Sketch-supervised Histopathology Tumour Segmentation: Dual CNN-Transformer with Global Normalised CAM

Yilong Li, Linyan Wang, Xingru Huang, Yaqi Wang†, Le Dong, Ruiquan Ge, Huiyu Zhou, Juan Ye, Qianni Zhang†

***Abstract*— Deep learning methods are frequently used in segmenting histopathology images with high-quality annotations nowadays. Compared with well-annotated data, coarse, scribbling-like labelling is more cost-effective and easier to obtain in clinical practice. The coarse annotations provide limited supervision, so employing them directly for segmentation network training remains challenging. We present a sketch-supervised method, called DCTGN-CAM, based on a dual CNN-Transformer network and a modified global normalised class activation map. By modelling global and local tumour features simultaneously, the dual CNN-Transformer network produces accurate patch-based tumour classification probabilities by training only on lightly annotated data. With the global normalised class activation map, more descriptive gradient-based representations of the histopathology images can be obtained, and inference of tumour segmentation can be performed with high accuracy. Additionally, we collect a private skin cancer dataset named BSS, which contains fine and coarse annotations for three types of cancer. To facilitate reproducible performance comparison, experts are also invited to label coarse annotations on the public liver cancer dataset PAIP2019. On the BSS dataset, our DCTGN-CAM segmentation outperforms the state-of-the-art methods and achieves 76.68 % IOU and 86.69 % Dice scores on the sketch-based tumour segmentation task. On the PAIP2019 dataset, our method achieves a Dice gain of 8.37 % compared with U-Net as the baseline network. The dataset, annotation and code will be published at https://github.com/skdarkless/DCTGN–CAM.**

***Index Terms*— Sketch supervision, tumour segmentation, transformer, global normalised CAM**

This research is supported in part by the National Natural Science Foundation of China (No. 62206242), the Zhejiang Provincial Natural Science Foundation of China (Grant No. LY21F020017).

Yilong Li, Xngru Huang, Qianni Zhang are with School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK (e-mail: {yilong.li,xingru.huang, qianni.zhang}@qmul.ac.uk).

Yaqi Wang is with the College of Media Engineering, Communication University of Zhejiang, Hangzhou, China (e-mail: wangyaqi@cuz.edu.cn).

Le Dong is with School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China (e-mail: ledong@uestc.edu.cn).

Rui n Ge is with School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China (e-mail: gespring@hdu.edu.cn).

Linyan Wang and Juan Ye are with the Second Affiliated Hospital of Zhejiang University, Hangzhou, China (e-mail: {linyan.wang, yejuan}@zju.edu.cn).

Huiyu Zhou is with School of Computing and Mathematic Sciences, University of Leicester, Leicester, UK (e-mail: hz143@leicester.ac.uk).

I. INTRODUCTION

CANCER is one of the most deadly diseases in the world. Despite tumour resection surgery, patients are at high risk of recurrence. Pathologists create stained histology slides using samples of the resected tumour tissue, to assess the effect of the pre-operation treatment regimen. The visual examination of histopathological images involves searching for specific medical features such as the tumour's shape, location and growth pattern [1]. In clinical practice, digital scanners [2] capture digitised whole slide images (WSIs), making the visual examination of histopathology slides easier and more flexible. Nevertheless, the time and efforts required for pathologists to visually analyse WSIs of every single case are enormous as a result of the large number of slides to be analysed, in contrast to the limited availability of specialised pathologists. Besides, visual evaluations are inherently subject to inter-observer and intra-observer variabilities. The inconsistent and imprecise output annotations may be not satisfied, leading to a negative impact on the actual diagnosis and future treatment planning [3].

Recent improvements in computer vision open new revenues for (semi-)automatic analysis of digital WSIs, saving significant time and resources in manual analysis. Tumour segmentation in histopathological images is heavily dependent on the quality and quantity of annotated ground truth boundaries, which, however, are costly and challenging to acquire. Specifically, tumour borders are complex, vague, and non-rigid, making it extremely difficult for even experienced pathologists to define.

In a more practical scenario, pathologists tend to mark tumour regions with a rough outline instead of drawing out every detail around the tumour border [4]. However, in normal deep-learning methods, the learned models generate predictions that are equivalent to the training labels. Accurate, detailed boundaries can only be predicted by segmentation models trained with accurate, detailed annotations. It becomes unusual, yet highly demanded capability, to segment tumour regions accurately using models trained on coarse markings. Particularly, factors such as intrinsic tumour variance, patient

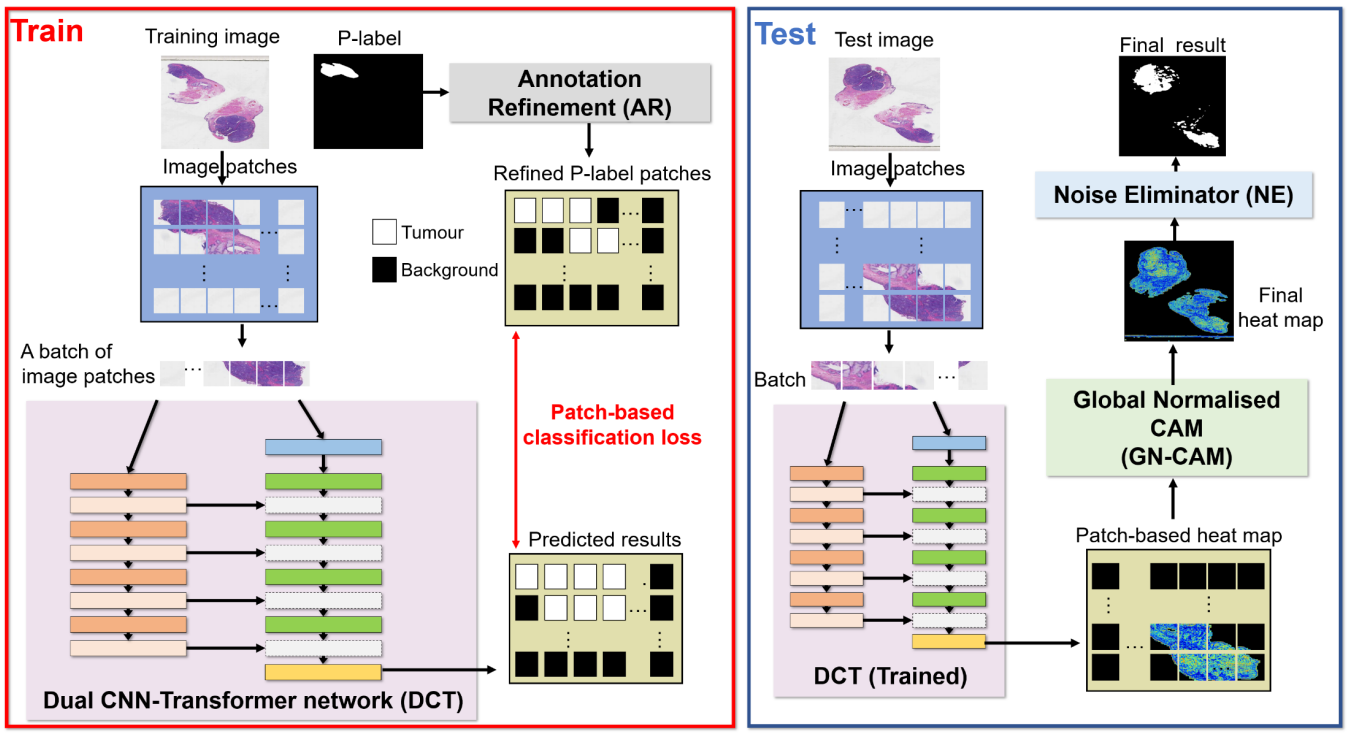


Fig. 1. The pipeline of the proposed sketch-supervised tumour segmentation method DCTGN-CAM. A Dual CNN-Transformer classification network (DCT) is trained by the tumour image patches and refined P-label patches, processed by an annotation refinement (AR), and supervised by a binary classification (cross-entropy) loss. In the test stage, the testing image patches are passed through the trained DCT classification network. A GN-CAM visualizer combines the local with global tumour heat maps simultaneously to create accurate tumour boundaries. The final denoising tumour segmentation results are obtained by a noise eliminator (NE).

variance and technical variance generated in scanning will further complicate training and result in unsatisfactory prediction results.

Weakly supervised methods try to solve model training with coarse annotations including category-based, sketch-based, bounding box-based, point-based, and interaction-based categories [5]. However, weakly supervised methods only train the model on the coarse annotations without consideration of subsequent annotation refinement. Thus, in this paper, we are motivated to boost supervision signals by refining coarse annotations. Additionally, fully convolutional networks, e.g., VGG [6], GoogleNet [7]and ResNet [8], have been adopted as the mainstream backbones to extract medical object features supervised by coarse annotations. Transformer-based methods have recently been proposed for the representation of global features and the segmentation of medical objectives on 2D and 3D images [9] . There have been several CNN-Transformer fusion studies that demonstrate the brilliant performance of global features on CNN structures. Inspired by these approaches, we propose a novel scheme to join the strength of CNN and Transformers in the context of sketch-based supervision for segmenting tumour regions.

This research is committed to developing a tumour segmentation model that can learn from the light annotation of coarse region boundaries, and once trained, is able to define accurate tumour boundaries with fine details on unseen histology images. To facilitate experiments and evaluation, we acquire two versions of annotations of tumour regions on our target datasets, a set of poor-quality labels (P-label) and a set of fine-quality labels (F-label). P-label can be obtained with relatively light efforts by pathologists and is used for training the models. In contrast, F-label requires significant time to prepare, and in this paper is only utilised as ground truth. The accuracy of F-labels is far better than that of the P-labels. Based on this scenario, we propose a framework that follows a sketch-supervised paradigm [10]. More specifically, it aims to generate accurate tumour region masks by models learned only from P-labels. The core of the framework entails a Dual CNN-Transformer network (DCT), supported by a Global Normalisation class activation map (GN-CAM). The main idea of the Dual CNN-Transformer structure is to integrate the advantages of CNN and Transformer, and provide descriptive joint global and local tumour representations. This dual network structure forms the foundation for the sketch-based tumour segmentation task. Fig. 1 introduces the training and testing pipelines of our DCTGN-CAM method. The overall framework contains four main components:

Annotation refinement (AR): Due to the poor quality of the P-label, the network training may be greatly limited. To alleviate this problem, the P-label is first pre-processed before being used for training. A classical colour-based segmentation is performed based on pixel grouping using the k-means algorithm, and the resulting region boundaries are characterised by their precision to colour, contrary to the semantic-guided manual annotation by experts. Therefore, the use of this method can greatly optimize the accuracy of the

P-label.

Dual CNN-Transformer Network (DCT-Net): The patch-based segmentation paradigm is a common resort to image segmentation in many applications [11], which transfers segmentation to a patch classification problem and establishes the segmentation borders based on patch class grouping. A common issue in this approach is that the information contained in each individual patch is limited, and the patch classification result is often non-ideal [12]. Intuitively, a feature extraction method that can take into account the information from a patch's neighbourhood and exploit the contextual visual cues is the key. As large false-positive patch exists on the sketch-based coarse annotations, showing that the global relationship among patches with different locations is one of the main challenges. In this regard, we propose a dual-branch structure that includes a local CNN branch and a Parallel SWIN transformer branch, to extract the feature relationship between patches while containing the suitability with the CAM module. The CNN branch is designed to extract accurate boundaries inside patches while the Transformer branch is to eliminate the effect of coarse false-positive patches according to the high-dimensional global representations. We proposed a structure that concatenates each convolution block with parallel SWIN transformer blocks, the output of this network is then input to the CAM module.

A **global-normalised CAM module (GN-CAM):** Class activation mapping can show the deep focus of the features. Therefore, this method can also be used to generate more specific thermal maps for deep features. Thus, we design a method to calculate patch-based segmentation annotation using the heat map generated by CAM. Since CAM is generally used to explain the results of patch-based classification tasks, we propose a fusion method for the results of two CAM. Ordinary CAM has a high recognition of the details of each patch, but this method lacks consistency and cannot cope with the connection between patches. We designed a Global Normalised CAM, which calculates the thermal map of CAM from WSI rather than each patch. This CAM ensures continuity in WSI and smooth boundaries of tumour sites. Finally, we fuse the results of two kinds of CAM.

A **Noise Elimination module (NE):** The patch-based network will face the problem of large boundary error, even after using the heat maps generated by CAM. Therefore, our NE module is based on matrix processing to optimize tumour boundaries and eliminate noise. This method can smooth the boundaries of patch-based segmentation results, and eliminate the noise caused by CAM heat maps to a large extent.

Our proposed method has more satisfying performances against the popular methods in this area, and our final result is even better than the primitive P-label while using F-label in the evaluation as ground truth. The average improvement of our proposed methods against the P-label is more than 20 %, which proves the success of our work as a sketch-supervised framework and also an ideal way of annotation improvement of poor-quality annotations. Our contributions are threefold:

- By calculating the intersection of cancer regions in unsupervised k-means and sketch annotations, annotations from experts are optimized to facilitate subsequent patch-based cancer localization.

- A dual-branch DCT classification method leverages the tumour features comprehensively. The proposed Parallel SWIN Transformer block ensures the consistency of global feature representation.

- A Global-normalised CAM is introduced to generate a whole-slide-based heat map from patch-based tumour classification predictions, which combine the local and global normalisation.

II. Related works

A. Tumour segmentation

Conventional image processing techniques, such as Otsu thresholding, Canny, Fuzzy C-mean and Watershed segmentation, do not work effectively when applied to tissue segmentation in histopathology images, as these methods cannot capture either the local low-level features along tumour boundaries or the global semantic features. Instead of relying on manually crafting features, deep convolution networks make a more straightforward choice by training the models to extract the most relevant and descriptive feature information. Patch-based classification networks like [13] and [14] achieve WSI image segmentation with low computational complexity while sacrificing boundary smoothing conditions. U-Net is one of the most widely used techniques in patch-based pathological image segmentation [15]. Its main idea is to capture global features on the shrinking path and achieve accurate positioning on the expanding path. However, U-Net does not fully consider the local dependence among pixels, especially when the target to be segmented has weak edges and sparse colouring. To address this issue, a fusion framework is proposed for promoting the accuracy of tumour edge segmentation [16]. Long-range dependencies can be modelled by conditional random fields, which can be exploited to post-process semantic segmentation predictions of the proposed network. However, this method is computational-intense and requires a large number of expert annotations for training.

B. Weakly supervised segmentation

Generally, sketch refers to sparse annotations that provide masks for small areas of pixels [17]. In existing methods, selective pixel loss is usually used for annotated pixels. For model training, some studies attempt to expand sketches or reconstruct the entire mask[18]. Pixel-relabeling requires iterative training. A number of works employ conditional random fields in post-processing [19], [20] or as a trainable layer to refine segmentation results without relabeling[21]. These methods, however, are not effective in providing better supervision for the training of models. More recent methods for evaluating and refining segmentation masks are developed, leading to more accurate predictions, such as a multi-scale attention gate proposed by Gabriele et al. [22], and a PatchGAN discriminator to leverage shape priors by Zhang et al. [23]. However, these methods require additional sources of mask data and are not applicable in more general scenarios.

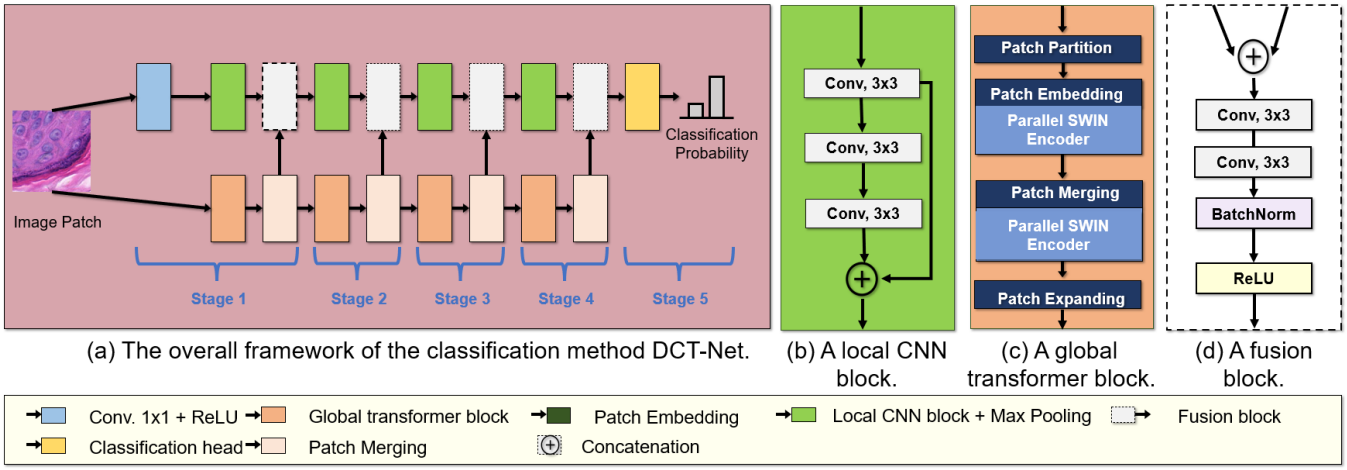


Fig. 2. The structure of the proposed Dual CNN-Transformer network (DCT). Subfigure (a) presents the details of five stages inside the DCT network. Each stage contains a local CNN block and a global transformer block except for the 5_{th} stage. Subfigure (b) shows the residual connection of the CNN block. Subfigure (c) illustrates the subblocks inside the proposed transformer block including the patch partition, the patch embedding, the parallel SWIN encoder, the patch merging and the patch expanding. It is noticeable that the patch partition and the patch embedding layer are only executed in the first stage. Subfigure (d) presents a simple but effective way to fuse the global and local tumour features of the image patches.

Initial cues are essential for weakly supervised segmentation tasks since they provide reliable priors to generate segmentation maps. Class activation map (CAM) can be a good auxiliary as it can provide the preliminary information of object localisation [24]. It highlights class-specific regions that can serve as the initial cues. In [24], the authors demonstrate that a CNN with a Global Average Pooling (GAP) layer has localization capabilities despite not being explicitly trained to do so. Our work in this paper takes inspiration from two algorithms of this kind, namely, CAM [24] and Grad-CAM [25], we intend to resolve their existing issues modelling global tumour representation in the whole slide level.

C. Attention mechanism and Transformers

Attention mechanisms are designed to discover and explore the key parts of a batch of data. All attention modules can be inserted into full convolution networks and extract global context information. Several existing works have applied Non-local modules to segmentation tasks. In [26], the authors introduce a global feature with the Non-local operation. In [27], the local features are integrated with their global dependence adaptive, which models semantic interdependence in spatial and channel dimensions respectively. Global-guided Local Affinity is proposed to play a crucial role in modelling capture context information [28]. Adaptive Context Modules with a pyramid structure are built to present global information. The above non-local-based attention models are not friendly to memory. In order to reduce computation costs, several related works are proposed later. Attention in CCNet collects all kinds of information near and far on crisscross paths with low computation complexity [29]. Inspired by Spatial attention and Non-local block, GCNet uses a simple non-local block with fewer memory requirements [30]. However, these models are embedded into convolution layers and sampling layers. Sampling layers like the pooling layer always lose the details of images, causing poor performances. Additionally, self-attention blocks like Non-local blocks can exploit global information integrated by channel and spatial dimensions, but with high computational complexity [31].

Simon et al. introduce a multi-task learning approach to segment and classify nuclei, glands, lumina, and various tissue regions in digitized pathology slides [32]. This method capitalizes on data from multiple independent sources, ensuring alignment in tissue type and resolution. By employing a single network, they achieve simultaneous predictions for multiple tasks. Rudiger et al., on the other hand, propose a method for merging model paths with different spatial scales while maintaining spatial relationships [33]. They employ a straightforward attention block that can be seamlessly incorporated into standard encoder-decoder networks at various levels. Additionally, they suggest the use of a context classification gate block as an alternative means of incorporating global context solely from diverse spatial scales. These studies exhibit multi-scale feature extraction and concatenation modules, although their complexity for each scale is high. These findings have inspired me to explore multi-scale tumor feature processing while keeping computational costs low.

Recently, the Vision Transformer has emerged as a novel approach by integrating a Non-local block to facilitate attentive interaction among different patch tokens [34]. In the domain of medical image segmentation, several techniques have been developed to address the limitations associated with capturing both global semantic information and local contextual details [35]. The TransUNet method leverages the self-attention mechanism to compute global context [36], while SETR replaces the coding component of conventional convolutional layers with Transformers, resulting in improved segmentation performance [37]. SwinU-Net, resembling Unet architecture, employs a hierarchical Swin Transformer with shifted windows as the encoder to extract contextual features [38]. Additionally, a symmetric Swin Transformer-based decoder with a patch expanding layer is designed to up-sample feature

maps and restore spatial resolution [39]. The MSHT model adopts a multistage hybrid design, combining Transformer blocks with convolutional neural networks (CNNs) to enhance spatial features and leverage the global modeling capabilities of Transformers [40]. SEGTRANSVAE combines an encoder-decoder architecture, a Transformer, and a variational autoencoder (VAE) branch, synergistically utilizing the strengths of CNNs, Transformers, and VAEs, making it a promising solution for medical image segmentation [41]. Nevertheless, the inference time of hybrid CNN-Transformer or pure Transformer structures is longer compared to CNNs, due to the increased computational resources required by Transformer blocks. Inspired by the success of these approaches, we incorporate Transformers into our network design to leverage their capabilities in our proposed hybrid models based on CNN and Transformer methods. Additionally, we also prioritize addressing the computation complexity to ensure efficient processing.

III. METHODS

Fig. 1 shows the overall framework architecture. Using an unsupervised Annotation refinement (AR) module, coarse annotations are first refined as much as possible. Then, supervised by a binary cross-entropy loss, the proposed Dual CNN-Transformer Network (DCT) simultaneously train a fully convolutional network and a transformer for patch classification. Tumour segmentation masks for a test image are inferred based on the patch classification output from DCT. Specifically, the proposed Global Normalised CAM (GN-CAM) calculates gradient-based heat maps derived from the final convolution layer of DCT. To produce the whole heat map with the same size as the WSI, all individual heat map patches are placed in order. Global normalization models the global tumour information over the whole heat map and ensures precise marking of tumour boundaries. Lastly, noise is eliminated using a convolutional CRFs-driven eliminator.

A. Annotation Refinement

P-label is the sketch-like coarse mask drawn by experts. Its boundaries are inexact, meaning that many non-tumour regions around the boundaries are likely to be included inside the mask, and some visual features in these vague regions are considered in training rather than only from the genuine tumour tissue. In contrast, F-label illustrates the tumour regions and boundaries accurately and requires significant time to prepare. In this paper, F-labels are only utilised as ground truth for performance testing.

To relieve these data challenges and improve the training data quality, an annotation refinement module is designed based on the K-means clustering algorithm, to refine the pixel memberships in marginal regions of tumours marked by P-labels. Following this unsupervised process, pixels with similar visual features are grouped together, while regions of distinct colours are better delineated by the mask boundaries. When experts annotate the coarse tumour masks, they tend to do it slightly excessively by including all tumour regions inside the mask, as well as some non-tumour tissues along the margin. Thus, the coarse P-label, denoted as Y_0, only roughly separates non-tumour regions from tumour regions roughly. The AR module is designed to preliminarily improve the coarse masks, by re-examining the tissue membership along the mask margin based on pixel colours. Unsupervised K-means clustering is applied to all pixels of WSIs, creating a new set of labels Y_1 to represent the tumour boundaries. Nevertheless, the Y_1 label sometimes includes some non-tumour pixels that have similar colour features to the tumour regions, while the original coarse annotation Y_0 normally does not include these regions unless they are in contact with the genuine tumour region. The region of Y_0 is usually much larger than the region of Y_1, while Y_1 may consist of some outlier, disconnected regions from the main tissue region. Thus, the refined tumour mask is obtained by finding the intersection of the two $\hat{Y} = Y_1 \cap Y_0$.

B. DCT Network for Patch Classification

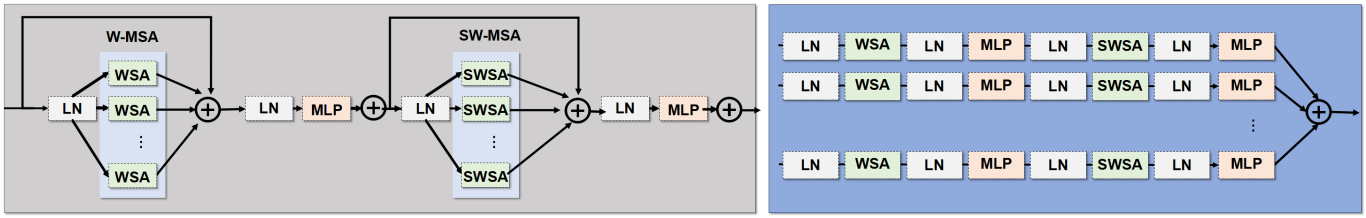
In existing unsupervised or weakly supervised, patch classification-based segmentation methods, VGG is commonly used as a CNN-based backbone for classification. VGG [6] shows substantial improvement through the use of deeper convolutional layers and small kernels, and it is popular in patch-based classification and segmentation tasks on weakly-supervised medical imaging. However, VGG has some internal design limitation that leads to network gradient vanishing and ignorance of long-term dependency among pixels. Besides the VGG structure, UNet is also taken as a commonly used CNN-based backbone to extract high-dimensional features. However, this pixel-wise segmentation structure may introduce more false-positive results when the annotations are coarse sketches, which is unacceptable for weakly supervised tumor segmentation tasks. Recently, several Transformer-based methods attempt to describe global features effectively. Therefore, the an intuitive idea is to exploit Transformers to complement for the lack of CNN structures. In light of the fact that most category-based weakly supervised approaches use VGG as their backbones, taking VGG as our base network makes it easy to demonstrate network improvements and comparisons with other approaches.

Algorithm 1 Annotation refinement by K-means

1: **repeat**
2: Compute the cluster centroids of background C_1 and tumor C_2, where x_i is one pixel of WSI, $x_{i'} \neq x_i$. The k^{th} cluster centroid is the vector of the feature means in the k^{th} cluster.
3: $C_1, C_2 = \min_{C_1,C_2} \left\{ \sum_{\mathrm{k}=1}^{2} \frac{1}{|\mathrm{C_k}|} \sum_{\mathrm{i},\mathrm{i}' \in C_k} \left(x_\mathrm{i} - x_{i'}\right)^2 \right\}$
4: Assign each observation x_i in the unsupervised label Y_1 to the cluster whose centroid is closest.

$$x_\mathrm{i} \in \begin{cases} C_1, & \mid x_\mathrm{i} - x_{C_1} \mid < \mid x_\mathrm{i} - x_{C_2} \mid \\ C_2, & others \end{cases}$$

5: **until** the cluster assignments stop changing.
6: Obtain the refined annotation $\hat{Y}$ by the intersection operation: $\hat{Y} = Y_1 \cap Y_0$.



(I) The SWIN Transformer encoder. (II) Our proposed Parallel SWIN encoder.

Fig. 3. Comparison between the existing SWIN Transformer and our proposed parallel SWIN Transformer block. Feature representation is continuous and independent in our parallel design rather than local layer normalisation (LN) and residual connection in each shifted-window-based/window-based multi-head self-attention module (SW/W-MSA). SW/WSA is the shifted-window-based/window-based self-attention module. MLP is the Multi-layer Perceptron.

Specifically, the VGG network prioritises shallow features (colour) over high-level features (morphological structure) in the pathological image classification task [42]. It means that the convolution networks like VGG lack a global understanding of a whole image, while, for the classification of pathological image patches, the extraction of global semantic features is the key to cancer recognition at the boundary. As a result, VGG cannot classify cancer tissues accurately, especially around cancer borders with complex visual characters. Recently, the emergence of Transformers shows a promising perspective in solving the problem of long-term dependence in the field of computer vision. To combine the strength of convolutional neural networks and Transformers, we propose a dual CNN-Transformer network, namely, the DCT net, which consists of two branches, a CNN branch, and a Transformer branch. In the CNN branch, we substitute the usual fully convolutional block structures as in VGG, with residual blocks to focus on local features. The transformer branch is designed to extract global semantic features that complement the local visual representations. This dual branch structure ensures a robust and precise tumour classification by modelling the local details and global tissue relationship simultaneously.

The dual-branch DCT classification is organised in 5 stages, as shown in Fig. 2. In stage 1, image patches are first passed through a local CNN block, then a global Transformer block, and a fusion block. Stages 2,3 and 4 follow the same structure but have different network parameters. These stages do not contain an extra channel-adjustment convolution layer in the CNN branch compared with stage 1. Finally, in stage 5 a classification head (Linear + BatchNorm + ReLu+ SoftMax) is proposed to generate classification vectors for each image patch according to the output feature of stage 4.

As shown in sub-figure (b) of Fig. 2, each local CNN block contains three convolution layers, and the kernel size of each layer is 3×3. By using a residual connection, the output features of the third convolutional layer are added to those of the first convolutional layer as the final output of the local tumour representation L. Compared with VGG structures, this CNN block ensures more stable gradient backpropagation for weakly supervised learning.

As shown in sub-figure (c) of Fig. 2, our proposed global transformer block has a similar structure to that of the SWIN Transformer block, which includes patch partition, patch embedding, parallel SWIN Transformer encoder, and patch expanding. Extracted from the coarsely annotated masks, some patch labels may be wrongly corresponded to meaningful visual features, leading to problems in the learning of such features and subsequently affecting the decision power of the network. Therefore, we hope to separate the calculation process of each feature as much as possible. Our expectation is that each head can independently extract a type of global feature such as global texture, global tumour colour distribution, or global tumour boundary. However, a SWIN Transformer fuses multi-head attention results for the normal window before calculating self-attention based on shift windows. Consequently, each head cannot model one type of global feature independently, resulting in redundant multi-head shift-window attention. To resolve this issue, we adjust the structure by using a stack of these sub-blocks to ensure the consistency of feature representation and fusing the features of each head after the cascade self-attention calculation on two windows (a normal window and a shift window), as shown in the sub-figure (b) of Fig. 3. Our design ensures a consistent representation of global features.

$$\left[\mathbf{x}_p^1; \mathbf{x}_p^2; \cdots; \mathbf{x}_p^N\right] = x\mathbf{E} \tag{1}$$

$$\mathbf{z}_0 = \left[\mathbf{x}_p^1; \mathbf{x}_p^2; \cdots; \mathbf{x}_p^N\right] + \mathbf{E}_{\text{pos}} \tag{2}$$

$$\mathbf{z}'_\ell = \text{MLP}\left(\text{LN}\left(\text{WSA}\left(\text{LN}\left(\mathbf{z_0}\right)\right)\right)\right) \tag{3}$$

$$\mathbf{z}_\ell = \text{MLP}\left(\text{LN}\left(\text{SWSA}\left(\text{LN}\left(\mathbf{z}'_\ell\right)\right)\right)\right) \tag{4}$$

$$\mathbf{G} = \sum \mathbf{z}_\ell, \ell = 1 \ldots T \tag{5}$$

Rather than flattening the patches x and mapping them by a trainable linear projection [34], we exploit a convolution operation $\mathbf{E}$ shown in equation (1), to project the image patch to a high-dimensional space and to split the image patch into smaller window-based patches simultaneously. The image patch $x \in \mathbb{R}^{H \times W \times 3}$ is transformed into a sequence of patches $x_p \in \mathbb{R}^{\frac{H}{P_H} \times \frac{W}{P_W} \times C}$, where (H, W) is the resolution of image patch, C is the adjusted channel number, (P_H, P_W) is the resolution of each resulting window-based image patch. Then we add the Random Position Embedding $\mathbf{E}_{\text{pos}}$ on these window-based patches by a 3D dropout operation, and obtain the embedding $\mathbf{z}_0$ by the equation (2). Then, the embedding $\mathbf{z}_0$ is passed through a Parallel SWIN encoder. As shown in sub-figure (II) of Fig. 3, a part of our Parallel SWIN encoder consists of Window-based Self Attention (WSA) and

Multilayer Perceptron (MLP) sub-blocks. The other parts of our Parallel SWIN encoder include Shift-Window-based Self Attention (SWSA) and MLP. Additionally, LayerNorm (LN) layers are applied before each sub-block. In Parallel SWIN, sub-blocks are computed exactly the same as in SWIN Transformer, so we will not repeat the computational details inside the WSA and SWSA sub-modules. Parallel SWIN makes feature representation more continuative since the T heads operate independently. As a result of one Parallel SWIN, a global representation called G is produced.

In each stage, the local CNN feature L is concatenated with the global Transformer feature G, and then passed through a convolutional layer, a BatchNorm layer, and a non-linear ReLu layer. Through this fusion module, the two views of tumour features are effectively combined, enabling accurate tumour extraction, and modelling of long-term dependence and local tumour details.

C. CAM Feature Extractor and Visualizer

Class Activation Mapping (CAM) [43] presents exemplary visualisation ability and has attempted to be utilized in sketch-based learning. The aim of CAM is to show interest in features for the target network and thus reveal the focus of the network on every patch. By using CAM, we can see that the network's interest in the tumour region is not only in the colour value, thus the output of CAM will grant the final result a significant improvement in accuracy.

However, CAM can only calculate and analyze tumour information within a patch. In the patch-based segmentation task, the proportion of a patch in a WSI is very small, and the relationship between patches is also very important because it plays a fundamental role in defining tumour region boundaries. Intuitively, we need to consider both information within the patches as well as that between the patches. To this end, we designed a Global Normalised class activation map (GN-CAM). Firstly, we calculate the heat map of CAM in two forms, i.e., the class activation heat map inside the patches and that of the whole WSI. The heat values are fused as the final results of GN-CAM. By taking into account the overall picture of WSI, this fusion result captures the changes of details within patches and prevents noises in visually confusing regions, like tumour boundaries.

As a first step, the global normalised CAM (GN-CAM) collects the gradient-guided information B^l flowing from the DCT-Net's last convolution layer and the output map of the DCT-Net y^{out}. The l^{th} and $(l+1)^{th}$ layers are the two convolutional layers inside the last fusion block of the DCT-Net. Denote i as the channel index of a feature map. Assuming the i^{th} feature map from the $(l+1)^{th}$ layer as y_i^{l+1} according to the gradient backpropagation, the i^{th} guided gradient map from the $(l+1)^{th}$ layer as R_i^{l+1}. The gradient feature of the $l+1$ layer is calculated by

$$y_i^{l+1} = \text{relu}\left(y_i^l\right) = \max\left(y_i^l, 0\right), \tag{6}$$

$$B_i^{l+1} = \frac{\partial y^{\text{out}}}{\partial y_i^{l+1}}. \tag{7}$$

The guided gradient map flowing out the l layer B^l is calculated by:

$$B_i^l = \left(y_i^l > 0\right) \cdot \left(B_i^{l+1} > 0\right) \cdot B_i^{l+1}. \tag{8}$$

We define an updateable queue O_1 to store all guided gradient-based features B from the same whole slide image. Then we decentralize the feature set B by a global normalisation and store the processed features $B^{'}$ in a new queue O_2. Additionally, every pixel $B_{i,m,n}^l$, $m \in M, n \in N$ is normalised locally by a patch-based level, where m, n is one position in the feature map B.

$$\mu_i^l = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N} B_{i,m,n}^l}{MN} \tag{9}$$

$$s_i^l = \sqrt{\frac{\sum_{m=1}^{M}\sum_{n=1}^{N}\left(B_{i,m,n}^l - \mu_i^l\right)^2}{(MN)^2}} \tag{10}$$

$$B_i^{l^{''}} = \frac{B_i^l - \mu_i^l}{s_i^l} \tag{11}$$

The locally normalised features from the same whole slide image are collected in a queue O_3. Each two corresponding normalised gradient maps from O_2 and O_3 are counted together and outputted as the final segmentation results M by:

$$M_i = \frac{B_i^{'} + B_i^{''}}{2}. \tag{12}$$

D. Noise Eliminator

The final refinement of tumour segmentation relies on the convolutional CRFs [44]. Consider an input $\mathbf{M}$(the probability map from the output of the GN-CAM) with shape $[b, c, h, w]$ where b, c, h, w denote batch size, number of classes, input height and width respectively. Assuming that two pixels $u = (p, q)$ and $v = (p + dp, q + dq)$ come from two conditionally independent distributions, where p and q are the image coordinates. $d(u, v) > t$ is a restraint called Manhattan distance, where t refers to the filter size. All pixels with a distance greater than t have a pairwise potential of zero. The Gaussian kernel matrix k_g is defined as

$$k_g[b, dp, dq, p, q] = exp(-\sum_{u=1}^{d}\frac{\omega}{2\delta_u^2}). \tag{13}$$

Denoting $\mathbf{E} \in R^{b\times c\times h\times w}$ as the final output of the CAM visualizer. A Gaussian kernel g can be calculated based on feature vectors $e_1, ..., e_d$ by Equ.13. ω is defined as:

$$\left|e_u^{(d)}[b, p, q] - e_u^{(d)}[b, p - dp, q - dq]\right|^2, \tag{14}$$

where δ_i is a learnable parameter. For a set of Gaussian kernels $\{g_1 ... g_S\}$, S is the number of kernels. We define the global kernel matrix $G = \sum_{a=1}^{S} w_r \cdot g_r$. In the combined message passing of all S kernels, the result $\mathbf{M}$ defined as:

$$\mathbf{M}[b, c, p, q] = \sum_{dp,dq\le t}\mathbf{G}[b, dp, dq, p, q] \cdot \mathbf{E}[b, c, p + dp, q + dq] \tag{15}$$

So the final tumour segmentation for one whole side image is given as the matrix $\mathbf{M}$.

TABLE I

A SUMMARY OF THE DATA STATISTICS IN THE PRIVATE DATASET (BSS) AND THE PUBLIC DATASET (PAIP2019), INCLUDING THE NUMBER OF WHOLE SLIDE IMAGES, AND THE NUMBER OF IMAGE PATCHES.

Dataset	Tumour	# Training image	# Training patch	# Validation image	# Validation patch	# Testing image	# Test patch
BSS [45]	BCC	30	600k	10	200k	10	200k
	SP	30	600k	10	200k	10	200k
	SKC	30	600k	10	200k	10	200k
PAIP2019 [46]	Resection	30	750k	10	250k	10	250k
	Biopsy	0	0	0	0	9	225k

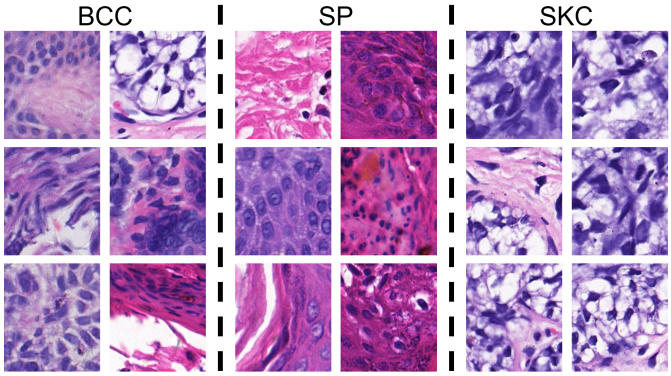


Fig. 4. Three types of tumour patch samples are extracted from the WSIs in the private BSS dataset, which contains basal cell cancer (BCC), the squamous papilloma (SP), and seborrheic keratosis cancer (SKC). These patches show a close look at tumour shapes and colours of different classes.

IV. EXPERIMENTS AND RESULTS

A. Data introduction

1) BSS dataset: The BSS dataset [47] is a private tumour dataset and has been adopted in our previous work. The BSS dataset contains 150 WSIs of squamous cell carcinoma including basal cell cancer (BCC), squamous papilloma (SP) and seborrheic keratosis cancer (SKC). All of the images on the BSS dataset are from the Second Affiliated Hospital of the Zhejiang University of China. It takes around 4 years to collect the dataset, make annotations and review the data in total. To protect the privacy of patients, all personal labels on scan images have been removed. For each scan, the invited experts roughly spent 60 minutes marking fine tumour labels (F-labels) and 5 minutes annotating coarse tumour labels (P-labels). After that, we invited 2 senior experts to spend one week checking whether all tumour regions are marked in the scans of the BSS dataset.

Several patch samples are cut from the WSIs and shown in Fig. 4. It is clear to observe that the whole tumour lesion is composed of multiple lobules as shown in Fig.4. Each lobule is covered with squamous epithelial cells. Fibrous vascular tissue in the centre is infiltrated with inflammatory cells. There is an obvious thickening of squamous epithelium, vacuoles in the cytoplasm, and an increase in goblet cells. Inside tumour cells, there was no obvious mitotic phase or nuclear heterogeneity.

2) PAIP2019 dataset: PAIP2019 dataset [45] has facilitated the development and benchmarking of cancer diagnosis and segmentation. This dataset contains 50 high-quality annotations for liver cancer WSIs and is first released by the PAIP Liver Cancer Segmentation Challenge, organised in conjunction with the Medical Image Computing and Computer Assisted Intervention Society (MICCAI 2019). Hepatocellular carcinoma (HCC) is a cancer of the internal organs. Most primary liver cancers are caused by hepatocellular carcinomas. There are a number of cellular and stromal components in HCC scans, including tumour cells, inflammatory cells, blood vessels, acellular stroma, tumour envelopes, fluids, mucus, or necrosis. This dataset contains all cases diagnosed between 2005 and 2018. All whole slide images were randomly arranged for training, validation, and test sets.

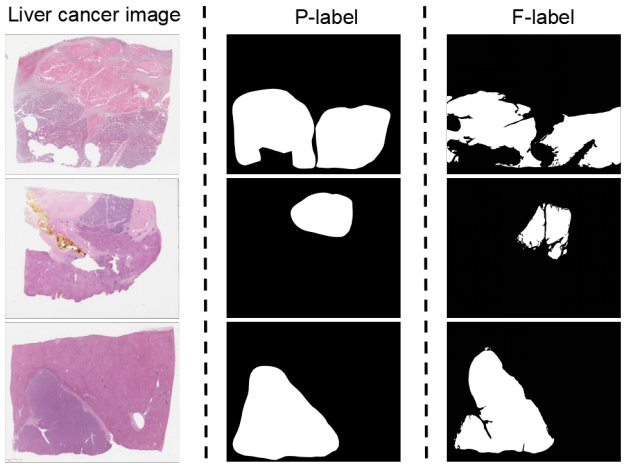


Fig. 5. Three whole slide image samples of the public dataset (PAIP2019 [45]). The first column is the original images; the second and third columns present the corresponding poor labels (P-labels) and the fine labels (F-labels) marked by experts.

The annotations published by PAIP2019 are of very high quality. However, in the course of clinical practice, the majority of tumour WSIs don't have precise annotations. In order to simulate this common situation, we invited two tumour experts who are also responsible for the marking of the private BSS dataset, to mark 50 WSIs with coarse annotations (P-label) for PAIP2019. The original high-quality annotations from the PAIP2019 dataset are used as F-labels in the evaluation process. The two versions of annotation are shown in Fig. 5.

To alleviate the potential impact of variations in coarse labels, we adopted a meticulous approach during the data preparation phase. Specifically, we enlisted the expertise of four histopathologists, who independently labeled the tumor datasets using coarse annotations. We carefully considered subjective independence and authoritative annotations by inviting four histopathologists to mark the BSS dataset and the remaining four experts to label the PAIP2019 dataset. The final coarse annotations were generated by aggregating the inputs

from all four experts, resulting in a more comprehensive and representative labeling scheme. In Section IV, we presented experimental results obtained from two different tumor datasets, namely the BSS dataset and the PAIP2019 dataset in Tables III and IV, respectively. These datasets encompassed a total of five tumor sub-categories, providing a diverse set of cases for evaluating the performance of our proposed method.

B. Evaluation matrics

The segmentation inference results are evaluated using Recall, Specificity, Accuracy, IOU and Dice. Assuming the positive sample is the tumour and the negative sample is the normal tissues. Six results are defined to demonstrate the relationship between ground truth and prediction results, including True Positive (TP), True Negative (TF), False Positive (FP) and False Negative (FN).

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

$$Specificity = \frac{TN}{TN + FP} \tag{17}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

$$IOU = \frac{TP}{FP + TP + FN} \tag{20}$$

$$Dice = \frac{2TP}{FP + 2TP + FN} \tag{21}$$

To verify the generalizability of our proposed method, the segmentation performances need to be inferred from both of the private BSS tumour dataset and the public PAIP2019 tumour dataset. We evaluate five performance metrics of four components in our proposed method.

C. Training details

In Tables II, III, and Table IV, we compare four state-of-the-art (SOTA) pixel-wise medical image segmentation methods and three classification-based image segmentation methods for the task of sketch-based tumor segmentation. The classification-based methods require the use of CAMs to visualize tumor distribution within patches, whereas the pixel-wise methods do not rely on CAMs. To ensure a fair comparison of these networks, we provide the necessary parameter settings for data preparation, model training, and inference.

For the data preparation in the experiments presented in Tables II, III, and Table IV, we randomly shuffled all Whole Slide Images (WSIs) from the BSS dataset. We allocated 20% of the WSIs for validation, 20% for testing, and the remaining 60% for training. The same data split ratio was applied to the resection scan of the PAIP2019 dataset. We saved all the patches in three sub-datasets for subsequent model training, validation, and testing. The number of split patches from each WSI ranged from 20,000 to 25,000 due to varying resolutions. Overall, the BSS dataset contained approximately 3 million patches, while the PAIP2019 dataset had 1.25 million patches. This demonstrates sufficient data support for model training and performance testing. For a detailed breakdown of the data preparation in the training, validation, and testing stages, please refer to Table I, which gives the details on the distribution of WSI images and patches in both datasets.

During model training, we utilized cross-entropy loss for all the networks listed in Tables II, III, and Table IV. The initial learning rate was set to 0.0001, and the SGD optimizer was employed. Each method was trained for 500 epochs using an Nvidia RTX 3080 GPU. To ensure consistent input sizes, we divided all images into (512, 512) patches, and a batch size of 64 was used.

For model inference, we tested all the trained models on 600k testing patches from the BSS dataset and 475K patches from the PAIP2019 dataset. A comprehensive comparison of segmentation performance for all classification-based models is presented in Table II. Table III provides a comparison of tumor segmentation performance for both pixel-wise and classification-based methods on the BSS dataset. Similarly, Table IV compares the tumor segmentation performances of these methods on the PAIP2019 dataset. It is important to note that the (GN-)CAMs shown in Tables III and IV were utilized during model inference and did not require additional training.

D. Sketch-based Tumour Segmentation Results on the BSS Dataset

1) The Annotation Refinement: We implement a series of ablation studies to figure out the gain when using our proposed annotation refinement module (AR) as shown in Tab. II. Experimental results show that the AR module is robust and efficient on two CAMs (CAM and GN-CAM) and three types of models (VGG, VF, and DCT). The base VGG+CAM+NE combination achieves a Specificity of 98.90 % after using the AR module. Especially, the segmentation precision increases from 83.97 % to 88.29 % after only adding the AR model to the base experimental configuration. Further, the AR module has more gains in Recall, IOU, and Dice metrics on the proposed GN-CAM visualizer compared to the base CAM. In detail, after using the AR module, the VGG+GN-CAM+NE combination achieves a Recall of 81.96 % (+8.64 %), an IOU of 70.83 % (+5.39 %), and a Dice of 82.62 % (3.92 %).

2) The Dual CNN-Transformer Classification network (DCT): The proposed DCT network is designed to classify whether a patch belongs to the ”tumour” class. Tab. II shows the qualitative comparison results between the existing methods and our proposed DCT. The proposed DCT network outperforms the compared VGG and VF methods, especially when using the GN-CAM module simultaneously. The best sketch-based tumour segmentation performances achieves Recall 84.44 %, Specificity 97.85 %, Accuracy 96.23 %, IOU 71.33 %, and Dice 83.12 %, which sets the experimental configuration of AR+DCT+GN-CAM+NE. A series of experimental evidence demonstrates the effectiveness of the proposed DCT

TABLE II

PERFORMANCE COMPARISON OF TUMOUR SEGMENTATION WITH DIFFERENT CLASSIFICATION NETWORKS, VISUALISER, ANNOTATION REFINEMENT (AR) AND THE NOISE ELIMINATOR (NE) IN THE BSS DATASET (%).

Model	CAM	AR	NE	Recall	Specificity	Accuracy	Precision	IOU	Dice
VGG [6]	CAM [43]	×	×	69.66	97.75	93.79	80.84	60.34	74.83
			✓	70.81	98.31	94.21	83.97	62.80	76.83
		✓	×	69.38	98.21	94.01	83.27	61.57	75.69
			✓	70.82	**98.90**	94.47	**88.29**	65.03	78.60
	GN-CAM(ours)	×	×	71.29	97.71	94.20	81.08	61.81	75.87
			✓	73.32	98.30	94.86	84.93	65.44	78.70
		✓	×	79.79	97.73	95.62	80.94	67.62	80.36
			✓	**81.96**	98.10	**96.18**	83.30	**70.83**	**82.62**
VF [47]	CAM [43]	×	×	73.73	97.67	94.61	78.27	62.64	75.93
			✓	76.44	97.70	95.00	77.28	64.32	76.86
		✓	×	78.89	97.79	95.59	81.79	67.56	80.31
			✓	80.56	98.32	96.16	85.07	71.00	82.75
	GN-CAM(ours)	×	×	76.80	98.06	95.22	82.60	66.40	79.59
			✓	79.45	**98.72**	95.95	**86.83**	71.16	82.97
		✓	×	83.00	97.61	96.02	80.75	69.50	81.86
			✓	**85.46**	98.00	**96.60**	82.95	**72.96**	**84.18**
DCT(ours)	CAM [43]	×	×	72.71	97.79	94.49	82.57	63.27	77.33
			✓	75.01	**98.79**	95.42	**88.51**	68.54	81.20
		✓	×	73.66	98.04	94.77	83.97	64.87	78.48
			✓	75.61	98.60	95.41	87.43	68.41	81.09
	GN-CAM(ours)	×	×	85.96	97.48	96.37	79.59	70.73	82.65
			✓	88.19	97.86	96.92	81.73	74.04	84.83
		✓	×	84.44	97.85	96.23	80.84	71.33	83.12
			✓	**88.28**	98.40	**97.08**	83.97	**76.68**	**86.69**

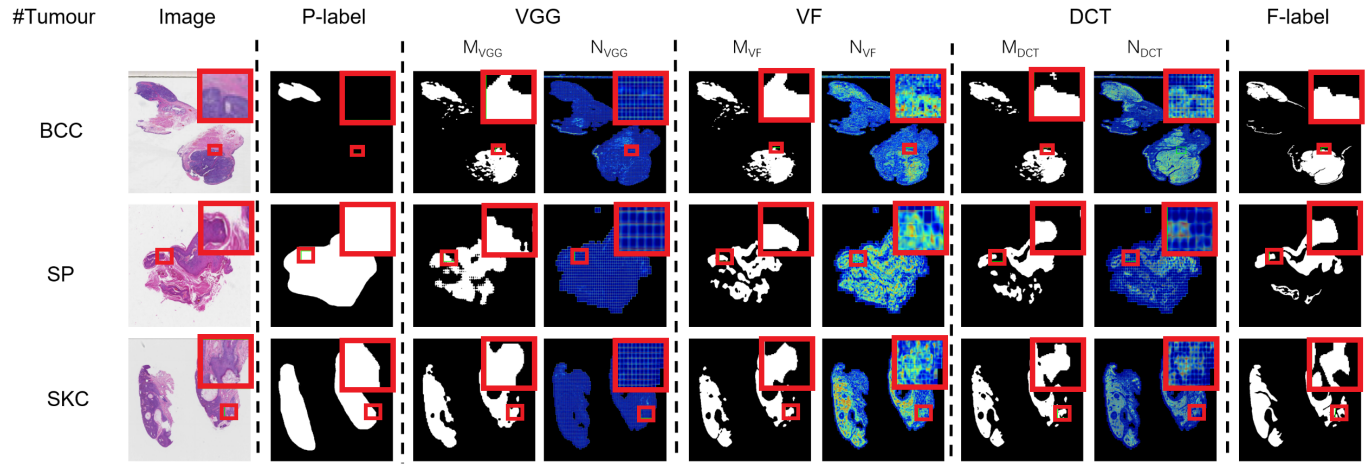


Fig. 6. Qualitative segmentation results of sketch-supervised based methods. Three types of binary tumour classification networks (VGG, VF and ours) are trained and tested along with AR, GN-CAM, and NE modules. Each type of network is trained three times with different types of images for distinguishing three types of skin tumours (BCC, SP and SKC). For one method i, M_i means the final tumour segmentation result and N_i is defined as the visualized heat map. The tumour segmentation performances of our proposed DCTGN-CAM method are closest to the fine labels annotated by experts. The upper left corner of each image presents enlarged tumour segmentation results.

for sketch-based tumour segmentation tasks. Compared with the recent VF-based method (VF+CAM) [47], our proposed method outperforms 14.55 % of Recall, 14.04 % of IOU and 10.76 % of Dice, which is a significant improvement on the BSS dataset.

Fig. 6 further illustrates the qualitative segmentation analysis among VGG, VF and our proposed DCT networks. We carefully draw three types or tumour predictions using three different methods, respectively. It can be found that the predictions of DCT are closest to the F-labels. For example, in the SKC WSI, the magnified image in the upper right corner shows that the enlarged tumour region looks like a ”horse” in the green-box-selected area of the F-label. It is clear to distinguish the shape and location of the ”horse” when using our DCT but is impossible to discriminate the ”horse” body in the predictions of VGG and VF methods. Qualitative results present that our proposed DCT is effective and outperforms other networks on sketch-based tumour segmentation.

3) GN-CAM: Inspired by the CAM concept, we design a GN-CAM module to represent local and global tumour features and to refine the tumour location on the patch level. We attempt to solve the non-negligible challenge for patch-based segmentation tasks: representing the global relationship between patches and the boundary of the generated WSI is not consistent. Our work involves generating and merging the output heat map patches by GN-CAM globally. Although

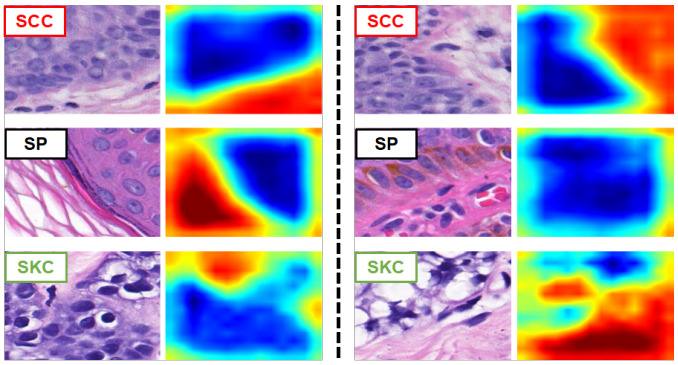


Fig. 7. Three types of predicted heat maps obtained by our proposed method DCTGN-CAM on the BSS dataset. The blue areas in the heat maps illustrate a relatively high probability of belonging to tumour regions.

we only train a binary classification network with sketch supervision, the proposed GN-CAM is capable to infer tumour location and boundaries precisely in the inference stage. Fig. 7 provides a close look at the predicted heat maps of tumour patches. Heat maps illustrate accurate tumour location and region information on three types of patches after using the proposed GN-CAM module. It proves that the GN-CAM module can illustrate precise tumour segmentation results based on the classification probabilities.

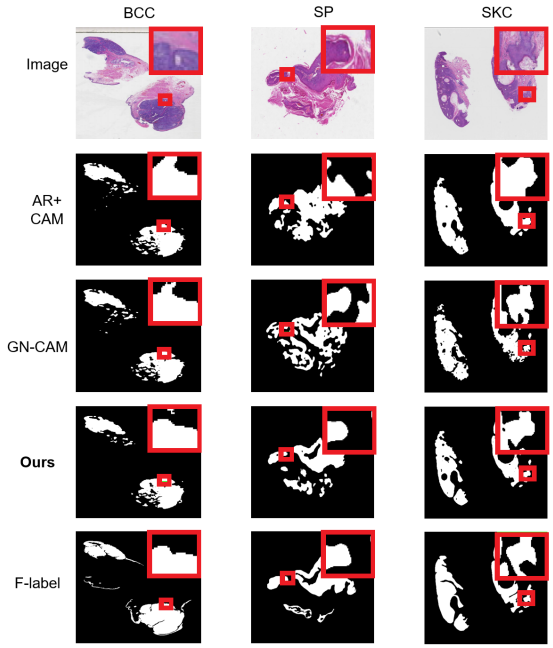


Fig. 8. Qualitative segmentation results of different visualizers in the sketch-supervised framework. Each method is trained three times with different types of images for distinguishing three types of skin tumours (BCC, SP and SKC). The tumour segmentation performances of our proposed DCTGN-CAM method are closest to the fine labels.

AR and GN-CAM modules complement each other clearly. Fig. 8 shows the segmentation improvement of AR and GN-CAM modules under the same DCT classification network. As compared with AR+CAM and GN-CAM, our segmentation results have fewer false-positive pixels than the GN-CAM, which means fewer normal tissues are classified as tumours due to the usage of AR. Furthermore, the boundaries of our predicted tumour regions are more precise than those predicted by AR+CAM, indicating the GN-CAM improves the boundary details effectively. Therefore, the AR and the GN-CAM work together to improve tumour segmentation performances.

4) The noise eliminator: Using patch-based segmentation has another challenge: jagged edges always appear at patch boundaries, and noise effects in non-tumour areas often look square or rectangular. As a result, taking threshold-based results as the final tumour segmentation results may lead to great visual errors. In our ablation study, the problem of noise is significantly relieved by the noise eliminator module. Fig. 9 shows the predicted patches focus on the noise eliminator. It is proven that the noise eliminator is responsible for eliminating false negative samples (filling voids within tumour tissue) and false positive samples (eliminating isolated noise areas outside the large tumour areas). Compared with the middle results directly from the output of the CAM, the final results processed by NE effectively improve the segmentation performances.

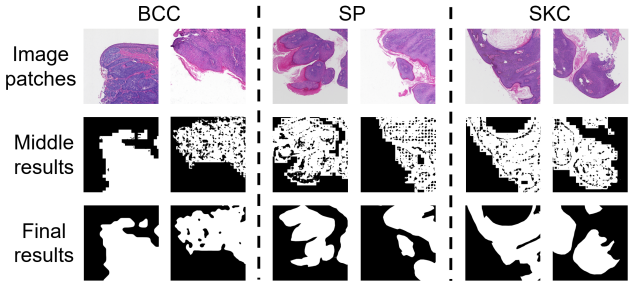


Fig. 9. Optimized segmentation results with the noise eliminator in our proposed DCTGN-CAM. The first row is the original image patches; The second row shows the middle results only processed by the binary threshold. The third row presents the final tumour predictions processed by the noise eliminator.

5) Comprehensive analysis: We compare our proposed method to the existing methods to analyze their performance comprehensively in the Tab. III. The P-labels are used to train the methods and the F-labels are used to evaluate them. Compared to CNN-based networks such as U-Net and AttnUNet, our method surpasses Transformer-based networks like SWinU-Net and hybrid CNN-Transformer networks, including TransUnet and MSHT, in terms of recall, specificity, accuracy, Intersection over Union (IoU), and Dice metrics. Furthermore, in contrast to the pure CNN methods U-Net and AttnUnet, the pixel-wise approaches presented in Tab. III, comprising U-Net, AttnUNet, TransUnet, and MSHT, demonstrate superior performance. This suggests that the global context feature encoding provided by the Transformer is more suitable for the sketch-based tumor segmentation task than pure CNNs. Furthermore, the highest performance metrics are achieved using classification-based segmentation methods, with 88.28% recall, 98.90% specificity, 97.08% accuracy, 76.68% IoU, and 86.69% Dice coefficient. Notably, our proposed methods account for four out of the five best performances As all learnable networks are supervised by coarse labels, the pixel-wise label will bring in large false-positive errors. In this case, segmentation-based methods like U-Net are ineffective.

TABLE III
TUMOUR SEGMENTATION PERFORMANCES ON THE BSS DATASET (%).

Auxiliary module	Network	Recall	Specificity	Accuracy	IOU	Dice
Pixel-wise segmentation methods						
	P-label	57.60	97.89	87.97	45.04	61.34
	U-Net [48]	47.90	98.09	85.91	44.21	60.83
	AttnUNet [49]	49.61	98.36	86.77	48.49	62.52
	TransUNet [9]	50.34	98.54	89.95	48.83	64.87
	SwinU-Net [50]	52.67	98.63	91.61	50.79	65.36
	MSHT [51]	54.84	98.77	92.05	52.21	66.79
Classification-based segmentation methods						
CAM [43]	VGG [6]	70.82	**98.90**	94.47	65.03	78.60
	VF [47]	80.56	98.32	96.16	71.00	82.75
	DCT (ours)	75.61	98.60	95.41	68.41	81.09
GN-CAM (ours)	VGG [6]	81.96	98.10	96.18	70.83	82.62
	VF [47]	85.46	98.00	96.60	72.96	84.18
	DCT (ours)	**88.28**	98.40	**97.08**	**76.68**	**86.69**

TABLE IV
SEGMENTATION PERFORMANCE ON THE PAIP2019 DATASET (%).

Network	Recall	Specificity	Accuracy	IOU	Dice
P-label	77.14	84.57	89.21	70.14	81.57
U-Net [48]	85.70	96.29	92.02	75.97	85.86
AttnUNet [49]	87.89	96.76	94.02	78.03	88.1
TransUNet [9]	89.28	96.73	94.44	80.45	89.23
SwinU-Net [50]	90.71	96.8	94.92	82.89	91.51
MSHT [51]	91.86	96.93	95.37	84.29	92.25
Ours	**95.42**	**97.23**	**96.57**	**89.10**	**94.23**

E. Sketch-based Tumour Segmentation Results on the PAIP2019 Dataset

Another ablation study is to compare our method with the non-classification method on the public PAIP2019 dataset, to further prove the efficiency of our methods. Tab. IV compares the P-label, U-Net and our method with F-label, respectively. U-Net is a pixel-wise method commonly used for tumour segmentation [48]. Experiment results show that U-Net has a 3-12% performance gain compared with P-label even if the U-Net training with only supervision on P-label. However, our method shows a large performance gain compared with U-Net when using the same experiment configurations. Specifically, our proposed method outperforms approximately 9% increase of Recall, 4% of Accuracy, 13% of IOU, and 12 % of Dice.

Although our proposed method has an exciting performance on the above private datasets, we still need to evaluate our method on the public datasets to verify the universality of our method. Compared with various types of skin cancers in the previous private dataset, the boundaries between the non-tumour tissues and tumour tissues are relatively smooth and hard to recognise. Tab. IV and Fig. 10 show the systematic evaluation results among the existing methods on the PAIP2019 dataset. It presents that our method (and our proposed modules) has better segmentation results on every evaluation metric. Our methods still obviously outperform the fully supervised network U-Net, proving the significant success of our methods on sketch-supervised tumour segmentation tasks. As all learnable networks are supervised by coarse labels, the pixel-wise label will bring in large false-positive errors. In this case, segmentation-based methods like U-Net are ineffective.

F. Clinic Application

It is noticeable that the P-label of the PAIP 2019 is a rough outline drawn based on the location of the tumour in the F-label. Similarly, the P-labels of the BSS dataset also have false positive samples only. Therefore, the premise of the good results for our method is the P-label should include all tumour regions.

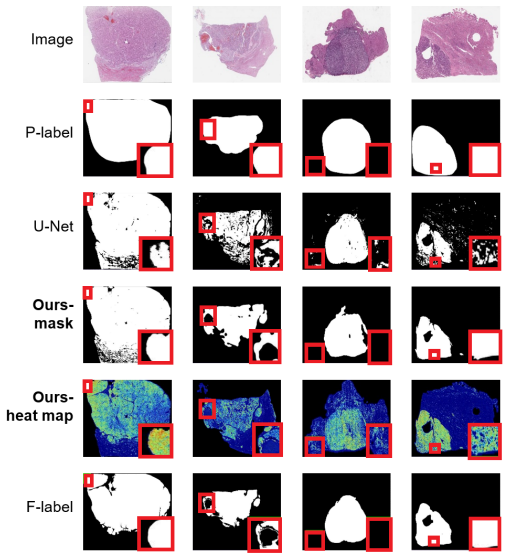


Fig. 10. Qualitative segmentation results on PAIP2019 dataset. Compared with the U-Net method, our method has more accurate tumour segmentation results in tumour boundaries for the sketch-supervised tumour segmentation task. The fifth row shows that the tumour regions have high responses after processing by our method.

It is possible to use our method in clinics as well. As a pathologist, all that needs to be done is to coarsely label the contours of the tumour boundaries in a short amount of time. The accurate tumour segmentation results can then be easily achieved by using the methods that we have suggested. Our work enables doctors to automatically obtain more accurate cancer segmentation results at a lower cost of labelling.

V. CONCLUSION

In this paper, we propose a framework for sketch-supervised tumour segmentation in histopathology, called DCTGN-CAM. Annotations from experts are optimized by calculating the intersection of cancer regions in unsupervised k-means and sketch annotations. The dual-branch DCT classification method leverages tumour features comprehensively. Parallel SWIN Transformer ensures the consistency of global feature representation. With a Global-Normalised CAM, a whole-slide heat map is generated from patch-based tumour classification predictions, which combine local and global normalization. A robust analysis of two tumour datasets shows that DCTGN-CAM is superior to weakly supervised tumour segmentation methods. This work is valuable and practical for computer-aided histopathology analysis. However, the multi-step design may cause influent feature flow or noise effects. To optimize this work in the future, an end-to-end approach might be more effective. Additionally, the adaptability of the front-end trainable model to the back-end CAM remains to be

studied. In the future, we will continue to optimize CAM visualization, lightweight the dual CNN-Transformer structure, and study the adaptability of CAM visualization in sketch-based segmentation tasks.

REFERENCES

[1] A. Paul and D. P. Mukherjee, "Mitosis detection for invasive breast cancer grading in histopathological images," *IEEE transactions on image processing*, vol. 24, no. 11, pp. 4041–4054, 2015.

[2] T. Vu, P. Lai, R. Raich, A. Pham, X. Z. Fern, and U. A. Rao, "A novel attribute-based symmetric multiple instance learning for histopathological image analysis," *IEEE transactions on medical imaging*, vol. 39, no. 10, pp. 3125–3136, 2020.

[3] M. U. Akram and A. Usman, "Computer aided system for brain tumor detection and segmentation," in *International conference on Computer networks and information technology*, pp. 299–302, IEEE, 2011.

[4] R. Dorent, S. Joutard, J. Shapey, S. Bisdas, N. Kitchen, R. Bradford, S. Saeed, M. Modat, S. Ourselin, and T. Vercauteren, "Scribble-based domain adaptation via co-segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 479–489, Springer, 2020.

[5] M. Zhang, Y. Zhou, J. Zhao, Y. Man, B. Liu, and R. Yao, "A survey of semi-and weakly supervised semantic segmentation of images," *Artificial Intelligence Review*, vol. 53, pp. 4259–4288, 2020.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[9] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *ArXiv*, vol. abs/2102.04306, 2021.

[10] X. Luo, M. Hu, W. Liao, S. Zhai, T. Song, G. Wang, and S. Zhang, "Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels supervision," in *MICCAI*, 2022.

[11] W. Cui, L. Zeng, B. Chong, and Q. Zhang, "Toothpix: Pixel-level tooth segmentation in panoramic x-ray images based on generative adversarial networks," *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1346–1350, 2021.

[12] Y. Li, Y. Wang, H. Zhou, H. Wang, G. Jia, and Q. Zhang, "Du-net based unsupervised contrastive learning for cancer segmentation in histology images," in *International Conference on Intelligent Robotics and Applications*, pp. 201–210, Springer, 2022.

[13] G. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-Van De Kaa, P. Bult, B. Van Ginneken, and J. Van Der Laak, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific reports*, vol. 6, no. 1, pp. 1–11, 2016.

[14] Y. Xu, Z. Jia, L.-B. Wang, Y. Ai, F. Zhang, M. Lai, E. I. Chang, *et al.*, "Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features," *BMC bioinformatics*, vol. 18, no. 1, pp. 1–17, 2017.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[16] Y. Li, Z. Xu, Y. Wang, H. Zhou, and Q. Zhang, "Su-net and du-net fusion for tumour segmentation in histopathology images," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 461–465, IEEE, 2020.

[17] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, p. 101693, 2020.

[18] W. Bai, H. Suzuki, C. Qin, G. Tarroni, O. Oktay, P. M. Matthews, and D. Rueckert, "Recurrent neural networks for aortic image sequence segmentation with sparse annotations," in *International conference on medical image computing and computer-assisted intervention*, pp. 586–594, Springer, 2018.

[19] Y. B. Can, K. Chaitanya, B. Mustafa, L. M. Koch, E. Konukoglu, and C. F. Baumgartner, "Learning to segment medical images with scribble-supervision alone," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 236–244, Springer, 2018.

[20] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[21] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 1529–1537, 2015.

[22] G. Valvano, A. Leo, and S. A. Tsaftaris, "Learning to segment from scribbles using multi-scale adversarial attention gates," *IEEE Transactions on Medical Imaging*, vol. 40, no. 8, pp. 1990–2001, 2021.

[23] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

[24] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

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

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

[28] J. He, Z. Deng, L. Zhou, Y. Wang, and Y. Qiao, "Adaptive pyramid context network for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7519–7528, 2019.

[29] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 603–612, 2019.

[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

[32] S. Graham, Q. D. Vu, M. Jahanifar, S. E. A. Raza, F. Minhas, D. Snead, and N. Rajpoot, "One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification," *Medical Image Analysis*, vol. 83, p. 102685, 2023.

[33] R. Schmitz, F. Madesta, M. Nielsen, J. Krause, S. Steurer, R. Werner, and T. Rösch, "Multi-scale fully convolutional neural networks for histopathology image segmentation: from nuclear aberrations to the global tissue architecture," *Medical image analysis*, vol. 70, p. 101996, 2021.

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

[35] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *Medical Image Analysis*, p. 102802, 2023.

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

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

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

[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

[40] T. Zhang, Y. Feng, Y. Zhao, G. Fan, A. Yang, S. Lyu, P. Zhang, F. Song, C. Ma, Y. Sun, *et al.*, "Msht: Multi-stage hybrid transformer for the rose image analysis of pancreatic cancer," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[41] Q.-D. Pham, H. Nguyen-Truong, N. N. Phuong, K. N. Nguyen, C. D. Nguyen, T. Bui, and S. Q. Truong, "Segtransvae: Hybrid cnn-transformer with regularization for medical image segmentation," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, IEEE, 2022.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

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

[44] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[45] Y. J. Kim, H. Jang, K. Lee, S. Park, S.-G. Min, C. Hong, J. H. Park, K. Lee, J. Kim, W. Hong, H. Jung, Y. Liu, H. Rajkumar, M. Khened, G. Krishnamurthi, S. Yang, X. Wang, C. H. Han, J. T. Kwak, J. Ma, Z. Tang, B. Marami, J. Zeineh, Z. Zhao, P.-A. Heng, R. Schmitz, F. Madesta, T. Rösch, R. Werner, J. Tian, E. Puybareau, M. Bovio, X. Zhang, Y. Zhu, S. Y. Chun, W.-K. Jeong, P. Park, and J. Choi, "Paip 2019: Liver cancer segmentation challenge," *Medical Image Analysis*, vol. 67, p. 101854, 2021.

[46] Y. J. Kim, H. Jang, K. Lee, S. Park, S.-G. Min, C. Hong, J. H. Park, K. Lee, J. Kim, W. Hong, H. Jung, Y. Liu, H. Rajkumar, M. Khened, G. Krishnamurthi, S. Yang, X. Wang, C. H. Han, and J. Choi, "Paip 2019: Liver cancer segmentation challenge," *Medical image analysis*, vol. 67, p. 101854, 2020.

[47] Y. Li, Y. Wang, L. Dong, J. Ye, L. Wang, R. Ge, H. Zhou, and Q. Zhang, "Light annotation fine segmentation: Histology image segmentation based on vgg fusion with global normalisation cam," in *International Workshop on Computational Mathematics Modeling in Cancer Analysis*, pp. 121–130, Springer, 2022.

[48] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[49] O. Oktay, J. Schlemper, L. L. Folgoc, M. J. Lee, M. P. Heinrich, K. Misawa, K. Mori, S. G. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," *ArXiv*, vol. abs/1804.03999, 2018.

[50] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," *ArXiv*, vol. abs/2105.05537, 2021.

[51] T. Zhang, Y. Feng, Y. Zhao, G. Fan, A. Yang, S. Lyu, P. Zhang, F. Song, C. Ma, Y. Sun, Y. Feng, and G. Zhang, "Msht: Multi-stage hybrid transformer for the rose image analysis of pancreatic cancer," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 4, pp. 1946–1957, 2023.