histopathological image segmentation

Semantic segmentation of tumor regions is a crucial task in histopathological image analysis. During segmentation, a region of a whole slide image (WSI) is used as input, and the model then segments the region using predetermined features. Despite recent developments in deep learning over the years, it was still a crucial and difficult task for researchers to segment the region of interest or cancerous region of histopathological images until the advent of vision transformers. Nowadays, transformer-based approaches have been used to solve a number of segmentation challenges, such as colon cancer segmentation [83], multi-organ nucleus segmentation [2], and nuclei segmentation [15, 50, 84, 85]. Some outstanding SOTA works are tabulated and detailed in Table 2, along with their associated network type, tissue type, dataset, challenge, highlight, etc. Figure 10 shows some examples of SOTA transformer architectures developed for histopathological image segmentation.

The U-shaped CNN-based methods, often known as UNet [38], have obtained remarkable success in a number of histopathological image segmentation challenges. Besides, UNets are constrained in modeling long-term dependencies because of the convolutional layers present in them. Hence, in order to solve this challenge, researchers have made tremendous efforts over the years to develop high-performance hybrid transformers integrated with the UNet backbone. One of the most logical ways of inserting a

Table 2 Transformer applications in histopathological image segmentation

Method	Tissue	Dataset	Challenge	Highlight	DSC / IoU / F1 (%)
Swin-MIL [83]	Intestine	Custom	Image annotation and lack of related information between instances	Transformer-based weakly supervised approach	–/–/ 99.90
MCTrans [84]	Cell	Pannuke	Inability of CNN-based methods to model long-term dependencies	Multi-compound transformer with CNN	68.90/– /–
TSHVNet [85]	Cell	CoNSeP and Pannuke	Difficulties in differentiating various classes of nuclei and separating nuclear instances with high clustering,	Integration of multiattention modules (transformer and SimAM)	85.6 /– /82.00
Diao *et al.* [86]	Colon	NPC2020	Insufficient global context encoding	Transformer-based network using TransUNet	83.30/73.00 /–
DS-TransU-Net [15]	Colon	GlaS	Ignoring the pixel-level intrinsic structural features inside each patch	Dual Swin transformer U-Net with standard U-shaped arch	87.19/78.45/–
TransAttU-net [87]	Colon	GlaS	Modeling long-range contextual dependencies and Computational costs	Transformer with Multi-level Attention-guided U-Net	89.11 / 81.13 /–
ATTransUNet [2]	Colon	GlaS and MoNuSeg	Heavy computational burden of paired attention modeling between redundant visual tokens	A transformer-enhanced hybrid architecture based on the adaptive token	89.63 / 82.55 /–
HiTrans [88]	Liver	PAIP 2019	The inherent heterogeneity of hepatocellular carcinoma	A hierarchical transformer encoder-based network	–/ / 75.13
TransWS [40]	Colon and breast	GlaS and Camelyon16	highlighting target regions roughly, sub-optimal solution and low efficiency	Transformer-based weakly supervised learning	– /–/ 85.20
TransNuSS [50]	Colon and breast	TNBC and MoNuSeg	The challenges of pre-training nuclei segmentation models with ImageNet due to morphological and textural differences	Self-supervised learning incorporated with vision transformer model	83.07 / 68.72 /–
NST [89]	Liver, Breast, Colon, etc.	GCNS and MoNuSAC 2020	The staining of WSI sections is not uniform and nuclei having different sizes and shapes	A gastrointestinal transformer-based network	79.60 / 66.30 /–

Table 2 (continued)

Method	Tissue	Dataset	Challenge	Highlight	DSC / IoU / F1 (%)
MedT [90]	Colon and cell	GlaS and MoNuSeg	Inherent inductive biases in CNNs and insufficiently annotated datasets	Gated axial-attention transformer-based model	–/ 69.61 / 81.02
SwinCup [48]	Colon and colorectal	GlaS	Inability of CNNs to model global context	Cascaded Swin transformer-based network	–/–/ 92.00
DHUnet [49]	Breast, liver, and lung	BCSS, WSSS-4LUAD, etc.	Inability of the transformer model to capture fine-grained details in pathological images	Dual-branch hierarchical global–local fusion network	93.07 / 87.04 /–

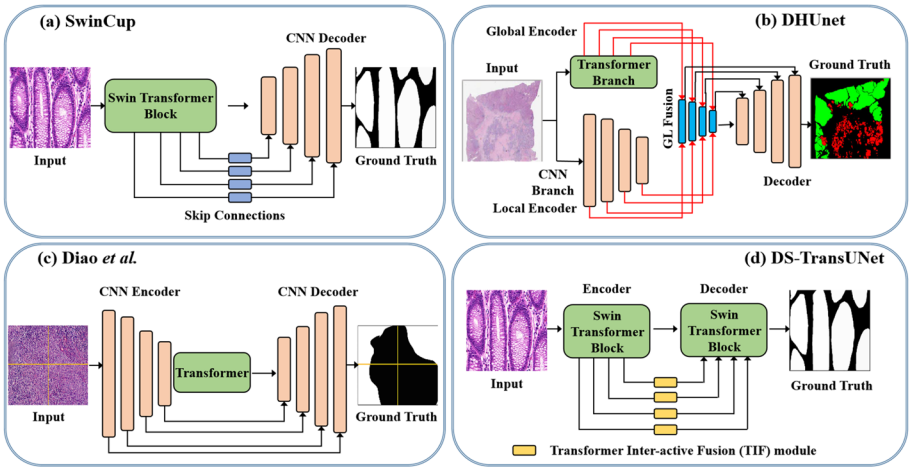

Fig. 10 Some examples of SOTA transformer architectures for histopathological image segmentation

transformer block into the U-shaped network is to place the entire transformer architecture between the encoder and decoder blocks so as to create long-range dependencies between high-level vision generalizations, as shown in Fig. 10. Some studies place the entire transformer architecture in the encoder part, while others place it in the decoder part. Methods such as TransNuSS [50], SwinCup [48], Diao et al. [86], DS-TransUNet [15], HiTrans [88], and DHUnet [49] [see Fig. 10(b)] are some examples of transformer-based U-shaped networks developed for histopathological image segmentation. In contrast to the various approaches mentioned above that incorporate transformer and U-shaped architectures within a single inference pathway, other studies looked into new ways of bridging transformers and CNNs for more accurate and robust segmentation. Although transformer-based architectures demonstrate the superiority of modeling long-range contextual information, their inability to capture local features still poses a lot of problems. Rather than cascading the transformer and convolution blocks, many studies recommend using the vision transformer and CNN as encoders that both accept histopathological images as input. After that, the embedded features are combined to link with the decoder. This approach benefits from simultaneously learning local and

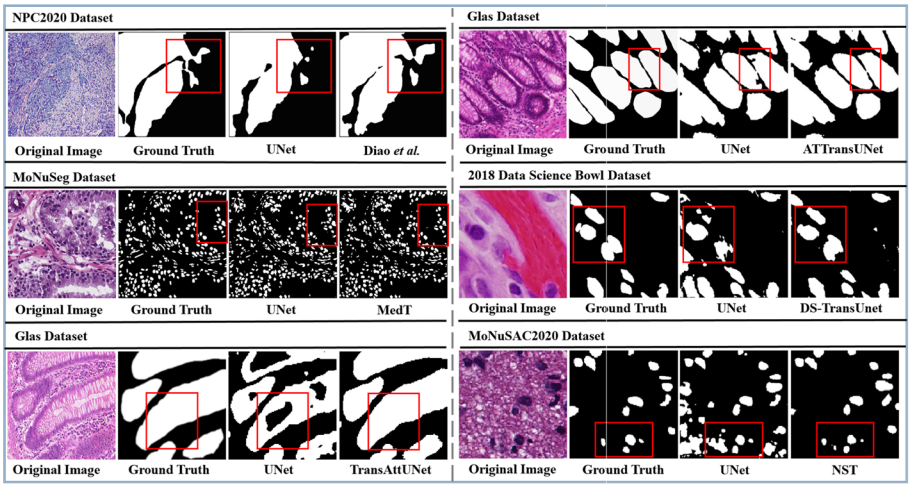

Fig. 11 Examples of segmentation results of popular Unet architecture [38] and transformer-based models (ATTransUNet [2], DS-TransUNet [15], MedT [90], Diao et al. [86], TransAttUnet [87], and NST [89]) on different histopathological datasets

global information and then stacking representations sequentially [89]. Other than using U-shaped transformer-based architectures, some methods, such as MCTrans [84], TransAttUnet [87], MedT [90] and TSHVNet [85] applied multi-scaling techniques for histopathological image segmentation. In addition, pure transformer-based architectures can also be applied to a variety of histopathological image segmentation tasks. With the exception of the UNet network variations already mentioned, using the transformer in conjunction with convolution blocks, TransWS [40] introduced a transformer-based weakly supervised learning method without convolution layers. The proposed approach was basically used to address the issues of low efficiency and sub-optimal solutions as well as the challenge of producing a high-quality class activation map that identifies the precise and integral target, leading to insufficient activation and undefined boundaries. Qian et al. [83] built a weakly supervised approach that inserts the transformer architecture into the MIL module to encode long-term or global dependencies. Figure 11 shows some visual segmentation results obtained from various transformer networks against the popular Unet architecture on different histopathological image segmentation datasets.

In summary, from the research papers surveyed in this section, we can conclude that the histopathological image segmentation domain has been slightly impacted by transformer-based architectures since the inception of the first ViT architecture, as shown in Table 2. In comparison to other medical imaging fields, we strongly believe that this is due to a lack of annotated histopathological segmentation datasets and the high computational cost of training WSIs. As stated above, the high computational cost involved with mining features at multiple intensities obstructs the applicability of multi-scale networks in histopathological image segmentation tasks. These multi-scale networks make use of processing input image information at several levels and obtain significantly better performance than single-scale networks. As a result, building efficient transformer-based models for multi-scale processing needs better attention. Besides, most of the recently developed transformer-based architectures are pre-trained mainly on the

ImageNet dataset for various downstream tasks. Hence, this technique is sub-optimal because of the huge domain gap between histopathological images and natural images. Recent ViT-based methods have largely focused on 2D histopathological image segmentation; therefore, building customized architectural frameworks by integrating temporal features for robust high-dimensional and high-resolution segmentation of WSI has not been fully investigated. Furthermore, with the development of ViT-based methods, we discovered that there is an urgent need to gather more varied and demanding histopathological image datasets. Although challenging and diverse datasets are also very important for evaluating the performance of transformers in other clinical settings, they are especially important for histopathological image segmentation because of the major influx of transformer-based approaches in this domain. To this end, we anticipate that these datasets will be crucial in determining the viability of ViT-based models for histopathological image segmentation.

Histopathological image detection and localization

The word "detection" has different meanings across many domains. As we mentioned earlier, it is frequently referred to as disease identification or diagnosis in clinical domains, whereas in the technical field, it simply refers to determining whether lesions or diseases are present. However, disease detection in histopathological images is often referred to as a technique for locating instances of diseases in a specific image and identifying the potential region of a tumor, such as mitosis detection from breast cancer images, and is generally an important aspect of disease identification. Disease diagnosis is one of the most challenging tasks for clinicians, so it is important to have a reliable CAD technique that can serve as a second observer and potentially speed up the diagnosis process. Following the success of CNN-based methods in histopathological image detection and localization, there have been a few attempts recently to improve performance using transformer-based architectures. These techniques are primarily based on the detection transformer (DETR) [33]. Transformer architectures used for detection tasks involving histopathological images often incorporate CNN blocks, where CNNs are mainly used to mine features from images while the transformers are used to improve the mined features for other subsequent tasks. A few outstanding SOTA works are tabulated and detailed in Table 3. Figure 12 shows some examples of SOTA transformer architectures developed for histopathological image detection.

Recently, Chen et al. [19] proposed a multi-scale ViT-based approach that makes use of a position-encoded ViT framework and a CNN with convolutional operation to mine global and local information. To tackle the large-scale context overflow challenges, Wenkang et al. [91] developed a novel transformer-based technique that integrates global and local context within an end-to-end module. In addition, Ali et al. [92] introduced a transformed-based CAD system by making use of deep CNN networks based on channel boosting techniques. Takagi et al. [18] proposed a ViT-based personalized attention mechanism network for gigapixel WSIs with clinical records. Liaqat et al. [95] developed a channel-boosted hybrid ViT-based network that makes use of transfer learning techniques to build boosted channels and uses both ViT and CNN models to analyze cancerous images. As shown in Fig. 12, RAMST [94] makes use of joint region attention and a multi-scale transformer network to alleviate the unstable predictions caused by

Table 3 Transformer applications in histopathological image detection and localization

Methods	Tissue	Dataset	Challenge	Highlight	ACC / F1 (%)
GasHis-transformer [19]	Stomach (Gastric)	HE-GHI-DS	Inability of CNN models to handle global information well	GasHis-transformer and LW-GasHis-transformer	97.97 / 97.97
PathTR [91]	Breast	CAMELYON16	Neglecting the intrinsic WSI global correlations among the patches	Context-Aware Memory ViT with a CNN Backbone	98.91 /–
PVTCB-Lymph-Det [92]	Colon, breast and prostate	LYSTO	Detecting lymphocytes automatically due to the presence of artifacts and morphological variations	Pyramid ViT-based network and convolution attention mechanism with ResNet-50	–/ 88.92
YOLOv5-transformer [93]	Breast, Colon, etc.	Custom	Accurate mitoses detection and morphological variations	Improved YOLOv5 transformer-based architecture	–/ 77.00
RAMST [94]	Stomach and colorectal	TCGA (CRC and STAD)	Unstable predictions caused by noisy patches and aggregation techniques	Joint regional attention and multi-scale transformer network	–
CB-HVTNet [95]	Colorectal, breast, etc.	LYSTO and NuClick	Insufficient feature representations	Channel-boosted hybrid ViT network	–/ 80.00
Hossain *et al.* [96]	Breast, etc.	TCGA and Custom	ViT-based network	ROI selection ViT-based network	96.10 /–
PersAM [18]	Lymph	Custom	Attention region estimation in digital pathological images	Personalized attention mechanism ViT network	83.13 /–

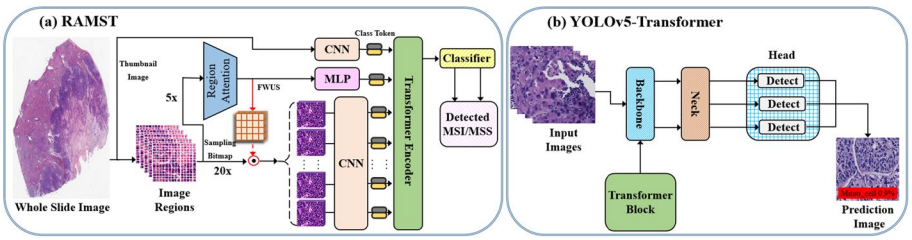

Fig. 12 Some examples of SOTA transformer architectures for histopathological image detection

noisy patches and aggregation techniques in WSIs. YOLOv5-transformer [93] built an improved transformer architecture that integrates transformer into the YOLOv5 model for mitoses detection. Hossain et al. [96], on the other hand, built a region of interest (ROI) selection ViT-based architecture to speed up the analysis of histopathological images and improve the detection accuracy of cancerous regions.

In summary, the number of new ViT-based architectures for the histopathological image detection and localization challenge task, as presented in Table 3 is lower than that of the classification task as reported in this survey paper. This is in comparison to the previous CNN-based methods that were promptly built for histopathological and

other clinical detection tasks. Some recent medical research papers demonstrate that the generic, class-agnostic detection system of multi-modal ViT-based models pre-trained on other images rather than medical images performs horribly on histopathological and other clinical datasets. Hence, evaluating the performance of multi-modal ViT-based architectures by pre-training them on modality-specific histopathological WSI datasets is a good research direction to investigate in the future.

Histopathological image survival analysis and prediction

Survival analysis and prediction is an arduous regression problem that aims to predict the time to an event, for example, the diagnosis of a disease or the relative risk of cancer death. Over the years, several techniques have been developed for survival analysis and prediction using histopathological WSIs. However, these techniques can be classified into two distinct classes: ROI-based and WSI-based approaches, respectively. Due to the high cost of computational resources, the majority of the existing literature has concentrated on regions of interest (tiles) chosen by pathologists from WSIs. Nowadays, a number of methods for histopathological image analysis have been proposed for a wide range of downstream tasks, using the detailed and dense annotations on WSIs. Recently, transformer-based architectures have demonstrated outstanding performance in predicting survival rates. A few outstanding SOTA works are summarized and detailed in Table 4. Fig. 13 shows some examples of SOTA transformer architectures developed for histopathological image survival analysis and prediction.

Transformer-based methods such as HiMT [100], MCAT [26], PG-TFNet [98], TransSurV [97], and SURVPATH [101] combine genomic data and histopathological images for survival analysis and prediction. As shown in Fig. 13, MCAT [26] introduced a multimodal co-attention Transformer network to learn an interpretable, dense co-attention mapping among genomic features and WSIs constructed in an embedding space. TransSurV [97] makes use of a Transformer-based multi-modal feature fusion network to extract useful predictive features from the multi-modal data. HiMT [100] introduced a hierarchical transformer-based network to mine the instant-level tile features at random from WSIs with varying magnification levels. AMIGO [3] created a multi-modal graph transformer architecture that predicts patient survival based on multi-modal histopathological images and shared related data. In addition, Huang et al. [25] designed a transformer technique for survival prediction based on the combination of tile features via an self-supervised learning (SSL) approach and a transformer. Shen et al. [99] make use of an explainable survival analysis framework coupled with a convolution-involved ViT-based network. More recently, Jaume et al. [101] introduced a memory-efficient multimodal-based transformer architecture that combines patch tokens and transcriptomics for patient survival prediction. Wang et al. [102] developed a pattern-perceptive survival transformer-based network that can statistically interpret the predictions as well as directly quantify the important histopathological patterns. HMCAT [104] introduced a hierarchical multi-modal co-attention Transformer-based network that addresses the challenges of the large size of histopathological WSIs and the significant disparity between the spatial scales of radiology images and histopathological WSIs. Shao et al. [103] make use of a hierarchical ViT-based architecture to completely investigate the contextual, spatial, and hierarchical relationships in the patient-level bag.

Table 4 Transformer applications in histopathological image survival analysis and prediction

Method	Tissue	Dataset	Challenge	Highlight	C-index (%)
TransSurv [97]	Colorectal	TCGA-CRC and NCT-CRC-HE	Inability of the previous models to extract useful predictive features from the multi-modal data	Transformer-based multi-modal feature fusion network	82.20
PG-TFNet [98]	Colorectal	TCGA-CRC	Inability to make use of the powerful representation learning capabilities of the neural networks	Transformer-based multi-modal feature fusion network	81.60
ESAT [99]	Lung	NLST and CHCAMS	Using a pre-selected subset of main patches or patch clusters as input instead of using the entire WSIs	Make use of the ViT backbone with convolution operations.	73.00
MCAT [26]	Bladder, Breast, Lung, Uterine	BLCA, UCEC, BRCA, BMLGG, LUAD	Computational complexity and large data heterogeneity gap between genomics and WSIs	Multimodal Co-Attention Transformer for Survival Prediction	65.30
HiMT [100]	Bladder, Breast, Lung, Brain, etc.	BLCA, BRCA, UCEC, LUAD, LGG, etc.	High computational cost of extracting patches from WSIs, which results in a large bag size	Hierarchical-based multi-modal Transformer framework	67.30
MaskHIT [82]	Breast, Lung, etc.	TCGA	Huge number of network parameters and insufficient labeled data	Masked pre-training of Transformers	61.20
SURVPATH [101]	Breast, Bladder, Stomach, etc.	TCGA	Capturing dense multimodal interactions between different modalities	Memory-efficient multimodal Transformer	62.90
Surformer [102]	Bladder, Breast, Lung, etc.	TCGA (BLCA, BRCA, LUAD, etc.)	Weak interpretability problems of the previous computational pathology model	Pattern-perceptive survival Transformer-based Network	68.70
HVTSurv [103]	Bladder, Breast, Lung, etc.	TCGA (BLCA, BRCA, LUAD, etc.)	The challenges of exploring contextual, spatial, and hierarchical interaction in the patient-level bag	Hierarchical ViT-based architecture	63.40
HMCAT [104]	Low Grade Glioma	TCGA-GBMLGG	The significant disparity between the spatial scales of radiology images and WSIs	Hierarchical multimodal co-attention transformer-based network	79.60
AMIGO [3]	Ovarian and bladder	InUIT and MIBC	ignoring specific details regarding the individual cells in a tile image	Sparse multi-modal graph Transformer-based network	61.00

Table 4 (continued)

Method	Tissue	Dataset	Challenge	Highlight	C-index (%)
SeTranSurv [25]	Breast, Lung, Ovarian	OV, LUSC, and BRCA	Ignoring the important role of spatial information in patches and the correlation between patches and WSIs	Integration of patch features through self-supervised learning and Transformer	70.50

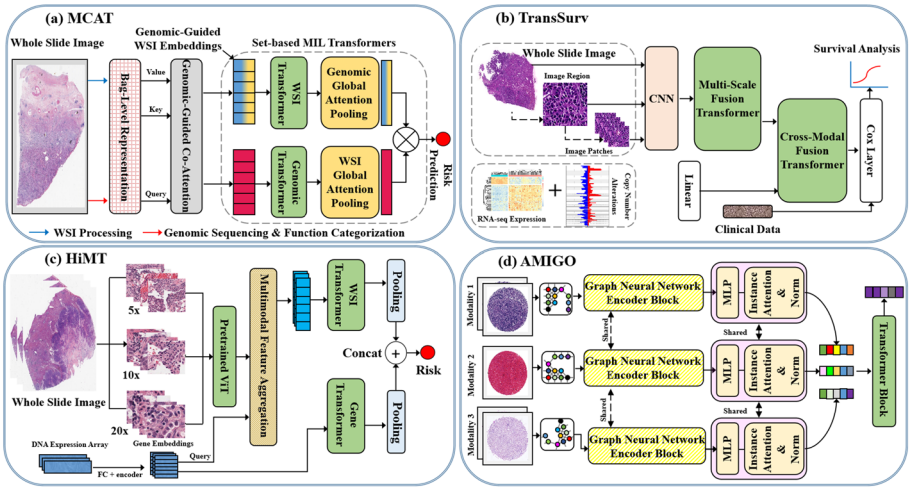

Fig. 13 Some examples of SOTA transformer architectures for histopathological image survival analysis and prediction

In summary, the number of new ViT-based architectures for the histopathological image survival and prediction task is lower compared to that of the classification tasks reported in this paper. This is in comparison to the previous CNN-based methods that were promptly built for histopathological and other clinical survival and prediction tasks. It is also important to note that despite the fact that there are several survey papers covering the applications of CNNs in histopathological image analysis [4–6], none of these studies have recently covered the use of Transformer architectures in survival analysis and prediction, despite the outstanding performance that these architectures have demonstrated over the last few years. We anticipate that this part will be a useful tool for researchers in this domain. In addition, we will briefly discuss some problems with transformer-based architectures for survival analysis and prediction below, along with some interesting future prospects.

As shown in Table. 4, transformer-based survival analysis and prediction architectures mostly rely on the concordance index metric (i.e., c-index) to evaluate the performance of the networks, which sometimes fails to accurately reflect clinical efficacy. Since the researchers currently depend only on the c-index metric as an evaluation metric, we believe that further effort is needed to develop more accurate clinical evaluation indicator to speed up the adoption of transformer-based survival analysis and prediction in clinical domains. Again, some of the transformer-based architectures

surveyed in this paper make use of histopathological images and genomic records for survival analysis and prediction. Hence, generating reports from other clinical or medical domains has its own challenges due to their unique nature and varied features. Besides, a few histopathological datasets, like TCGA ,[1] are available that consist of different cancer types together with clinical records. This dataset has the potential to be a valuable baseline for evaluating the performance of future multimodal transformer-based architectures for survival analysis and prediction. We suggest that in the future, transformer-based architectures tailored to particular tissues to predict patient survival should be investigated, with a focus on building challenging and varied datasets of different tissues.

Histopathological image representation

Due to memory and processing time constraints, histopathological images are often divided into smaller tiles (such as 256×256 pixels), and features are then mined concurrently from each tile. The representation of a histopathological WSI using information from multiple tiles, however, is a developing field of study with limited results that have been published, particularly in the context of clinical prediction and prognosis. Recently, several studies have been developed for learning multi-scale representations of images using transformer-based models, which can also be employed in convolutional pipelines in order to construct global representations of images. A few outstanding SOTA works are tabulated and summarized in Table 5. Figure 14 shows an example of the SOTA transformer architecture developed for histopathological image representation.

In order to learn high-resolution image representations from histopathological images, HIPT [22], made use of a ViT-based hierarchical image pyramid network, CD-Net [105] proposed a Transformer-based pyramidal context-detail network, and H2T [108] employed a handcrafted histological Transformer. As presented in Fig. 14, HIPT [22] uses two levels of self-supervised learning to take advantage of the natural hierarchical structure present in histopathological WSIs. The proposed architecture was pre-trained across 33 different cancer types by making use of 10,678 histopathological slides, 104 M 256×256 images, and 408,218 4096×4096 images. In addition, DSCA [106] built a dual-stream Transformer architecture with cross-attention to address the challenges of the unseen semantical disparity in multi-resolution feature fusion and the high computational complexity of histopathological WSI visual representation. ViT-AMCNet [20] makes use of an end-to-end transformer-based network with adaptive model fusion and a multi-objective optimization technique to address the challenges of poor interpretability and weak inductive bias ability for the laryngeal tumor grading task. Chan et al. [107] built a heterogeneous-graph edge attribute transformer-based network that can benefit from both node and edge heterogeneity.

In summary, since the number of publications and transformer applications in histopathology image representation is currently limited, as shown in Table. 5, it is challenging to draw any conclusions at this time. However, as the current transformer-based architectures give better results on histopathological image representation tasks, we anticipate further development in this domain in the near future.

[1] https://www.cancer.gov/tcga.

Table 5 Transformer applications in histopathological image representation

Method	Tissue	Dataset	Challenge	Highlight	ACC / AUC (%)
CD-Net [105]	Breast, Lung	TCGA (LUAD and LUSC)	Inability to leverage the rich multi-resolution information	Transformer-based pyramidal context-detail network	91.10 / 95.80
DSCA [106]	Lung, breast and brain	NLST, TCGA (BRCA and LGG)	High computational complexity and unnoticed semantic gap in multi-resolution feature fusion	A dual-stream Transformer network with cross-attention framework	-
HIPT [22]	Breast, lung, stomach, cell	IDC, LUAD, CCRCC, PRCC, CHRCC, and STAD	The structure of phenotypes in tumors and learning a good representation of a WSI	Hierarchical image pyramid Transformer with two levels of self-supervised learning	–/ 98.00
HEAT [107]	Colon, breast and esophageal	CAMELYON16, TCGA (COAD, BRCA, and ESCA)	The challenges of extracting diverse interactions between various cell types	Heterogeneous-graph edge attribute Transformer-based network	99.90 / 99.90
H2T [21]	Lung, breast and kidney	TCGA-NSCLC, CPTAC-LUAD, etc., BRCA, RCC and ACDC	High discordance on how a tissue sample and higher predictive power that comes at the cost of interpretability	Handcrafted histological Transformer-based network for unsupervised representation WSIs	-
ViT-AMCNet [20]	Laryngeal, breast, brain	Laryngeal cancer, breast cancer, brain cancer	Problems of poor transformer generalization bias and poor AMC interpretive ability	ViT-based network with adaptive model fusion and multi-objective optimization	95.14 / 96.17

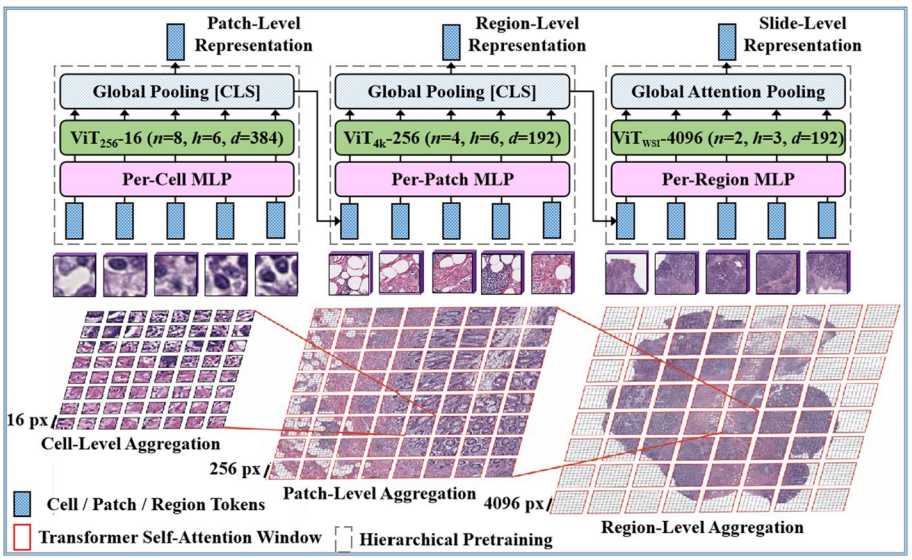

Fig. 14 A schematic illustration of the HIPT [22] transformer architecture for histopathological image representation

Transformer applications in other histopathological imaging tasks

This section briefly discusses the use of transformers in other histopathological gigapixel image domains, such as cross-modal retrieval analysis, image generation, image synthesis, and so on. Dingyi et al. [23] introduced a cross-modal retrieval dual-transformer architecture that can simultaneously execute four retrieval tasks at a time for the histopathology dataset across diagnosis reports and WSIs, respectively. MedViTGAN [24] developed a conditional GAN transformer-based network that can aid researchers in producing synthetic histopathological images for other downstream tasks in an end-to-end approach. In addition, a ViT-based network to enhance the use of contextual information found in histopathological images was proposed in [109]. The network is made up of two variations of ViT-based architecture (PREViT and ClusterViT) to improve the local context of the tissue patch features by adding prior knowledge to the network. Xu et al. [110] developed a Transformer architecture for high-quality histopathological image synthesis that combines ViT and diffusion autoencoders. The authors introduced a conditional denoising diffusion implicit model (DDIM) into the architecture, which was improved by integrating a ViT model as a semantic encoder, allowing it to comprehensively encode sophisticated phenotypic layouts particular to histopathology.

To this end, since the current transformer-based architectures give better results on other histopathological image tasks, we anticipate further development in these fields in the near future.

Discussion

Transformer architectures have been effectively used in a wide range of clinical tasks, including histopathological image analysis, as demonstrated in Tables 1, 2, 3, 4, and 5. Despite their strong performance due to the attention mechanisms incorporated into them, there are a number of challenges that could prevent transformer models from performing well, especially in real-world clinical applications. One of these challenges is the lack of supervised clinical information provided by experts in order to develop a supervised Transformer architecture, which is critical in training a transformer model. Therefore, in this section, we will discuss the recent research directions in addressing this challenge using the SOTA transformer models in different learning settings and also compare transformers with CNNs based on recently published papers.

Different learning settings with transformer architectures

In this section, we present and discuss different learning settings that are often used with transformer architectures for histopathological image analysis, including weakly supervised learning, self-supervised learning (SSL), multi-task learning (MTL), and multimodal learning (MML).

(i) Weakly supervised learning for histopathological imaging: The creation of a weakly supervised transformer-based architecture addresses the urgent need for histopathological image annotation, which is generally labor-intensive and also time-consuming. On the other hand, weakly supervised learning is an arduous task where a huge number of instances occur within each bag while only a slide label is provided. Multiple instance learning (MIL), which is also a subset of weakly supervised learning, has shown better results in a number of downstream tasks in previous studies. Despite recent improvements, they still have some drawbacks. One is that they concentrate on the regions that are easily distinguished as positive for the diagnosis while neglecting the positives that make up a small proportion of the WSI. For the purpose of obtaining more discriminative features, several studies have developed a number of weakly supervised approaches for histopathological image analysis, including segmentation [40, 83], and classification [2, 17, 45, 56, 111] tasks. Besides, the weakly supervised ViT-based MIL technique was further adopted for colorectal cancer lymph node metastasis (LNM) prediction [56] and can be used to reduce the doctor's workload and accelerate diagnostic operations.

(ii) Self-supervised learning for histopathological imaging: Supervised learning techniques heavily rely on pathologists to manually annotate several regions on WSIs before they can be used to train any network. However, this approach often requires a significant amount of annotated datasets for transformer architectures to be successfully trained in many computer vision tasks, and some of these datasets are uncommon in real clinical settings. Therefore, such problems were addressed by self-supervised learning (SSL) techniques. The primary objective of SSL is to enhance the performance of different downstream tasks by conveying knowledge from the associated unsupervised upstream task and pre-training the network by making use of its self-contained features in the untagged data. The standard method for training SSL ViT architectures involves pre-training the architecture primarily on ImageNet and then fine-tuning it on the targeted histopathological image dataset. This will generally improve the performance of

transformers in contrast to CNNs and allow for the attainment of SOTA accuracy. A significant amount of studies have attempted to apply the SSL approach for a variety of objectives in histopathological image analysis, including representation [22], classification [64, 75], and survival analysis and prediction [25] tasks.

(iii) Multi-task learning for histopathological imaging: Multi-task learning (MTL) is a technique in which a shared network learns multiple tasks at the same time. It has shown better performance than single-task learning techniques in increasing the learning capabilities of a deep learning architecture. Such techniques offer many benefits, such as preventing overfitting through the use of shared representations, speeding up learning by utilizing auxiliary information, and increasing data efficiency. However, building transformer architectures with multiple tasks assists in increasing the network generalization ability, which is crucial in histopathological image analysis. Recently, MTL has been commonly applied to transformer-based networks to tackle various downstream tasks in computer vision, and a commonly used technique is to combine a segmentation and classification task into a single network [40, 85]. In addition, Ali et al. [92] introduced a transformed-based CAD system by making use of deep CNN networks based on channel boosting techniques to improve the learning capability of the entire network. Wang et al. [45] built a weakly supervised transformer architecture by integrating graph neural networks and transformers for basal cell carcinoma classification and detection.

(iv) Multi-modal learning for histopathological imaging: Multi-modal learning (MML) is an approach that aims to build and develop models that can integrate data from multiple modalities, such as image data, genomic data, and clinical records. Over the years, research advancements in MML have grown rapidly in a number of computer vision tasks, particularly histopathological image analysis. It involves utilizing a single model to learn representations from various modalities. Using data from multiple modality sources, on the other hand, provides additional clues for disease diagnosis. Several studies have investigated the integration of genomic data and histopathological images for survival analysis and prediction using transformer-based architectures [26, 100, 101]. Takagi et al. [18] proposed a ViT-based personalized attention mechanism network for histopathological images with clinical records. AMIGO [3] created a multi-modal graph transformer architecture that predicts patient survival based on multi-modal histopathological images and shared related data. Cai et al. [60] created a frequency-domain transformer architecture that integrates frequency and spatial domains for histopathological lung cancer image analysis and subtype determination.

In summary, transformer architecture is regarded as a promising technique for fusing computer vision and NLP tasks. However, there is still a need to develop more accurate and robust CAD systems for real-time clinical settings where multiple data types, such as imaging, clinical, and laboratory records, are regarded as multiple sources of information.

Comparison of transformers and CNNs on different downstream tasks

Over the years, CNN-based architectures have been dominant in many research fields prior to the development of vision transformers (ViTs), including the field of histopathological image analysis. Many studies have also been conducted in this domain to ascertain whether CNN-based architectures can still work on ViT-based architectures.

Recently, ViT-based architectures have been shown to be capable of producing better results than CNNs, especially when pre-trained on a large number of datasets. In comparison to CNNs, ViTs have a weaker inductive bias, and as a result, they allow for more flexible feature detection. The performance comparison between ViTs and CNN-based models has received tremendous attention, as ViTs have excelled in a number of benchmarks, as shown in Fig. 11. Nguyen et al. [112] comprehensively evaluated six frequently used Transformer-based architectures for cancer segmentation. The results obtained, with the exception of Swin-UNet [113], show that Transformer-based architectures typically outperform CNN-based techniques because of their capacity to encode global context. Besides, this is one of the first studies to systematically compute the performance of transformer-based approaches on histopathological image segmentation. For the task of tumor detection and tissue type identification in digital pathology WSIs, Deininger et al. [114] compared the patch-wise classification result of the ViT DeiT-Tiny to the SOTA CNN-based ResNet18 model. Due to the limited number of annotated slide images, the authors further compared the two architectures by pre-training them on a large number of unlabeled WSIs using SOTA self-supervised techniques. The obtained results demonstrate that the ViT slightly outperformed the ResNet18 for three out of the four tissue types investigated in the study for tumor detection, while the ResNet18 architecture slightly outperformed the ViT for the remaining tasks. In addition, Springenberg et al. [115] conducted an extensive evaluation of deep learning architectures for histopathological image classification by comparing Transformers and CNNs, respectively. The study produced concrete architecture recommendations for medical practitioners as well as a generic approach for quantifying architecture quality based on complementary conditions that can be applied to future network architectures.

In summary, many previous studies and SOTA on histopathological image analysis have not completely shown that transformer-based architectures can outperform CNN-based architectures in all ramifications, especially in few-shot and low-resolution histopathological image analysis. Thus, developing hybrid architectures with convolutions, similar to approaches in computer vision, has been adopted in most current research works. In addition, apart from the excellent results achieved in most publications surveyed in this paper, as demonstrated in Tables 1, 2, 3, 4, and 5, transformer architectures are computationally expensive and require a large amount of data for training. Therefore, we anticipate further development in reducing transformer computational complexity in the near future.

Other challenges and future directions

We primarily reviewed the current SOTA Transformer-based methods for histopathological image analysis. There are still a number of open challenges to be addressed in the future, despite the excellent and outstanding results produced. (1) The first challenge is the intensiveness of annotations. Transformer-based architectures often need a large number of annotated datasets and can produce better results when trained on huge datasets, but their performance reduces when data or annotations are limited. To solve this problem, SSL techniques offer better and more interesting solutions. Transformers, on the other hand, can improve their capacity for representational learning by making use of unlabeled data and proxy tasks like reconstruction and contrastive learning. A

significant number of studies have applied the self-supervised approach for a variety of objectives in histopathological image analysis [22, 25, 64, 75] and have shown better performance. Some of these approaches have demonstrated that training networks using large-scale unlabeled 2D images is advantageous when fine-tuning them with small-scale datasets. However, we find that pre-training is computationally expensive, and future research should focus on simplifying and analyzing the efficiency of the pre-training model as well as fine-tuning it for small-scale datasets. (2) The second challenge is the scalability of the task. The heterogeneous nature of histopathological images makes representational learning very difficult. Studies in the past have mainly concentrated on resolving specific histological tasks, and transformer architectures perform better at learning heterogeneous tasks, especially when SSL techniques are adopted [22]. Again, the advanced scaling operations also give the transformer-based architectures the capacity to handle multi-domain and multi-scale tasks [22]. In addition, networks may fit a variety of datasets by scaling up transformer architectures, and researchers can modify a network at training time to move from a low-data scheme to larger dimensions. (3) The third challenge is the scalability of the data. Most ViT-based architectures, such as the original ViT [8], perform poorly when trained on small-scale datasets because they lack inductive bias. However, If there is enough training data, transformer architectures can overcome inductive bias challenges by employing different pre-training techniques. Besides, pre-training techniques [22, 69, 82] have also shown better performance in increasing the generalization ability of transformer architectures for histopathological imaging. Moreover, gathering large-scale datasets in the histopathological imaging domain is sometimes impractical due to time-consuming manual annotations and patient privacy concerns. Since gathering large-scale datasets across different imaging modalities still poses a lot of challenges, it is therefore essential to build transformer architectures that are less data-demanding for histopathological imaging applications by incorporating inductive bias mechanisms into transformer models, and we hope to see further research addressing this challenge in the near future.

(4) The fourth challenge is computational complexity. As shown in the previous sections, transformer-based architectures are computationally expensive due to the computation of the self-attention mechanism, which is usually quadratic to the size of the input image. This issue appears to be less of a problem with natural images, but with histopathological images, it is a significant difficulty. Again, this is because histopathological images such as WSIs come in gigapixels and are larger in size compared to natural image datasets. Unlike natural images such as ImageNet, which have fewer pixels, histopathological WSIs can be as huge as 150,000 x 150,000 pixels [22]. Compared with training strategies used for natural imaging models, transformer-based architectures for histopathological image analysis are typically more compacted and sometimes trained using patched input or even smaller batch sizes. The majority of SOTA transformer-based methods for histopathological image analysis are either built upon the already-existing transformer networks [15, 48, 51, 75] or make use of CNNs for feature extraction before being fed into a transformer [56, 61, 63, 81, 82]. Moreover, several studies have suggested that Softmax may be circumvented to linearize the computation of the self-attention mechanism, although none of these techniques have been used and applied to histopathological imaging yet. Therefore, we hope to see further research in

this particular direction. (5) The fifth challenge is the combination of data from different sources. An emerging research area such as imaging genomics has opened up new possibilities for cancer detection and prediction. Using data from multiple modality sources, on the other hand, provides additional clues for disease diagnosis. Some of the Transformer-based architectures surveyed in this paper make use of histopathological images and genomic records for different downstream tasks [100, 101]. However, generating reports from other clinical or medical domains has its own challenges due to their unique nature and varied features. Therefore, how to properly incorporate data from multiple sources for more accurate disease identification and prediction is another interesting and promising future research direction. (6) The sixth challenge is the black box and its interpretability. Over the years, several studies have been conducted on histopathological imaging using various deep learning techniques. Deep learning methods such as CNN sometimes function as black box solutions and are typically more difficult to explain. transformers, on the other hand, make use of a self-attention mechanism that imitates some human behaviors but still functions as a black box and is unable to reveal how different factors are combined to generate results. Given the importance of network interpretability in histopathological image analysis, it is critical to investigate the interpretability of transformer-based architectures. One of the common methods for visualizing Transformer architectures is to calculate relevancy scores from either single or shared attention blocks. The MHSA module in transformer architectures establishes a direct link between tokens, providing an intuitive guide for decision-making. Recently, visual language pre-training [69] has also been adopted for histopathological imaging, and the majority of the WSI-level diagnosis or prediction networks are computed in a black box, making it impossible for humans to understand which region of the slide has the greatest influence on the final prediction. Hence, in order to make the networks more understandable, it is preferable to construct a transformer-based architecture that can identify discriminant patches from the histopathological WSI that generate clinical or medical results.

Conclusion

Transformer architectures are now dominating almost all of the field of computer vision, with a rapid increase in the field of histopathological imaging. In this survey paper, we carry out a thorough review of the applications of transformer architectures in histopathological image analysis. In particular, we survey the applications of transformers in histopathological image classification, segmentation, detection, survival analysis and prediction, and representation and discuss their drawbacks. We found out that the majority of the existing transformer architectures can be naturally and easily applied to histopathological imaging challenges without significant modifications. As a matter of fact, many advanced approaches such as multi-task learning (MTL), weakly supervised learning, multi-modal learning (MML), and model enhancement across various domains are rarely investigated. In addition, we also provided unsolved research problems for further investigation. To this end, despite the outstanding performance of the transformer-based architectures in a number of papers reviewed in this survey, we anticipate that there will be much more exploration of transformers in histopathological image analysis to further increase the efficiency of clinicians, decrease subjectivity, and enhance patient

safety. Moreover, the majority of the diseases reviewed in this paper focused more on histopathological image analysis, and it is expected that in the future, it will be extended to other imaging modalities where multiple data types, such as imaging, clinical, and laboratory records, are regarded as multiple sources of information. We hope that this survey paper provides readers in this domain with a comprehensive idea of transformers.

Author contributions
CCA conceived and wrote the manuscript. JN read and checked the manuscript strictly; HL edited and supervised the manuscript; QS and LY prepared figures and tables. XZ did the final supervision. All authors contributed to the article and approved the submitted version.

Funding
This paper was supported in part by the National Natural Science Foundation of China under Grants 62001063, 61971072 and U2133211, and in part by the Fundamental Research Funds for the Central Universities under Grant 2023CDJXY-037.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 18 July 2023 Accepted: 15 September 2023
Published online: 25 September 2023

References

1. Shakarami A, Nicolè L, Terreran M, Dei Tos AP, Ghidoni S. Tcnn: A transformer convolutional neural network for artifact classification in whole slide images. Biomed Signal Process Control. 2023;84: 104812.
2. Li X, Pang S, Zhang R, Zhu J, Fu X, Tian Y, Gao J. Attransunet: An enhanced hybrid transformer architecture for ultrasound and histopathology image segmentation. Comput Biol Med. 2023;152: 106365.
3. Nakhli R, Moghadam PA, Mi H, Farahani H, Baras A, Gilks B, Bashashati A. Sparse multi-modal graph transformer with shared-context processing for representation learning of giga-pixel images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11547–11557. 2023
4. Srinidhi CL, Ciga O, Martel AL. Deep neural network models for computational histopathology: a survey. Med Image Anal. 2021;67: 101813.
5. Wemmert C, Weber J, Feuerhake F, Forestier G. Deep learning for histopathological image analysis. deep learning for biomedical data analysis: techniques, approaches, and applications, 153–169. 2021.
6. Hong R, Fenyö D. Deep learning and its applications in computational pathology. BioMedInformatics. 2022;2(1):159–68.
7. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Adv Neural Inform Process Syst **30** 2017.
8. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: transformers for image recognition at scale. ArXiv. 2020. abs/2010.11929
9. Prakash A, Chitta K, Geiger A. Multi-modal fusion transformer for end-to-end autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7077–7087 2021.
10. Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C. Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF International Conference on computer vision, pp. 6836–6846 2021.
11. George A, Marcel S. On the effectiveness of vision transformers for zero-shot face anti-spoofing. In: 2021 IEEE International Joint Conference on biometrics (IJCB), pp. 1–8 2021.
12. Atito S, Awais M, Wang W, Plumbley MD, Kittler J. Asit: Audio spectrogram vision transformer for general audio representation. arXiv preprint arXiv:2211.13189 2022.
13. Gupta A, Tripathi R, Jang W. Modeformer: Modality-preserving embedding for audio-video synchronization using transformers. In: ICASSP 2023-2023 IEEE International Conference on acoustics, speech and signal processing (ICASSP), pp. 1–5 2023.
14. Mehta S, Rastegari M. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178 2021.

15. Lin A, Chen B, Xu J, Zhang Z, Lu G, Zhang D. Ds-transunet: dual swin transformer u-net for medical image segmentation. IEEE Trans Instru Measure. 2022;71:1–15.
16. Stegmüller T, Bozorgtabar B, Spahr A, Thiran J-P. Scorenet: Learning non-uniform attention and augmentation for transformer-based histopathological image classification. In: Proceedings of the IEEE/CVF winter Conference on applications of computer vision, pp. 6170–6179 2023.
17. Li Z, Cong Y, Chen X, Qi J, Sun J, Yan T, Yang H, Liu J, Lu E, Wang L, et al. Vision transformer-based weakly supervised histopathological image analysis of primary brain tumors. iScience. 2023;26(1): 105872.
18. Takagi Y, Hashimoto N, Masuda H, Miyoshi H, Ohshima K, Hontani H, Takeuchi I. Transformer-based personalized attention mechanism for medical images with clinical records. J Pathol Inform. 2023;14: 100185.
19. Chen H, Li C, Wang G, Li X, Rahaman MM, Sun H, Hu W, Li Y, Liu W, Sun C, et al. Gashis-transformer: a multi-scale visual transformer approach for gastric histopathological image detection. Pattern Recogn. 2022. 130: 108827.
20. Huang P, He P, Tian S, Ma M, Feng P, Xiao H, Mercaldo F, Santone A, Qin J. A vit-amc network with adaptive model fusion and multiobjective optimization for interpretable laryngeal tumor grading from histopathological images. IEEE Trans Med Imaging. 2022. 42(1):15–28.
21. Vu QD, Rajpoot K, Raza SEA, Rajpoot N. Handcrafted histological transformer (h2t): unsupervised representation of whole slide images. Med Image Anal. 2023. https://doi.org/10.1016/j.media.2023.102743.
22. Chen RJ, Chen C, Li Y, Chen TY, Trister AD, Krishnan RG, Mahmood F. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16144–16155. 2022.
23. Hu D, Xie F, Jiang Z, Zheng Y, Shi J. Histopathology cross-modal retrieval based on dual-transformer network. In: 2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE), pp. 97–102. 2022.
24. Li M, Li C, Hobson P, Jennings T, Lovell BC. Medvitgan: End-to-end conditional gan for histopathology image augmentation with vision transformers. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 4406–4413 2022.
25. Huang Z, Chai H, Wang R, Wang H, Yang Y, Wu H. Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, vol. 12908. Springer, pp. 561–570 2021.
26. Chen RJ, Lu MY, Weng W-H, Chen TY, Williamson DF, Manz T, Shady M, Mahmood F. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4025 2021;
27. Li J, Chen J, Tang Y, Wang C, Landman BA, Zhou SK. Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. Med Image Anal. 2023. https://doi.org/10.1016/j.media.2023.102762.
28. Pinckaers H, Bulten W, Laak J, Litjens G. Detection of prostate cancer in whole-slide images through end-to-end training with image-level labels. IEEE Trans Med Imaging. 2021;40(7):1817–26.
29. Shen Y, Ke J. Sampling based tumor recognition in whole-slide histology image with deep learning approaches. IEEE/ACM Trans Comput Biol Bioinform. 2021;19(4):2431–41.
30. Senousy Z, Abdelsamea MM, Gaber MM, Abdar M, Acharya UR, Khosravi A, Nahavandi S. Mcua: multi-level context and uncertainty aware dynamic deep ensemble for breast cancer histology image classification. IEEE Trans Biomed Eng. 2021;69(2):818–29.
31. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 2014.
32. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp. 7132–7141. 2018.
33. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: European Conference on Computer Vision. Springer, pp. 213–229 2020
34. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022. 2021.
35. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357. 2021.
36. Bao H, Dong L, Wei F. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 2021.
37. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on machine learning, pp. 6105–6114. 2019.
38. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: medical image computing and computer-assisted intervention–MICCAI 2015, pp. 234–241, Springer, 2015.
39. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp. 770–778 2016.
40. Zhang S, Zhang J, Xia Y. Transws: Transformer-based weakly supervised histology image segmentation. In: Machine Learning in Medical Imaging, Springer, pp. 367–376 2022.
41. Yin P, Yu B, Jiang C, Chen H. Pyramid tokens-to-token vision transformer for thyroid pathology image classification. In: 2022 Eleventh International Conference on image processing theory, tools and applications (IPTA), pp. 1–6 2022.
42. Dwivedi VP, Bresson X. A generalization of transformer networks to graphs. arXiv preprint arXiv:2012.09699 2020.
43. Nakhli R, Moghadam PA, Mi H, Farahani H, Baras A, Gilks B, Bashashati A.. Sparse multi-modal graph transformer with shared-context processing for representation learning of giga-pixel images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11547–11557 2023.
44. Reisenbüchler D, Wagner SJ, Boxberg M, Peng T. Local attention graph-based transformer for multi-target genetic alteration prediction. In: medical image computing and computer assisted intervention–MICCAI Springer, pp. 377–386. 2022.2022.

45. Yacob F, Siarov J, Villiamsson K, Suvilehto JT, Sjöblom L, Kjellberg M, Neittaanmäki N. Weakly supervised detection and classification of basal cell carcinoma using graph-transformer on whole slide images. Sci Rep. 2023;13(1):1–10.
46. Zheng Y, Gindra RH, Green EJ, Burks EJ, Betke M, Beane JE, Kolachalama VB. A graph-transformer for whole slide image classification. IEEE Trans Med Imaging. 2022;41(11):3003–15.
47. Ding S, Li J, Wang J, Ying S, Shi J. Multi-scale efficient graph-transformer for whole slide image classification. arXiv preprint arXiv:2305.15773 2023.
48. Zidan U, Gaber MM, Abdelsamea MM. Swincup: Cascaded swin transformer for histopathological structures segmentation in colorectal cancer. Expert Syst Appl. 2023;216: 119452.
49. Wang L, Pan L, Wang H, Liu M, Feng Z, Rong P, Chen Z, Peng S. Dhunet: Dual-branch hierarchical global-local fusion network for whole slide image segmentation. Biomed Signal Process Control. 2023;85: 104976.
50. Haq MM, Huang J. Self-supervised pre-training for nuclei segmentation. In: medical image computing and computer assisted intervention–MICCAI 2022, Springer, pp. 303–313. 2022
51. Tummala S, Kim J, Kadry S. Breast-net: Multi-class classification of breast cancer from histopathological images using ensemble of swin transformers. Mathematics. 2022;10(21):4109.
52. Mehta S, Lu X, Wu W, Weaver D, Hajishirzi H, Elmore JG, Shapiro LG. End-to-end diagnosis of breast biopsy images with transformers. Med Image Anal. 2022;79: 102466.
53. Chen Y, Shao Z, Bian H, Fang Z, Wang Y, Cai Y, Wang H, Liu G, Li X, Zhang Y. dmil-transformer: Multiple instance learning via integrating morphological and spatial information for lymph node metastasis classification. IEEE J Biomed Health Inform. 2023. https://doi.org/10.1109/JBHI.2023.3285275.
54. Zhou X, Tang C, Huang P, Tian S, Mercaldo F, Santone A. Asi-dbnet: an adaptive sparse interactive resnet-vision transformer dual-branch network for the grading of brain cancer histopathological images. Interdiscip Sci Comput Life Sci. 2023;15(1):15–31.
55. Ding M, Qu A, Zhong H, Lai Z, Xiao S, He P. An enhanced vision transformer with wavelet position embedding for histopathological image classification. Pattern Recognition. 109532. 2023.
56. Tan L, Li H, Yu J, Zhou H, Wang Z, Niu Z, Li J, Li Z. Colorectal cancer lymph node metastasis prediction with weakly supervised transformer-based multi-instance learning. Med Biol Eng Comput. 2023. https://doi.org/10.1007/s11517-023-02799-x.
57. Barmpoutis P, Yuan J, Waddingham W, Ross C, Hamzeh K, Stathaki T, Alexander DC, Jansen M. Multi-scale deformable transformer for the classification of gastric glands: The imgl dataset. In: Cancer Prevention Through Early Detection, Springer, pp. 24–33. 2022.
58. Alsaafin A, Safarpoor A, Sikaroudi M, Hipp JD, Tizhoosh H. Learning to predict rna sequence expressions from whole slide images with applications for search and classification. Commun Biol. 2023;6(1):304.
59. Gao Z, Hong B, Zhang X, Li Y, Jia C, Wu J, Wang C, Meng D, Li C. Instance-based vision transformer for subtyping of papillary renal cell carcinoma in histopathological image. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp. 29–308. 2021
60. Cai M, Zhao L, Hou G, Zhang Y, Wu W, Jia L, Zhao J, Wang L, Qiang Y. Fdtrans: Frequency domain transformer model for predicting subtypes of lung cancer using multimodal data. Comput Biol Med. 2023;158: 106812.
61. Zheng Y, Li J, Shi J, Xie F, Jiang Z. Kernel attention transformer (kat) for histopathology whole slide image classification. In: International Conference on medical image computing and computer-assisted intervention, Springer, pp. 283–292. 2022.
62. Li H, Yang F, Zhao Y, Xing X, Zhang J, Gao M, Huang J, Wang L, Yao J. Dt-mil: deformable transformer for multi-instance learning on histopathological image. In: medical image computing and computer assisted intervention–MICCAI 2021, Springer, pp. 206–216. 2021.
63. Wang Y, Guo J, Yang Y, Kang Y, Xia Y, Li Z, Duan Y, Wang K. Cwc-transformer: a visual transformer approach for compressed whole slide image classification. Neural Comput Appl. 1–13. 2023
64. Wang X, Yang S, Zhang J, Wang M, Zhang J, Huang J, Yang W, Han X. Transpath: Transformer-based self-supervised learning for histopathological image classification. In: medical image computing and computer assisted intervention–MICCAI 2021. 186–195. 2021.
65. Shao Z, Bian H, Chen Y, Wang Y, Zhang J, Ji X, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Adv Neural Inform Process Syst. 2021;34:2136–47.
66. Zhu H, Lin M, Xu Z, Yao Z, Chen H, Alhudhaif A, Alenezi F. Deconv-transformer (dect): A histopathological image classification model for breast cancer based on color deconvolution and transformer architecture. Inform Sci. 2022;608:1093–112.
67. Zhang J, Kapse S, Ma K, Prasanna P, Saltz J, Vakalopoulou M, Samaras D. Prompt-mil: Boosting multi-instance learning schemes via task-specific prompt tuning. arXiv preprint arXiv:2303.12214. 2023.
68. Xiong C, Chen H, Sung J, King I. Diagnose like a pathologist: Transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification. arXiv preprint arXiv:2301.08125. 2023.
69. Lu MY, Chen B, Zhang A, Williamson DF, Chen RJ, Ding T, Le LP, Chuang Y-S, Mahmood F. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19764–19775. 2023.
70. Ding S, Wang J, Li J, Shi J. Multi-scale prototypical transformer for whole slide image classification. arXiv preprint arXiv:2307.02308. 2023.
71. Yu J, Ma T, Fu Y, Chen H, Lai M, Zhuo C, Xu Y. Local-to-global spatial learning for whole-slide image representation and classification. Computer Med Imaging Graph. 2023;107: 102230.
72. Zou Y, Chen S, Sun Q, Liu B, Zhang J. Dcet-net: Dual-stream convolution expanded transformer for breast cancer histopathological image classification. In: 2021 IEEE International Conference on bioinformatics and biomedicine (BIBM), pp. 1235–1240. 2021.
73. Wang L, Liu J, Jiang P, Cao D, Pang B. Lgvit: Local-global vision transformer for breast cancer histopathological image classification. In: ICASSP 2023 - 2023 IEEE International Conference on acoustics, speech and signal processing (ICASSP), pp. 1–5. 2023.

74. Pan L, Wang H, Wang L, Ji B, Liu M, Chongcheawchamnan M, Yuan J, Peng S. Noise-reducing attention cross fusion learning transformer for histological image classification of osteosarcoma. Biomed Signal Process Control. 2022;77: 103824.
75. Cai H, Feng X, Yin R, Zhao Y, Guo L, Fan X, Liao J. Mist: Multiple instance learning network based on swin transformer for whole slide image classification of colorectal adenomas. J Pathol. 2022;259(2):125–35.
76. Zhang H, Chen H, Qin J, Wang B, Ma G, Wang P, Zhong D, Liu J. Mc-vit: Multi-path cross-scale vision transformer for thymoma histopathology whole slide image typing. Front Oncol. 2022;12: 925903.
77. Wang Z, Yu L, Ding X, Liao X, Wang L. Lymph node metastasis prediction from whole slide images with transformer-guided multiinstance learning and knowledge transfer. IEEE Trans Med Imaging. 2022;41(10):2777–87.
78. Wang Z, Yu L, Ding X, Liao X, Wang L. Shared-specific feature learning with bottleneck fusion transformer for multimodal whole slide image analysis. IEEE Trans Med Imaging. 2023. https://doi.org/10.1109/TMI.2023.3287256.
79. Wang X, Yang S, Zhang J, Wang M, Zhang J, Yang W, Huang J, Han X. Transformer-based unsupervised contrastive learning for histopathological image classification. Med Image Anal. 2022;81: 102559.
80. Fu B, Zhang M, He J, Cao Y, Guo Y, Wang R. Stohisnet: A hybrid multi-classification model with cnn and transformer for gastric pathology images. Computer Methods Program Biomed. 2022. https://doi.org/10.1016/j.cmpb.2022.106924.
81. Zhao Y, Lin Z, Sun K, Zhang Y, Huang J, Wang L, Setmil Yao J. Spatial encoding transformer-based multiple instance learning for pathological image analysis Medical Image Computing and Computer assisted intervention-MICCAI. Berlin: Springer; 2022.
82. Jiang S, Hondelink L, Suriawinata AA, Hassanpour S. Masked pre-training of transformers for histology image analysis. arXiv preprint arXiv:2304.07434 2023.
83. Qian Z, Li K, Lai M, Chang EI-C, Wei B, Fan Y, Xu Y. Transformer based multiple instance learning for weakly supervised histopathology image segmentation In Medical Image Computing and computer assisted intervention-MICCAI. Berlin: Springer; 2022.
84. Ji Y, Zhang R, Wang H, Li Z, Wu L, Zhang S, Luo P. Multi-compound transformer for accurate biomedical image segmentation medical image computing and computer assisted intervention-MICCAI. Berlin: Springer; 2021.
85. Chen Y, Jia Y, Zhang X, Bai J, Li X, Ma M, Sun Z, Pei Z, Tshvnet, et al. Simultaneous nuclear instance segmentation and classification in histopathological images based on multiattention mechanisms. BioMed Res Int. 2022;2022. https://doi.org/10.1155/2022/7921922.
86. Diao S, Tang L, He J, Zhao H, Luo W, Xie Y, Qin W. Automatic computer-aided histopathologic segmentation for nasopharyngeal carcinoma using transformer framework computational mathematics modeling in cancer analysis. Berlin: Springer; 2022.
87. Chen B, Liu Y, Zhang Z, Lu G, Kong AWK. Transattunet: multi-level attention-guided u-net with transformer for medical image segmentation. arXiv preprint arXiv:2107.05274. 2021.
88. Guo Z, Wang Q, Müller H, Palpanas T, Loménie N, Kurtz C. A hierarchical transformer encoder to improve entire neoplasm segmentation on whole slide image of hepatocellular carcinoma. arXiv preprint arXiv:2307.05800. 2023.
89. Li Z, Tang Z, Hu J, Wang X, Jia D, Zhang Y. Nst: a nuclei segmentation method based on transformer for gastrointestinal cancer pathological images. Biomed Signal Process Control. 2023;84: 104785.
90. Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM. Medical transformer: gated axial-attention for medical image segmentation medical image computing and computer assisted intervention-MICCAI. Berlin: Springer; 2021.
91. Qin W, Xu R, Jiang S, Jiang T, Luo L. Pathtr: Context-aware memory transformer for tumor localization in gigapixel pathology images. In: Proceedings of the Asian Conference on Computer Vision, pp. 3603–3619. 2022.
92. Ali ML, Rauf Z, Khan AR, Khan A. Channel boosting based detection and segmentation for cancer analysis in histopathological images. In: 2022 19th International Bhurban Conference on applied sciences and technology (IBCAST), pp. 1–6 2022.
93. Yücel Z, Akal F, Oltulu P. Mitotic cell detection in histopathological images of neuroendocrine tumors using improved yolov5 by transformer mechanism. Signal Image Video Process. 1–8 2023.
94. Lv Z, Yan R, Lin Y, Wang Y, Zhang F. Joint region-attention and multi-scale transformer for microsatellite instability detection from whole slide images in gastrointestinal cancer medical image computing and computer assisted intervention-MICCAI. Berlin: Springer; 2022.
95. Liaqat Ali M, Rauf Z, Khan A, Sohail A, Ullah R, Gwak J. Cb-hvtnet: A channel-boosted hybrid vision transformer network for lymphocyte assessment in histopathological images. arXiv e-prints. 2305. 2023.
96. Hossain MS, Shahriar GM, Syeed MM, Uddin MF, Hasan M, Shivam S, Advani S. Region of interest (roi) selection using vision transformer for automatic analysis using whole slide images. Sci Rep. 2023;13(1):11314.
97. Lv Z, Lin Y, Yan R, Wang Y, Zhang F. Transsurv: Transformer-based survival analysis model integrating histopathological images and genomic data for colorectal cancer. IEEE/ACM Transactions on Computational Biol Bioinform 1–10. 2022.
98. Lv Z, Lin Y, Yan R, Yang Z, Wang Y, Zhang F. Pg-tfnet: Transformer-based fusion network integrating pathological images and genomic data for cancer survival analysis. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 491–496. 2021.
99. Shen Y, Liu L, Tang Z, Chen Z, Ma G, Dong J, Zhang X, Yang L, Zheng Q. Explainable survival analysis with convolution-involved vision transformer. Proc AAAI Conf Artif Intell. 2022;36:2207–15.
100. Li C, Zhu X, Yao J, Huang J. Hierarchical transformer for survival prediction using multimodality whole slide images and genomics. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 4256–4262, 2022.
101. Jaume G, Vaidya A, Chen R, Williamson D, Liang P, Mahmood F. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. arXiv preprint arXiv:2304.06819 2023.
102. Wang Z, Gao Q, Yi X-P, Zhang X, Zhang Y, Zhang D, Liò P, Bain C, Bassed R, Li S, et al. Surformer: An interpretable pattern-perceptive survival transformer for cancer survival prediction from histopathology whole slide images. SSRN 4423682. 2023.
103. Shao Z, Chen Y, Bian H, Zhang J, Liu G, Hvtsurv Zhang Y. Hierarchical vision transformer for patient-level survival prediction from whole slide image. Proc AAAI Conf Artif Intell. 2023;37:2209–17.

104. Li Z, Jiang Y, Lu M, Li R, Xia Y. Survival prediction via hierarchical multimodal co-attention aransformer: a computational histology-radiology solution. IEEE Trans Med Imaging. 2023. https://doi.org/10.1109/TMI.2023.3263010.
105. Kapse S, Das S, Prasanna P. Cd-net: Histopathology representation learning using pyramidal context-detail network. arXiv preprint arXiv:2203.15078. 2022.
106. Liu P, Fu B, Ye F, Yang R, Dsca Ji L. A dual-stream network with cross-attention on whole-slide image pyramids for cancer prognosis. Expert Syst Appl. 2023;227: 120280.
107. Chan TH, Cendra FJ, Ma L, Yin G, Yu L. Histopathology whole slide image analysis with heterogeneous graph representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15661–15670. 2023.
108. Vu QD, Rajpoot K, Raza SEA, Rajpoot N. Handcrafted histological transformer (h2t): unsupervised representation of whole slide images. Med Image Anal. 2023;85: 102743.
109. Wood R, Sirinukunwattana K, Domingo E, Sauer A, Lafarge MW, Koelzer VH, Maughan TS, Rittscher J. Enhancing local context of histology features in vision transformers Artificial Intelligence over infrared images for medical applications and medical image assisted biomarker discovery. Berlin: Springer; 2022.
110. Xu X, Kapse S, Gupta R, Prasanna P. Vit-dae: Transformer-driven diffusion autoencoder for histopathology image analysis. arXiv preprint arXiv:2304.01053 2023.
111. Myronenko A, Xu Z, Yang D, Roth HR, Xu D. Accounting for dependencies in deep learning based multiple instance learning for whole slide imaging. Berlin: Medical Image Computing and Computer Assisted Intervention-MICCAI. Springer; 2021.
112. Nguyen C, Asad Z, Deng R, Huo Y. Evaluating transformer-based semantic segmentation networks for pathological image segmentation medical imaging 2022. Image Process. 2022;12032:942–7.
113. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M. Swin-unet: Unet-like pure transformer for medical image segmentation. In: Computer Vision–ECCV 2022 Workshops, Springer, pp. 205–218. 2023.
114. Deininger L, Stimpel B, Yuce A, Abbasi-Sureshjani S, Schönenberger S, Ocampo P, Korski K, Gaire F. A comparative study between vision transformers and cnns in digital pathology. arXiv preprint arXiv:2206.00389. 2022.
115. Springenberg M, Frommholz A, Wenzel M, Weicken E, Ma J, Strodthoff N. From cnns to vision transformers–a comprehensive evaluation of deep learning models for histopathology. arXiv preprint arXiv:2204.05044. 2022.

Publisher's Note