Expert Systems With Applications 216 (2023) 119452

Contents lists available at ScienceDirect

Expert Systems With Applications

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

SwinCup: Cascaded swin transformer for histopathological structures segmentation in colorectal cancer

Usama Zidan [a], Mohamed Medhat Gaber [a,b], Mohammed M. Abdelsamea [a,c,*]

[a] School of Computing and Digital Technology, Birmingham City University, Birmingham, B4 7XG, UK
[b] Faculty of Computer Science and Engineering, Galala University, Egypt
[c] Department of Computer Science, Faculty of Computers and Information, University of Assiut, Egypt

ARTICLE INFO

Keywords:
Transformers
Histology image analysis
Gland segmentation
Deep learning
Self-supervision

ABSTRACT

Transformer models have recently become the dominant architecture in many computer vision tasks, including image classification, object detection, and image segmentation. The main reason behind their success is the ability to incorporate global context information into the learning process. By utilising self-attention, recent advancements in the Transformer architecture design enable models to consider long-range dependencies. In this paper, we propose a novel transformer, named Swin Transformer with Cascaded UPsampling (SwinCup) model for the segmentation of histopathology images. We use a hierarchical Swin Transformer with shifted windows as an encoder to extract global context features. The multi-scale feature extraction in a Swin transformer enables the model to attend to different areas in the image at different scales. A cascaded upsampling decoder is used with an encoder to improve its feature aggregation. Experiments on GLAS and CRAG histopathology colorectal cancer datasets were used to validate the model, achieving an average 0.90 (F1 score) and surpassing the state-of-the-art by (23%).

1. Introduction

Colorectal cancer (CRC) is currently one of the most significant public health issues. The main pathway to diagnose CRC is through a pathology slide examination. Pathology is the space where a disease is detected or diagnosed using a series of microscopic slide images. These images exhibit a variety of characteristics and patterns. A key differentiator of these images is their sense of scale in comparison to normal natural images. A typical workflow would see a pathologist examine a tissue on a slide image, navigating through multiple regions to identify different tissue types. This navigation process is happening frequently per slide image, assessing the intrinsic details of single tissues and zooming out to get the general context of the patterns a tissue group is forming.

Processing microscopic images digitally facilitated many research opportunities in the field of computer-aided image analysis (Abdelsamea et al., 2022). This digitisation effort enabled the development of computational techniques for tissue segmentation, biomarker quantification, and tissue type classification in histology images. Fig. 1 shows example patches from different Whole Slide Images (WSIs) with very different characteristics and colour profiles. These images present a distinct series of challenges that differ from those found in standard image recognition problems. Utilising recent breakthroughs in image analysis and applying them directly to WSI proved challenging due to the high resolution of WSI in comparison to natural images (WSI can be 1000 times larger). As a result, the majority of approaches use a patch-based approach, extracting smaller patches from a WSI and classifying them on a smaller scale. A smaller patch image, on the other hand, contains less context of a wide range of texture patterns that may be beneficial for classification. A downsampled version of the input image is generally iterated through to enhance the receptive field while still maintaining image size constraints. Consequently, spatial resolution diminishes as a result of this approach. Therefore, segmentation performance from an input patch image is deemed challenging due to the trade-off between the extent of the field of vision and the resolution of the input image.

Even though the number of papers about machine learning (ML) methods used in CAD systems is growing, there are no many papers about detecting and classifying colorectal lesions from WSI at the same time. This puts colorectal cancer (CRC) behind diseases like breast and prostate cancer.

Despite the limited number of publications on colorectal WSI diagnosis, there are a variety of different articles focusing on CRC classification, leveraging information from smaller tissue patches that may be used as the foundation for comprehensive diagnostic methods. The

* Corresponding author at: School of Computing and Digital Technology, Birmingham City University, Birmingham, B4 7XG, UK.
E-mail addresses: usama.zidan@bcu.ac.uk (U. Zidan), mohamed.gaber@bcu.ac.uk (M.M. Gaber), mohammed.abdelsamea@bcu.ac.uk (M.M. Abdelsamea).

https://doi.org/10.1016/j.eswa.2022.119452
Received 26 April 2022; Received in revised form 16 December 2022; Accepted 20 December 2022
Available online 30 December 2022

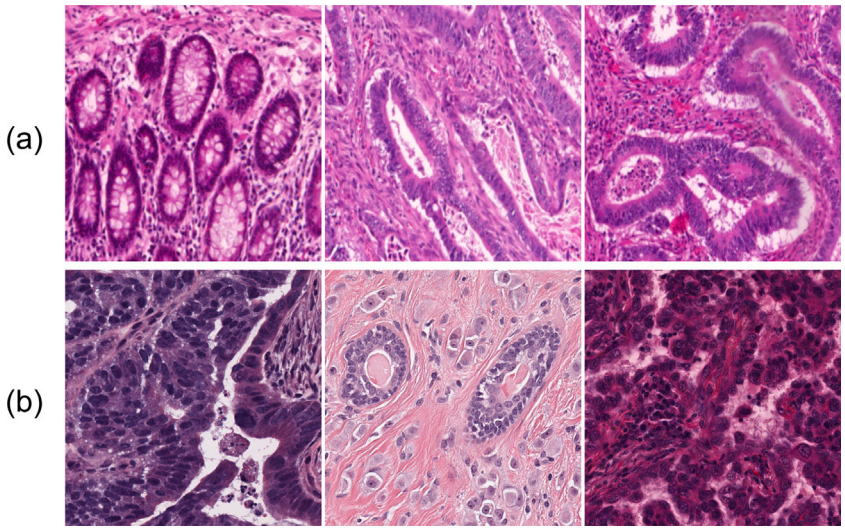

Fig. 1. Histology images show various complex Histopathological structures. As a result, developing learning algorithms for WSI presents unique challenges. WSI's huge gigapixel images constrained CNNs to widen their local bias. (a) different patches of colorectal cancer at stages T3 and T4. (b) H&E stained tissue from different organs with multiple nuclei scattered throughout the patch.

algorithms used to handle medical imaging tasks were based on mathematical models built by hand by subject matter experts. The convolution operator has long been dominant in the image processing field. Convolutional Neural Networks (CNNs) represent complicated images by stacking convolutional layers. With encoder–decoder architecture, CNNs can be used for semantic segmentation (Hussain et al., 2021; Ronneberger, Fischer, & Brox, 2015). However, these models lack a global understanding of local features. Several attempts to construct long-range associations by fusing intermediate features or employing a dilated convolutional kernel (Chen, Papandreou, Schroff, & Adam, 2017; Kushnure & Talbar, 2021; Zhou, Siddiquee, Tajbakhsh, & Liang, 2018). The use of CNNs for CRC diagnosis has been explored in the literature with varying degrees of detail. For instance, Iizuka et al. (2020) proposed a network based on the Inception-v3 model to classify H&E colorectal WSI into non-neoplastic, adenoma, and adenocarcinoma. A slide is divided into patches and classified using the network. Then, all patches are aggregated using an RNN to arrive at a final diagnosis. Song et al. (2020) produced an approach for segmenting colorectal adenomas using a modified DeepLab-v2 network. They introduced a series of skip connections that combine the upsampled lower layers with the higher layers. These results show impressive accuracy and high potential, yet they still suffer from shortcomings that need to be addressed. In medical imaging, image-processing experts have long tried to accommodate many of the challenging aspects that an image-processing pipeline encounters when images from the medical domain are the target. These include pre-training strategies to allow for the best transfer of knowledge to specific loss functions to supervise the boundaries of cells in an input image. CNNs still remain applicable to a subset of medical imaging where there is an availability of large labelled datasets. Moreover, CNN-based techniques are typically more difficult to interpret and frequently serve as "black box" solutions. Therefore, the medical imaging community has increased its efforts to combine the capabilities of hand-crafted and CNNs-based approaches, resulting in information-guided CNNs models. These efforts still fail to establish global structure relationships, which are critical for medical image segmentation.

Attention models have been widely used in a variety of domains in recent years, including image processing, speech recognition, and natural language processing. The primary idea behind the attention mechanism is to teach the network how to dismiss irrelevant information while focusing on vital information. The advances made possible by the attention mechanism enable neural networks to focus on important input features by leveraging their feature representations and focusing on task-relevant features. Additionally, by examining the attention maps, they enable us to interpret the world as seen by the neural network. Due to the widespread popularity of the attention mechanism, current research has focused on ways to improve its utilisation. The primary objective of attention is to locate features in images using an additional layer of weights that change the current representation. Then, through the process of learning, a neural network can determine which regions of the image demand attention. Numerous models have effectively improved performance by utilising this mechanism in a variety of ways. As an example of attention models, Attention U-net (Oktay, Schlemper, et al., 2018) introduced the incorporation of attention gates (AGs) to overcome the lack of global information. These gates can generate soft attention that suppresses irrelevant areas in the feature maps and focus on the salient features that are helpful to the task. Their light computational cost and ability to improve the sensitivity and accuracy of the model make them a useful addition to any model for dense label prediction. Yang and Yang (2023) recently showed a fusion of a CNN and Swin transformer model combined in a pyramid network for breast lesion segmentation. They designed an interactive channel mechanism to fuse the features from CNN layer and transformer layer, focusing only on the tumour-related regions and assigning high weights to these features. Their approach showed high boundary detection of breast lesions and high segmentation performance.

While several recent approaches have attempted to bridge the gap between local feature modelling and global context awareness in a variety of ways, the primary pitfalls that such models fall into are either an extension in the complexity dimension or a decrease in performance. When applied to the healthcare system, the major purpose of these models is to aid in the efficient and successful diagnosis of the target region. Models will need to be simple in architecture and very accurate in order to provide meaningful help to medical practitioners.

To alleviate the insufficiency of CNNs in global context modelling, this work offers a novel transformer-based model for the segmentation of histopathological structures in colon images, termed Swin Transformer with Cascaded UPsampling (SwinCup), to address the limitations of CNNs in global context modelling. SwinCup was designed to encode multi-scale global information while also incorporating a self-attention mechanism to handle the significant visual diversity of Histopathological features in colon images. SwinCup's encoder–decoder architecture is based on Swin Transformers and includes a cascaded upsampling decoder. We illustrate the performance of the model on two histology datasets as well as the effects of training on similar problem domains. SwinCup outperformed the state-of-the-art in segmenting

Histopathological structures, thanks to the global context represented by our model. In addition, we investigate the effects of supervised and self-supervised pipelines on the model's segmentation performance. We demonstrate an improvement in performance compared to previous supervised approaches, demonstrating the ability and performance of extracting labels to supervise self-supervision techniques and the importance of having pseudo labels during the self-supervision process. In summary, our main contributions are as follows:

- We build a Swin Transformer encoder block to model the global context of tumour-related regions
- We design a cascaded upsampler that utilises the supervised multi-scale features from the encoder to aid in the detection of the boundary of the tumour region
- We study the effects of different pre-training recipes and their effects on transformer-based methods using test downstream histology datasets.
- We demonstrate a modified self-supervised pipeline for training on WSI data

The paper is organised as follows: Section 2 discusses current advances in medical image analysis. Section 3 details the approaches employed, with a particular emphasis on the framework established for colorectal cancer segmentation. Finally, Sections 4 and 5 present the findings of this work, followed by a description of both the self-supervised and fully supervised pipelines.

2. Related work

Before the introduction of deep learning-based techniques, hand-engineered filers were used to segment and aggregate information from images. These techniques have shown great results (Dash et al., 2022; Subhan et al., 2022) yet are error-prone, time consuming and difficult to build with new or different datasets. Transformers have recently been able to challenge the state-of-the-art methodology of utilising convolutional layers to achieve the highest accuracy.

The Transformer architecture (Vaswani et al., 2017), which was initially used for natural language processing (NLP) tasks, has recently attracted considerable interest in the field of computer vision. An automatic cancer diagnosis is vitally important for physicians. Despite its significance, it is a particularly difficult task because of the lack of contrast between tumour and healthy tissues and the varying shape of the tumour during its progression. With the progress of imaging technology and machine learning algorithms, however, automatic diagnosis of malignant regions has become more practical. ViT (Dosovitskiy, Beyer, Kolesnikov, et al., 2021) models are especially well-suited for image processing due to their capacity to capture global relationships and discriminate numerous organs. Transformer-based network architectures designed for vision applications require extensive training data to be effective. In contrast to datasets for visual applications, the amount of data samples for medical imaging is significantly smaller. Studies have been made to utilise the locality of CNNs and the global awareness of transformers, leading to hybrid models. TransUNet (Chen et al., 2021) is one of the first transformer-based medical image segmentation architectures proposed. The encoder employs a CNN-Transformer hybrid architecture, with numerous upsampling layers in the decoder providing the final segmentation mask. Other models such as nnformer (Zhou, Guo, & Zhang, 2021) utilised a mix of two intertwined stems of convolution and self-attention to capitalise on the capabilities of both. By reversing the order of the convolutional embedding and transformer layers, they can exploit a low-level, high-resolution embedding created by a convolutional block and the long-range dependencies of high-level objects created by a transformer block.

TransBTS (Wang et al., 2021) is another attempt to exploit Transformer in 3D CNNs to aid in MRI brain tumour segmentation. The TransBTS model utilised a CNN as a feature extractor as a downsampling block that results in compact volumetric feature maps. These features capture a local context in the volume. A transformer block processes the features as tokens for global feature modelling. Combining the capabilities of Transformers and CNNs, hybrid architecture-based techniques adequately represent the global context and capture local information for precise segmentation. Although TransBTS is among the state-of-the-art transformer-based models, a direct comparison with SwinCup is not possible since TransBTS has been designed to segment 3D images.

Other methods process the input image information at one scale only and have seen widespread applications in medical imaging. For instance, a Transformer based network architecture has been proposed in Valanarasu, Oza, Hacihaliloglu, and Patel (2021), where Gated-Axial attention is used to train the model effectively on medical images. The authors propose parallel branches where each branch operates in a different context. Local patches can be attended to using a deeper branch while the global features can be captured using another branch. The Medical Transformer (MedT) utilities gated axial self-attention in the encoder for medical image segmentation without the need for pre-training. A key modification to the self-attention mechanism is the use of gates in the attention module to allow it to decide whether the learned positional encoding is informative about the current task based on how much data is available. The authors' two-branch approach processes images both on a local and global scale which intuitively mirrors a pathologist's workflow. The trade-off here lies in the model's requirement of a larger set of data to train it from scratch and given the large sizes of histological images, feeding it as an input to this network will prove difficult due to the huge overhead needed to downscale large resolution images to the required image size for this model.

2.1. Swin transformer

Vision transformer (ViT) (Dosovitskiy et al., 2021), while novel and versatile, suffers from two setbacks. Visual entities vary greatly across different scenes, leading to inconsistent performance. It also introduces quadratic complexity as the image resolution increases, as does the complexity of the model. Therefore, benefiting from a design that has worked for ages, authors of the Swin Transformer (Liu, Lin, Cao, et al., 2021) fitted the transformer block into the familiar deep CNNs architecture. The Swin Transformer proposes non-overlapping windows and a hierarchical architecture as an attempt to tackle the issues with the original ViT transformer.

Similarly to the beginning of ViT, the image is split into non-overlapping patches and embedded through a patch embedding layer. With 4 stages, the input resolution is downsampled at each stage and the number of patches inside each window is increased to form a hierarchical design, a mechanism that adds locality to the attention mechanism. To allow for global feature interaction, the windows are shifted at each transformer block, maintaining a context of the whole image. Moreover, the hierarchy allows the model to gain a better receptive field the deeper the depth of the network.

By iterating through the image on different, downsampled, scales and applying self-attention on local windows, the authors were able to model images in a similar style that convolutions do. Furthermore, Shifted Windowed Multi-headed Self-attention (SW-MSA) allows for local window interaction, giving the model both local attention to fine-grained features and global contextual awareness of the whole image.

DS-TransUNet (Lin et al., 2022) improved Swin-UNet (Cao et al., 2021) by adding an additional encoder to accommodate multi-scale inputs and introducing a fusion module to efficiently construct global dependencies between features of multiple scales using the self-attention method.

Swin Transformer model approaches the image problem from a different point of view than usual CNNs by working from 'outside-to-inside', capturing the global context, and narrowing attention layer by layer to capture the fine details. These contributions produce models

that work well on medical images, specifically histology images. A sub-domain where the pathologists usually start the examination by viewing slides globally and zooming in for details on the cellular level.

Hybrid architecture-based approaches (Chen et al., 2021; Lin et al., 2022; Yang & Yang, 2023) combine the strengths of transformers and CNNs to effectively model global context and capture local features for accurate segmentation. Yan et al. (2022) propose an Axial Fusion Transformer UNet (AFTer-UNet) that contains a computationally efficient axial fusion layer between the encoder and decoder. Zhou, Guo, and Zhang (2021) proposed nnFormer which surpasses Swin-UNet by over 7% (dice score) on the Synapse dataset (Landman et al., 2015). They interleaved convolutional layers with transformer layers in one encoder, fusing features at each block.

3. Methodology

Attention mechanisms have been utilised in many ways over the past few years, but recently, transformers are new neural network architectures that heavily rely on the concept of attention, specifically self-attention. A transformer layer contains a self-attention module and a feed-forward network. Each input token is multiplied by three weight matrices to produce the Query (q), Key (k), and Value (v) vectors. By taking the dot product of the query and the key vectors we attend the query vector to each feature in the key vector. We divide by the square root of the embedding dimension (C) and then apply a softmax layer. The result is a probability representing each token's contribution to the current token. The final step is to multiply the softmax result by the Value vector to get the new token representation. Thus, self-attention can be formulated as:

$$SA(q.k, v) = softmax(\frac{q \cdot k^T}{\sqrt{d_k}})v \tag{1}$$

where $\sqrt{d_k}$ is the dimensionality of the query/key–value sequence. Multiple heads of attention project the current embedding differently, resulting in distinct representations of the image, each with a unique attention result. This is done using by dividing the input sequence into different sub-groups, and performing attention operations on each sub-group. This multi-headed attention mechanism allows for global context between distinct features of the input sequence, as a result of the continual interaction between tokens. One issue with employing such a pipeline is the lack of order in the network. Before applying attention, transformers add a positional embedding. The model learns a pattern from these vectors that helps it determine the position of each word or the distance between words in a sequence.

Our framework's initial stage includes parsing WSIs and segmenting them into distinct patches. The patching process is preceded by automatic tissue segmentation to identify where in the WSI to start the patch extraction process. The WSI is reduced to low resolutions and converted to HSV colour space. A binary mask is produced for tissue areas by thresholding saturation of H&E stain. A few hyperparameters are required, most notably the size of the image patch and the number of patches to generate. The next step begins by segmenting each slide's tissue region and dividing it into several smaller patches (e.g. 256 × 256 pixels) that may be used directly as inputs to a fully convolutional autoencoder. Our primary pre-training datasets comprise around 250,000 images.

3.1. SwinCup architecture

The final model used in the downstream task is our proposed SwinCup model consisting of a classic encoder–decoder architecture where the main building blocks are an encoder, decoder, and skip connections. The encoder layers are all based on the Swin Transformer layers. These layers begin by splitting an image into small non-overlapping patches, each size of 4 × 4. The dimension of one patch then becomes 48 (4 × 4 × 3). The patches are then passed through a linear layer to project them in an arbitrary dimension (C). The standard transformer block attends each token to all other tokens, leading to quadratic complexity. However, in the Swin block of SwinCup, the flattened patches are re-projected to an image plane and a self-attention mechanism is applied in localised windows, each with M × M patches. This is known as window multi-head self-attention (W-MSA) and can be computed as:

$$\begin{aligned} z'_l &= \text{W-MSA}\ (LN(z_{l-1})) + z_{l-1}, \\ z_l &= MLP(LN(z'_l)) + z'_l \end{aligned} \tag{2}$$

Interaction between windows through shifting windows is what allows the features representation to be learned in a global context. The window layout in shifted window multi-head self-attention (SW-MSA) is shifted towards the upper-left of the image. This modifies the window layout so that each window can be made up of several other sub-windows while still retaining the same number of patches. The output of SW-MSA can be formulated as:

$$\begin{aligned} z'_{l+1} &= \text{W-MSA}\ (LN(z_l)) + z_l, \\ z_{l+1} &= MLP(LN(z'_{l+1})) + z'_{l+1} \end{aligned} \tag{3}$$

A patch merging mechanism is then used to downsample the input by reducing the number of tokens as the network goes deeper and increases the dimension of the feature maps, leading through multiple layers to a hierarchical feature representation. This process merges each group of 2 × 2 neighbouring patches followed by a linear layer on the concatenated features. Therefore, downsampling of 2x is achieved through patch merging.

The final product of the encoder is multi-level feature maps each in a smaller resolution than the one before but richer in higher-level features. These feature maps are fed into the decoder that processes them to get the final prediction of the network.

The decoder is a cascaded upsampler (CUP) which consists of multiple upsampling steps to decode the hidden feature and get the final segmentation mask. The decoder starts by cascading multiple upsampling blocks that start with the lowest resolution feature map, upsampling it, and concatenating it with the coming skip concatenating from the encoder. Each block consists of a 2x upsampling operator, a 3 × 3 convolutional layer, and a ReLU layer. Similarly to U-Net, we concatenate the decoder's features with a skip connection. Using CNNs, the cascaded upsampling approach is employed to recover the resolution from the previous layer. The encoder, for example, produces feature maps with the dimensions $\frac{H}{16} \times \frac{w}{16} \times D$. Then, to attain the full resolution, we use cascaded multiple upsampling blocks. Each block has two 3 × 3 convolution layers, a batch normalising layer, a ReLU layer, and a resolution of H × W as well as an upsampling layer. The combined performance of both the encoder and decoder form the traditional u-shaped architecture, enabling feature aggregation at different resolution levels via skip-connections, as seen in Fig. 2.

Our framework's last component includes running our pre-trained backbone on downstream tasks. We conducted this test using the GlaS and CRAG datasets. To test the effectiveness of our model on WSI colorectal datasets, we devised two pre-training schemes. We tested the model in a supervised fashion on the target dataset. We then

Table 1
Comparison results with state-of-the-art models on colorectal cancer segmentation datasets. Models were pre-trained using Imagenet weights. SwinCup shows great performance in segmenting colorectal cancer from slide patches.

Dataset	CRAG			GLAS		
Model	F1	Recall	Precision	F1	Recall	Precision
DeepLab	0.71	0.73	0.77	0.84	0.89	0.83
AttUnet	0.51	0.62	0.47	0.83	0.88	0.82
TransUnet	0.78	0.78	0.47	0.88	0.89	0.85
SwinUnet	0.67	0.63	0.63	0.79	0.71	0.71
SwinCup	**0.89**	**0.90**	**0.89**	**0.92**	**0.92**	**0.792**

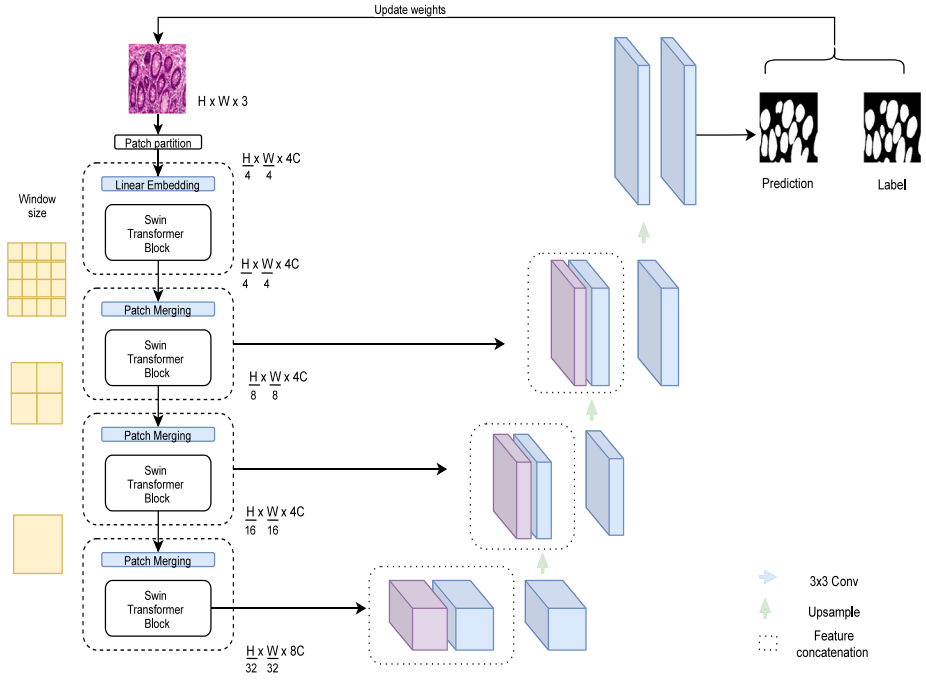

Fig. 2. An encoder divides an input image into tokenized patches that are recombined by the transformer layers. Each level applies self-attention in distinct windows that permit token interaction and accumulation. We employed a depth of [2, 2, 18] and a number of [4, 8, 16] attention heads. The last layer downsamples the image to fit within one self-attention window. The final segmentation mask is constructed by upsampling feature maps in a cascaded fashion.

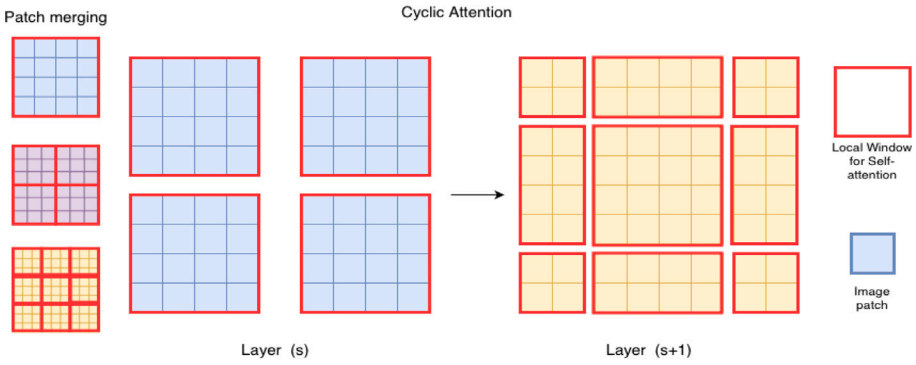

Fig. 3. Assuming an input size of 224 × 224 and a window size of 7 × 7. Each red square represents a window. Each window contains 49 patches. The patch merge mechanism downsamples the image and expands the field of attention to encompass the whole image. Cyclic window-attention moves the windows each layer to allow for cross attention between image regions.

experimented with different pre-training schemes and propose a new modified self-supervised pipeline that beats state-of-the-art techniques, evident in the downstream task (Table 3).

3.2. Pre-training

Self-supervised learning (SSL) is a class of unsupervised learning that has recently gained significant attention. SSL does not require labelled data but uses an auxiliary task to learn from unlabelled data. In particular, we employed a self-supervised pretraining scheme to learn from unlabelled data. We used the Swin Transformer architecture as the backbone for our model, and we pretrained the model on unlabelled data. We then fine-tuned the model on the downstream task of colorectal cancer histopathological structure segmentation. We transferred the parameters from the online branch of the self-supervised workflow to the SwinCup model and evaluated the model's performance with and without the pre-training procedure (Table 1). Additionally, comparisons with various self-supervised procedures were conducted.

We employed the swin transformer as our primary backbone in the self-supervision pipeline. We utilise Momentum Contrast v2 (MoCo v2) (Chen, Fan, Girshick, & He, 2020), a self-supervised learning technique that employs contrastive loss to create an embedding space that contains many augmented representations of the same image. A batch of augmented tiles is processed by two encoders. The contrastive loss then brings adjacent pairs of matching tiles closer together and separates adjacent pairs of differing tiles. Only the first of these two encoders are used to back-propagate gradients. The second encoder's weights are updated using an exponential moving average (EMA) of the weights of the first encoder. We modified the loss function to take into account the generated labels from the previous step. These labels allow the model to recognise that some samples outside the current image are of the same class and thus have to be mapped to a similar location in the embedding space. The original MoCo implementation lacks this aspect, as it assumes all data points in training are separate objects and need to be mapped separately. In Histopathology data, however, this assumption does not apply as different non-adjacent

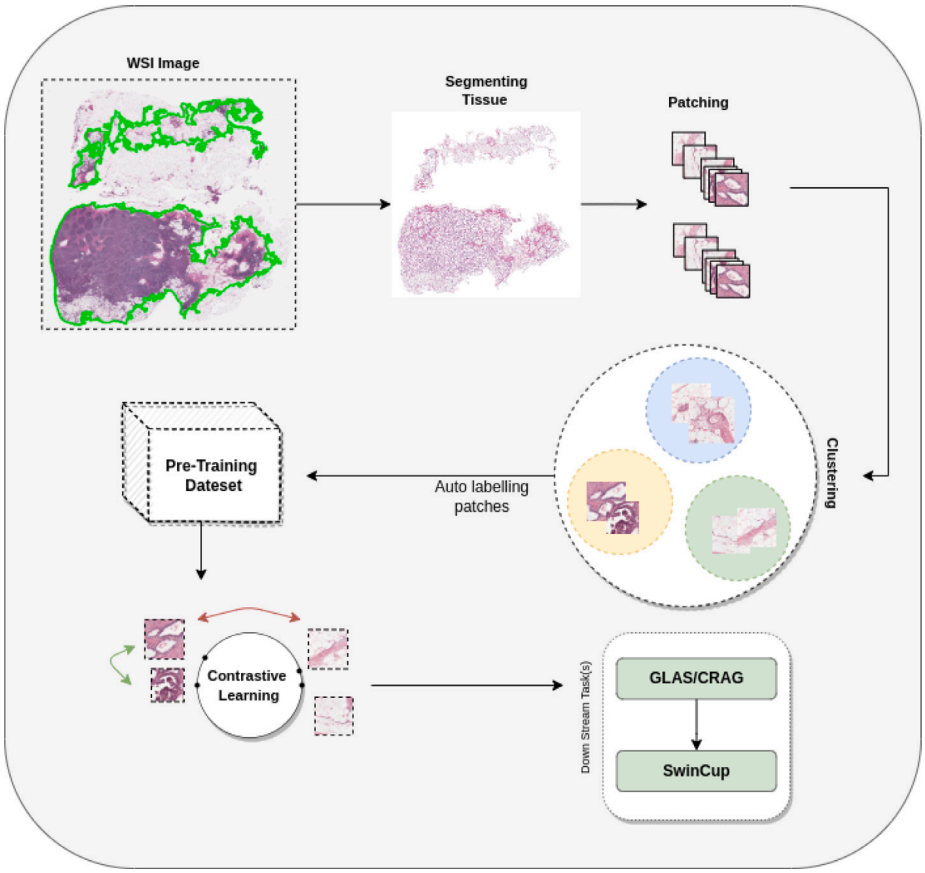

Fig. 4. An overview of the framework for pre-training models for histopathology images. The pre-training datasets used in this work are formed of large H&E stained slides which are used as initial training. The pipeline begins with acquiring WSI data and extracting patches. Then a clustering algorithm is applied to generate pseudo labels that are then fed into a self-supervised training process (pre-training phase). The downstream task begins with the pre-trained weights and fine-tunes them based on the new data.

patches can look different but be classified under the same class. To accommodate for this limitation, our loss function takes into account the augmented versions of the input, in addition to, a queue filled with other representations that are of similar (positive) tiles or other (negative) tiles. The model then learns an embedding space that has images of the same objects in the same cluster as their positive samples but apart from all the negative samples. (see Fig. 3).

4. Experiments

All models were implemented using the deep learning toolbox PyTorch (Paszke, Gross, Massa, et al., 2019), and all experiments were run on a single RTX 8000 GPU with 48 GB memory. The network is trained for 60 and 150 epochs, on GlaS and CRAG datasets, respectively, with a batch size of 8. Adam optimisation was used with β 1 and β 2 set to 0.99. A cyclic learning rate with a maximum learning rate of 0.00001 was used to facilitate convergence. The objective function used is a combination of Dice loss and Binary Cross-Entropy weighted equally. During training, data augmentations were used to prevent overfitting and improve the model's generalisation. The following augmentations were used: flipping, rotation, and changes in contrast and intensity, a binary mask was the output of the network with argmax applied to get the highest probability for each pixel.

4.1. Datasets

To evaluate the performance of our model, we utilised multiple Histopathological microscopy image datasets, both for pre-training and downstream testing. (see Figs. 5 and 6).

4.1.1. GlaS

GlaS (Sirinukunwattana, Pluim, Chen, et al., 2016) is a colon histology dataset created as part of the MICCAI'2015 challenge to advance automated techniques for quantifying gland morphology. It consists of 16 Histopathological slides of colorectal cancer at stage T3 or T4 from which 165 patch images were generated. Due to the fact that each sample is processed independently in the laboratory, significant variation in stain dispersion and tissue structures exists between subjects. The GlaS dataset is divided into two subsets into our studies: 132 images for training and 33 images for testing.

4.1.2. GRAG

Colorectal Adenocarcinoma Grading (CRAG) dataset contains 38 WSIs scanned with an Omnyx VL120 scanner with a pixel resolution of 0.55 μm/pixel (20 × objective magnification). The 38 WSIs are from different patients and are mostly of size 1512 × 1516 pixels, with the corresponding ground truth at the instance level. During training, the CRAG dataset (Awan et al., 2017) is split into 173 training images and 40 test images.

4.1.3. Pre-training data

During the pre-training phase, two different H&E stained datasets were used. The PESO (Bulten et al., 2019) dataset consists of 62 prostate cancer (PCa) whole-slide images at a pixel resolution of 0.48 mu/pixels, resulting in a total of 62 slides. Since the resolution of these slides is significantly higher than natural images, patch extraction is performed at a resolution of 1024 × 1024 of the whole slide. This was done to alleviate the computational cost and to have enough training images. The final version consisted of 15,000 image patches. We evaluate the patches based on their tissue percentage and colour

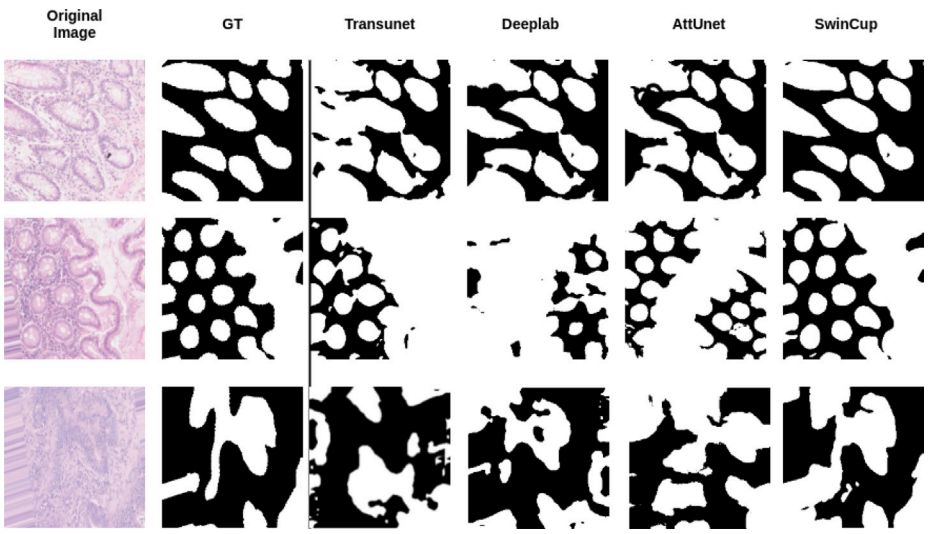

Fig. 5. Qualitative results of the different methods used tested on CRAG dataset. Differences can be seen between the convolution-based networks and the attention Transformers. CNNs perform well on similar-looking structures such as the ones in the first 2 rows but deteriorate as the gland shape changes dramatically. Global context-aware models can aggregate information properly given current and further context, as seen in the SwinCup results.

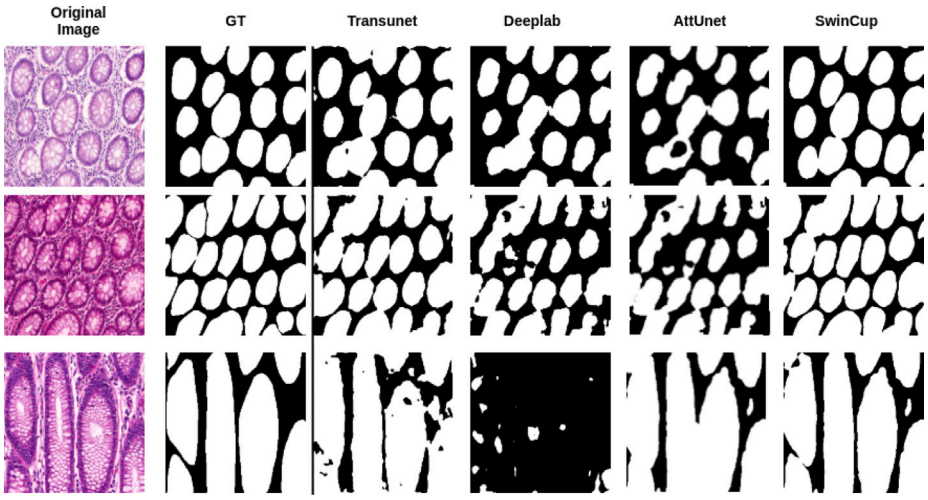

Fig. 6. Qualitative results of the different methods used tested on GlaS dataset. Similar differences can be seen between the convolution-based networks and the attention Transformers. Homogeneous gland shapes are easily captured but more complex shapes require a global awareness of the boundaries.

profile. Depending on the threshold in the algorithm, we may change how much tissue is in the patches, giving us more control over the number of images to generate. Tupac (Veta, Heng, et al., 2019) is a tumour proliferation dataset consisting of 500 breast cancer cases from The Cancer Genome Atlas. Each case is represented with one whole-slide image and is annotated with a proliferation score based on mitosis counting by pathologists, and a molecular proliferation score. We randomly selected 10 WSIs from the archive and began 'patch-extraction' to extract 256×256 patches from each slide. This resulted in around 249 thousand patches from training.

5. Discussion

The advantage of the transformer-based neural network is that its global self-attention mechanism enables the learning of a more flexible attention structure. It allows models to have a more nuanced understanding of the relationships between the tokens in the sequences on which they are training. When one token attends to another in the context of a specific sequence, it indicates that they are closely linked and have influence on one another. The effectiveness of utilising transformer models on medical images can be shown in the results of the experiments presented in this study. This contributes to current efforts to apply image transformers to the vision domain and enables us to investigate image transformer-based modelling.

Global awareness is critical in medical image analysis. Histopathology images contain a high degree of variability in terms of patterns, which necessitates that models be exposed to a vast quantity of data in order to learn parameters that are invariant to subtle changes and more aware of the general structure of objects in the image. Individual cells may not reflect a great deal of information, but when grouped together, they can highlight a more general pattern that is picked up by a pathologist and given a label to describe their state.

Windowed-Multi Head Self Attention (W-MSA) solves the inherit complexity problem in self-attention by reducing the computation to equally sized regions in the image. This is done using two critical aspects in Swin transformers. In the beginning, self-attention computes attention interaction between various tokenized patches in fixed-size windows. This method makes information aggregation more dynamic

in specific areas and makes the attention computations linear as the number of patches within a given window is constant. Furthermore, dynamically learning region-specific weights allows for greater problem-specific adaptation. Second, sliding windows enable models to attend to many regions simultaneously, aggregate information from other locations on the image, and assign weights based on these aggregated values.

To demonstrate the effectiveness of our proposed architecture on colorectal semantic segmentation, we report results on gland segmentation in Table 1. In general, SwinCup demonstrated competitive performance against a mixture of convolution-based and hybrid models in testing. Additionally, we verified our method's efficacy on the CRAG dataset, demonstrating overall higher performance in contrast to other approaches and emphasising our method's strong generalisation capacity across diverse datasets. The multi-scale self-attention mechanism iterates through the image, interacting with many windows at each level to capture the image's long-range patterns. Transformer-based networks benefit from this process since they require fewer layers to attain global context-awareness.

The purpose of attention is to accumulate information that is pertinent to the current point of view. Utilising such mechanisms on a pixel-by-pixel basis provides the advantage of being aware of surrounding structures and their effect on the current representation. This results in a high computational complexity, which is addressed by windowed attention calculations. By focusing attention on a small number of windows, we imitate the convolutional process and emphasise local feature details. Shifted-window attention enables inter-window interaction, reinforcing the learning process with global awareness. At each step, the windows expand to encompass a larger portion of the image, providing a more complete perspective of the growing structures. Given that attention is capable of estimating the convolutional kernel, SwinCup benefits from the ability to learn local and global characteristics that are relevant to the present task. This notion is shown in Fig. 1, where several cellular architectures are depicted. A local context is required to distinguish the border of the gland from other cellular structures, while a global view is required to assess the overall shape of the gland.

Previous studies (Zeiler & Fergus, 2014) have demonstrated that features in a network have a hierarchical structure, with lower levels capturing local characteristics such as corners or edge/colour patterns, while higher layers tend to capture more complicated patterns that can an object or parts of an object. Although the use of high-level features can help identify glands and other cellular structures, it does not contribute much to low-level attributes such as texture.

Fig. 5 shows qualitative results of the test set in the CRAG data set. A notable difference is seen in the DeepLab network as it performs well on the similar-looking structures such as the ones in the first 2 rows but deteriorates as the gland shape changes dramatically. This supports the notion that a more global context-aware model can aggregate information properly given the current and further context.

To confirm the effectiveness of our model, we used the Paired t-test as a statistical significance assessment to determine if there is a statistically significant difference between the test results obtained before and after the application of our model. The paired t-test determines the significance of the difference between two populations when the distribution of the differences between the samples is not normal. In the first test dataset (CRAG), a p-value of 0.0101 was obtained in the F1 metric, a p-value of 0.0066 in the recall metric, and a p-value of 0.0129 in the precision metric. Furthermore, we observed a p-value of 0.0026 in F1, a recall p-value of 0.0195, and a precision p-value of 0.0012 in the GLAS dataset. All reported p-values are less than 0.05, indicating that we reject the null hypothesis. This indicates that the true mean of the test findings between the two populations is not equivalent. In other words, SwinCup significantly increases the effectiveness of the underlying model.

Table 2
Downstream results of the self supervised pipeline. Multiple Histopathology datasets were used to examine the impact of the size of data in pre-training on the downstream CRAG dataset.

Dataset	Dice	Precision	F1	Recall	#Classes
ImageNet (200)	0.76	0.86	0.85	0.84	200
Medmnist (32k)	0.71	0.83	0.81	0.80	91
Medmnist (290k)	0.71	0.82	0.81	0.79	91
ImageNet	0.69	0.82	0.80	0.79	1000
TUPAC	0.69	0.82	0.80	0.78	4
PESO	0.67	0.81	0.79	0.77	3
Medmnist (15k)	0.67	0.80	0.78	0.76	15

Table 3
Comparing SwinCup performance when pre-trained using other self-supervised pipelines and the proposed pipelines in Section 3.2. Our proposed pipeline shows great improvements owing to the modified loss function and model architecture.

Model	GLAS			CRAG		
	F1	Recall	Precision	F1	Recall	Precision
Byol	0.58	0.65	0.74	0.61	0.64	0.73
Random	0.60	0.67	0.60	0.60	0.66	0.60
MOCO V2	0.75	.78	0.80	0.75	0.76	0.79
SimCLR	0.74	0.75	0.81	0.74	0.75	0.80
SwinCup	**0.77**	**0.77**	**0.82**	**0.77**	**0.80**	**0.81**

5.1. Supervised pre-training on similar data

Table 4 demonstrates the effects of pre-training pipeline layout in Fig. 4. Using weights from a model trained on the ADE20k dataset gives a performance boost (20%) compared to random initialisation (71%), although the ADE20k dataset is a scene parsing dataset that has been highly annotated (Table 4). Despite the fact that these weights are learned from a completely different application area with drastically different patterns than histology, it performs fairly well when compared to the random initialisation. This supports the notion that more fine-grained datasets lead the model to develop a diverse array of feature extractors that can transfer to other tasks. Furthermore, this is also supported when high modality similarity between the pre-training and target datasets exist, which is the case with the PESO dataset (92%) (Taher, Haghighi, Feng, Gotway, & Liang, 2021; Wen, Chen, Deng, & Zhou, 2021).

5.2. Self-supervising swin transformers

To show the efficacy of our self-supervised pipeline, we pre-trained various swin transformer models on the PESO and TUPAC dataset and evaluated performance on a downstream task of colorectal semantic segmentation (Table 3). The transformer model is trained during pre-training to map matching inputs to the same space in the embedding dimension. The mechanism through which this mapping occurs varies among all evaluated frameworks. This enables us to evaluate the ability of various mapping mechanisms and determine if the addition of pseudo labels improves the pre-training process.

Table 3 shows the average results from cross-validation runs on all pre-trained models on the CRAG dataset. We see an improvement in using our pseudo-labels in the pre-training manifest in the downstream task. Other self-supervised pipelines rely on mapping single images and their augmentation to the same representation.

Self-supervision enables the ability to learn low-level task-agnostic features that can be easily adapted to multiple tasks without requiring large amounts of annotations. Even though compared to supervised techniques self-supervision does not fair well, we hypothesise that this is due to the directly learned high-level features, in supervised pipelines, that are more domain-specific.

Fine-grained datasets have been shown to aid self-supervised pipelines in learning more adaptable parameters. The key differentiating

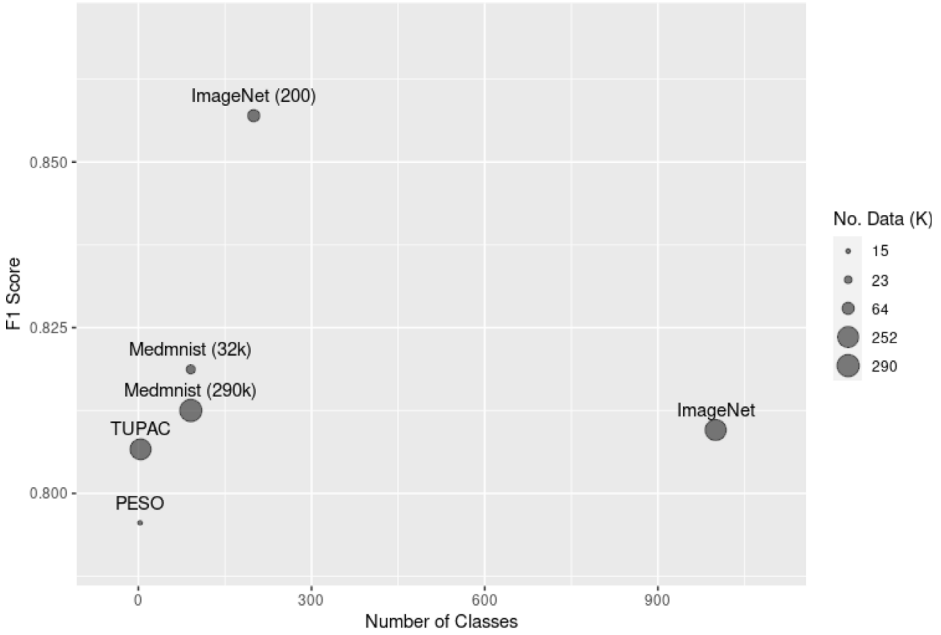

Fig. 7. Self-supervised pre-training on multiple medical image datasets of different sizes compared to ImageNet. Generally, ImageNet pre-training is still the best option but, as seen above, the full ImageNet can be a deterrent to performance. This can be due to the very high number of classes making the transferable weights outside of the downstream task domain.

Table 4
A summary of the effects of different pre-training datasets in the supervised pipeline. Downstream GlaS data.

Pre-training dataset	F1	Recall	Precision
None	0.71	0.74	0.78
ADE20K	0.89	0.9	0.9
PESO	**0.92**	**0.92**	**0.92**

Table 5
A summary of the effect of different decoders used with the Swin Transformer encoder. Tested on GlaS datasets.

Model	F1	Recall	Precision
SwinCup w/UpperNet	0.83	0.84	0.85
SwinCup w/Cup	0.90	0.91	0.89

aspect between histology images and natural images, such as the ones in ImageNet, is the variability in the 'scene' in which common objects are found in the image. While natural images are composed of different objects interacting together with various coarse details apparent in the images that make the boundaries between objects more clear, medical images emphasise a more subtle pattern behaviour where visual differences between subordinate classes are often subtle and deeply embedded within local discriminative parts. These patterns are usually small and coupled with local variations in texture which is what a pathologist is trained to recognise. These subtle patterns make it harder for the model to recognise without enough exposure to diverse datasets that allow the model the ability to learn descriptive feature extractors that can detect more diverse and subtle patterns. We can see an improvement in performance with more diverse data added. Fig. 7 illustrates the performance changes when changing the pre-training dataset. ImagNet still remains the most impactful on results when fine-tuning the model. We can also report that more diverse medical data improve performance, with more classes contributing mainly to the difference in performance. The full ImageNet dataset contains 1K+ classes and performs worse than the 200-class counterpart. This suggests a cutoff threshold where too much out-of-domain data can result in 'over-fit' features that do not contribute to the downstream task.

5.3. Ablation study

During the process of this study different decoders were tested to evaluate the feature aggregation of the Swin transformer extracted features. Table 5 shows a summary of different decoders experimented with.

5.3.1. UPerNet decoder

UpperNet head is designed based on a Feature Pyramid Network (FPN) (Lin, Dollar, Girshick, et al., 2017), which exploits a top-down architecture to extract multi-level feature representations in an inherent and pyramidal hierarchy. The FPN module has shown great performance in scene-parsing datasets yet when it comes to medical image segmentation, it seems that the complexity of the module hinders its performance.

5.3.2. Cascaded UPsampler

The efficiency of cascaded upsampler is remarkable given its simplicity, with feature maps simply being upsampled through bilinear interpolation rather than time-consuming deconvolution or transformer-based modules that add additional computational complexity. Its effectiveness and simplicity inspired this design for SwinCup. Using the cascaded upsampler boosts the performance by 2% over the FPN module.

5.3.3. Size of pre-training data

Table 2 shows the results of self-supervised pre-training on different datasets and their effect on the downstream task. We observe an increase in performance when using diverse datasets (such as ImageNet). MedMnist (Yang et al., 2021) also demonstrates an advantage over pure histology datasets (PESO & TUPAC). This is in line with work in literature (Wen et al., 2021) where more fine-grained data sets in the pre-training allow more transferable weights to be learned by the model.

These results illustrate a narrowing gap between self-supervised and supervised learning. It remains to be seen what effects these modifications will have on other medical domains and modalities. There is an obvious limitation here due to domain gabs and a difference in input

data configurations such as CT. Thus, this remains a potential direction that can be studied in the future.

5.4. Conclusion

Medical tasks are accompanied by a lack of huge amounts of labelled data, due to the expensive nature of such endeavours. Machine learning pipelines that can harness other resources to allow for a more robust implementation in a clinical workflow are essential. This work attempts to bridge the gap between the clinical and technical world of machine learning by presenting a framework for training transformer-based models on challenging medical datasets. In this work, we present SwinCup, a hierarchical Swin Transformer encoder–decoder-based pipeline for histological Structures Segmentation in colorectal cancer. We demonstrate the effect of a global context approach to medical images and emphasise the impact of pre-training on similar domains for pathology-related problems. Experiments on colorectal slide images show that SwinCup outperforms other state-of-the-art methods in gland segmentation. The simple cascaded decoder used in this work demonstrates the effective results of hierarchical feature extraction derived from self-attention in the medical domain. In our experiments, we show how self-supervised learning boosts performance by 20% compared to random initialisation without a need for expert labelling. In-domain pre-training (92%) has been found to give better results compared to out-of-domain pre-training (89%). It should be noted that the models developed in this study can be easily extended to other medical datasets and will be left for future work to investigate the implementation of SwinCup in different modalities. SwinCup provides a framework that integrates multi-level local features with global awareness of the learned structures. The experimental results and performance on the test datasets demonstrate the model's performance. This methodology helps medical practitioners diagnose target locations in a quick and accurate manner. By applying the model in both supervised and self-supervised pipelines, we also demonstrate a recently developed direction in the literature on learning techniques. In conclusion, we presented a model capable of learning distinctive characteristics from histological slides with great performance and efficient architecture. Furthermore, we introduced a modification to the learning strategy that pushes self-supervised learning closer to supervised learning performance.

CRediT authorship contribution statement

Usama Zidan: Methodology, Software, Validation, Visualization, Writing – original draft. **Mohamed Medhat Gaber:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Mohammed M. Abdelsamea:** Conceptualization, Methodology, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Abdelsamea, M. M., Zidan, U., Senousy, Z., Gaber, M. M., Rakha, E., & Ilyas, M. (2022). A survey on artificial intelligence in histopathology image analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *12*(6), Article e1474.

Awan, R., Sirinukunwattana, K., Epstein, D. B. A., Jefferyes, S. D. R., Qidwai, U. A., Aftab, Z., et al. (2017). Glandular morphometrics for objective grading of colorectal adenocarcinoma histology images. *Scientific Reports*, *7*.

Bulten, W., Bándi, P., Hoven, J., Loo, R. v. d., Lotz, J., Weiss, N., et al. (2019). Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard. *Scientific Reports*, *9*(1), 1–10.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., et al. (2021). Swin-Unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537.

Chen, X., Fan, H., Girshick, R., & He, K. (2020). Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021). Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Chen, L. -C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

Dash, S., Verma, S., Bevinakoppa, S., Wozniak, M., Shafi, J., & Ijaz, M. F. (2022). Guidance image-based enhanced matched filter with modified thresholding for blood vessel extraction. *Symmetry*, *14*(2), 194.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An image is worth 16 × 16 words: Transformers for image recognition at scale. In *9th International conference on learning representations*.

Hussain, R., Karbhari, Y., Ijaz, M. F., Woźniak, M., Singh, P. K., & Sarkar, R. (2021). Revise-Net: Exploiting reverse attention mechanism for salient object detection. *Remote Sensing*, *13*(23), 4941.

Iizuka, O., Kanavati, F., Kato, K., Rambeau, M., Arihiro, K., & Tsuneki, M. (2020). Deep learning models for histopathological classification of gastric and colonic epithelial tumours. *Scientific Reports*, *10*(1), 1–11.

Kushnure, D. T., & Talbar, S. N. (2021). MS-UNet: A multi-scale unet with feature recalibration approach for automatic liver and tumor segmentation in CT images. *Computerized Medical Imaging and Graphics*, *89*, Article 101885.

Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., & Klein, A. (2015). MICCAI multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI multi-Atlas labeling beyond cranial vault—Workshop challenge*: *vol. 5*, (p. 12).

Lin, A., Chen, B., Xu, J., Zhang, Z., Lu, G., & Zhang, D. (2022). DS-TransUNet: Dual swin transformer U-Net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1–15.

Lin, T., Dollar, P., Girshick, R., et al. (2017). Feature pyramid networks for object detection. In *2017 IEEE conference on computer vision and pattern recognition* (pp. 936–944). Los Alamitos, CA, USA: IEEE Computer Society.

Liu, Z., Lin, Y., Cao, Y., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision* (pp. 9992–10002). IEEE.

Oktay, O., Schlemper, J., et al. (2018). Attention U-Net: Learning where to look for the pancreas. CoRR arXiv:1804.03999.

Paszke, A., Gross, S., Massa, F., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems* (pp. 8024–8035).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Sirinukunwattana, K., Pluim, J. P. W., Chen, H., et al. (2016). Gland segmentation in colon histology images: The GlaS challenge contest. CoRR arXiv:1603.00275.

Song, Z., Yu, C., Zou, S., Wang, W., Huang, Y., Ding, X., et al. (2020). Automatic deep learning-based colorectal adenoma detection system and its similarities with pathologists. *BMJ Open*, *10*(9), Article e036423.

Subhan, F., Aziz, M. A., Khan, I. U., Fayaz, M., Wozniak, M., Shafi, J., et al. (2022). Cancerous tumor controlled treatment using search heuristic (GA)-based sliding mode and synergetic controller. *Cancers*, *14*(17), 4191.

Taher, M. R. H., Haghighi, F., Feng, R., Gotway, M. B., & Liang, J. (2021). *A systematic benchmarking analysis of transfer learning for medical image analysis*. DART/FAIR@MICCAI.

Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., & Patel, V. M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. In *Lecture notes in computer science*: *vol. 12901*, *Medical image computing and computer assisted intervention* (pp. 36–46). Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Veta, M., Heng, Y. J., et al. (2019). Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge. *Medical Image Analysis*, *54*, 111–121.

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., & Li, J. (2021). Transbts: Multimodal brain tumor segmentation using transformer. In *International conference on medical image computing and computer-assisted intervention* (pp. 109–119). Springer.

Wen, Y., Chen, L., Deng, Y., & Zhou, C. (2021). Rethinking pre-training on medical imaging. *Journal of Visual Communication and Image Representation*, *78*, Article 103145.

Yan, X., Tang, H., Sun, S., Ma, H., Kong, D., & Xie, X. (2022). After-Unet: Axial fusion transformer Unet for medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3971–3981).

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., et al. (2021). MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. arXiv preprint arXiv:2110.14795.

Yang, H., & Yang, D. (2023). CSwin-PNet: A CNN-swin transformer combined pyramid network for breast lesion segmentation in ultrasound images. *Expert Systems with Applications*, *213*, Article 119024.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In D. J. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Lecture notes in computer science*: *vol. 8689*, *Computer vision - ECCV 2014 - 13th European conference* (pp. 818–833). Springer.

Zhou, H., Guo, J., & Zhang, Y. (2021). nnFormer: Interleaved transformer for volumetric segmentation. CoRR arXiv:2109.03201.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018). UNet++: A nested U-Net architecture for medical image segmentation. In *Lecture notes in computer science*: *vol. 11045*, *Deep learning in medical image analysis and multimodal learning for clinical decision support* (pp. 3–11). Springer.