IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2022.Doi Number

Polyp Segmentation with the FCB-SwinV2 Transformer

Kerr Fitzgerald[1], Jorge Bernal[2], Aymeric Histace[3], Bogdan J. Matuszewski[1]

[1] Computer Vision and Machine Learning (CVML) Group, University of Central Lancashire, Preston, United Kingdom
[2] Computer Vision Center and Computer Science Department, Universitat Autònoma de Barcelona, Barcelona, Spain
[3] ETIS UMR 8051, CY Paris Cergy University, ENSEA, CNRS, Cergy, France

Corresponding author: Kerr Fitzgerald (e-mail: kffitzgerald@uclan.ac.uk).

This work was supported by the Science and Technology Facilities Council (Grant Number: ST/S005404/1) and by MCIN/AEI/10.13039/501100011033 (Grant Numbers: PID2020-120311RB-I00 and RED2022-134964-T)

ABSTRACT Polyp segmentation within colonoscopy video frames using deep learning models has the potential to automate colonoscopy screening procedures. This could help improve the early lesion detection rate and in vivo characterization of polyps which could develop into colorectal cancer. Recent state-of-the-art deep learning polyp segmentation models have combined Convolutional Neural Network (CNN) architectures and Transformer Network (TN) architectures. Motivated by the aim of improving the performance of polyp segmentation models and their robustness to data variations beyond those covered during training, we propose a new CNN-TN hybrid model named the FCB-SwinV2 Transformer. This model was created by making extensive modifications to the recent state-of-the-art FCN-Transformer, including replacing the TN branch architecture with a SwinV2 U-Net. The performance of the FCB-SwinV2 Transformer is evaluated on the popular colonoscopy segmentation benchmarking datasets Kvasir-SEG, CVC-ClinicDB and ETIS-LaribPolypDB. Generalizability tests are also conducted to determine if models can maintain accuracy when evaluated on data outside of the training distribution. The FCB-SwinV2 Transformer consistently achieves higher mean Dice and mean IoU scores when compared to other models reported in literature and therefore represents new state-of-the-art performance. The importance of understanding subtleties in evaluation metrics and dataset partitioning are also demonstrated and discussed.

Code available: https://github.com/KerrFitzgerald/Polyp_FCB-SwinV2Transformer

INDEX TERMS Medical image processing, Polyp segmentation, Deep learning, SwinV2, Transformer

I. INTRODUCTION

Colorectal cancer is the second lead cause of cancer-related deaths worldwide. In 2020, more than 930,000 deaths occurred due to colorectal cancer with more than 1.9 million cases being diagnosed. It is estimated that by 2040 there will be 3.2 million new cases of colorectal cancer and that the number of deaths will increase to 1.6 million [1]. Colorectal cancer often arises from small benign polyps which progress over time to become malignant. Colonoscopy is widely considered as the gold standard among polyp screening and removal procedures. The procedure is performed using a colonoscope, a long, flexible tube with a camera and light at the end. The colonoscope is inserted through the patients rectum and into the colon, allowing clinicians to navigate through the colon and visually inspect targeted regions for abnormalities in real time. Additionally, the colonoscope can have an instrument channel which allows surgical tools to remove identified polyps, a procedure known as polypectomy.

Colonoscopy procedures do have limitations as studies estimate that between 17% and 28% of polyps are missed [2] [3] [4]. Missed polyps can significantly impact patient health and it is predicted that improving polyp detection rates by 1% would reduce the risk of colorectal cancer development by approximately 3% [5]. High demands on healthcare systems are also increasing the pressure and workloads placed upon colonoscopy clinicians [6]. Computer aided systems to support clinicians in improving the detection rate and characterization of polyps have therefore undergone significant research in recent years. Due to their excellent performance, deep learning models now dominate this research area.

The goal of polyp segmentation in colonoscopy images is to accurately identify and delineate polyps from the surrounding healthy tissue of the colon. Accurate segmentation of polyps allows for a comprehensive assessment of a polyp's texture, shape and relative size which can be crucial in assessing the polyp's malignancy or potential to develop into a malignancy. However, variability in patients, colonoscopy procedures (e.g. position and angle of colonoscope) and polyp morphologies cause images of polyps to differ in shape, size, color and texture. The task of automatic polyp segmentation using deep learning models remains challenging due to this variability and is exacerbated by the limited availability of polyp image databases.

In recent years, deep learning models for the semantic segmentation of polyps have predominantly been composed of Convolutional Neural Networks (CNNs) or Transformer Networks (TNs). Subsequently, hybrid semantic segmentation models which combine the benefits of both CNN and TN architectures have been developed. Representative examples of current state-of-the-art models for polyp segmentation are the Fully Convolutional Branch-TransFormer (FCN-Transformer) [7] and DUCK-Net [8]. The DUCK-Net model is a Fully Convolutional Network (FCN) whilst the architecture of the FCN-Transformer combines the benefits of both TNs and CNNs by running a model of each type in parallel and combining the outputs which are then passed onto a prediction head for processing. The TN architecture used in the FCN-Transformer is the Pyramid Vision Transformer Version 2 (PVTv2) [9].

Motivated by the aim of improving the performance of polyp segmentation models and their robustness to data outside of the training distribution, we propose a new CNN-TN hybrid model named the FCB-SwinV2 Transformer. This model has been created by making extensive modifications to the FCN-Transformer [7]. These include changes to the Fully Convolutional Branch (FCB) of the FCN-Transformer (including the use of an increased number of channel dimensions and a residual post normalization approach) and the replacement of the PVTv2 model within the Transformer Branch (TB) with a SwinV2 [10] based U-Net. The reasoning behind the TB replacement is due to the unique shifted window-based self-attention mechanism employed by SwinV2 models. This mechanism excels in capturing complex hierarchical structures and should help to capture relevant information across various polyp morphologies to improve segmentation performance.

Our main contributions include:

- A novel CNN-TN hybrid deep learning model named the FCB-SwinV2 Transformer, created by making extensive modifications to the previous state-of-the-art FCN-Transformer [7].
- A performance comparison (including generalizability testing where possible) of the FCB-SwinV2 Transformer with high performing models on popular colonoscopy segmentation benchmarking datasets including Kvasir-SEG [11], CVC-ClinicDB [12] and ETIS-LaribPolypDB [13]. The FCB-SwinV2 Transformer achieves state-of-the-art performance.
- An examination of common issues within polyp segmentation literature relating to dataset partitioning and averaging methodologies used to calculate performance metrics. Experimental proof of the critical importance of such issues is provided.

II. Related Work

This section provides an overview of the relevant work on the semantic segmentation of polyps. Fully Convolutional Networks (FCNs), Transformer Networks (TNs), and CNN-TN hybrid architectures are described. Summaries on recent state-of-the-art models are provided.

A. Fully Convolutional Networks (FCNs)

One of the most influential deep learning models for medical image segmentation is U-Net [14]. The original U-Net model was a Fully Convolutional Network (FCN) which consisted of an encoder and decoder. The encoder used in the original U-Net is a CNN. CNNs consist of various stacked layers, each designed to sequentially process the input data. These layers apply filters through convolution operations, use pooling to reduce dimensions, and employ activation functions to introduce non-linearities. This structure effectively extracts and refines features at each stage. The hierarchical nature of CNNs allows the encoder to synthesize abstract representations and patterns, capable of representing specific shapes or entire objects. The decoder uses transposed convolutions which recover the spatial dimensions of the image by up-sampling the compressed feature maps from the encoder. Skip connections link decoder layers to corresponding encoder layers, thereby reintroducing the spatial information lost during down-sampling. This design enables the decoder to effectively utilize both high-level and low-level features from the encoder, ensuring accurate reconstruction of segmentation maps. U-Net was designed as a 'one-stage' model, where images are directly processed to produce a segmentation map. This contrasts with two-stage [15] models where regions are first identified and then further analyzed for semantic segmentation. Recent state-of-the-art polyp semantic segmentation models follow a one-stage approach for efficiency and direct processing capability.

Since the introduction of U-Net, numerous FCNs for semantic segmentation of polyps have evolved from the original U-Net architecture [8] [16] [17] [18] [19] [20] [21] [22] [23]. These models often incorporate advanced

ImageNet pre-trained CNNs into the encoder, significantly improving the efficiency and accuracy of feature extraction by leveraging pre-learned image representations. Examples of high performing FCN models are summarized below.

The Parallel Reverse Attention Network (PraNet) [22] employs a parallel partial decoder to extract high-level features and generate a global map for initial segmentation guidance. It then uses a reverse attention module to mine polyp boundary information and employs a recurrent cooperation mechanism which iteratively refines the segmentation by aligning the initial predictions with polyp boundaries. The Multi-Scale Residual Fusion Network (MSRF-Net) [23] employs unique Dual-Scale Dense Fusion (DSDF) blocks to allow multi-scale information exchange and maintain high resolution. MSRF-Net is therefore able to capture both high-level and low-level features which results in the prediction of accurate segmentation maps. The current state-of-the-art FCN polyp segmentation model is DUCK-Net [8]. This model uses the innovative DUCK convolutional block which can apply various filter sizes in parallel. This allows adaptive selection of the most effective filter size for each network stage. This allows the general localization of polyps whilst precisely delineating their boundaries. In contrast to many other FCN segmentation models, DUCK-Net does not use any form of encoder pre-training, demonstrating the high power of its feature extraction capabilities.

FCNs do have inherent drawbacks which can limit their performance for polyp segmentation. Due to their limited receptive field size, FCNs can have a reduced understanding of the global contextual information contained within an image. This can cause FCN models to struggle with scale variability and can lead to poor generalization performance.

B. Transformer Networks (TNs) and Hybrid Models

The introduction of the Vision Transformer (ViT) [24] revolutionized computer vision research by using a Transformer Network (TN) to conduct image classification. Like CNNs, TNs are composed of stacked layers that sequentially process input data. However, TNs apply self-attention mechanisms instead of convolution operations. TNs typically work by splitting images into a number of fixed-size patches. The patches are then flattened and positional embeddings are added to ensure the spatial relationship between patches is maintained. Layers of the transformer network use the self-attention mechanism to calculate attention scores for all pairs of image patches. This allows the network to assess the relative importance of each patch with regard to every other patch, allowing the network to capture global contextual information across entire images. As information progresses through successive layers, the TN is able to focus on different aspects of the image patches simultaneously, enhancing its feature extraction capability and understanding of relationships between image regions. The self-attention mechanisms ability to capture global contextual information and examine specific image regions enhances the TNs ability to understand the shape and texture of regions which may be relevant for polyp segmentation.

Since the introduction of the ViT, numerous polyp semantic segmentation models which are composed of fully TN based [25] [26] [27] or CNN-TN hybrid architectures [7] [28] [29] [30] [31] [32] have been developed. Many of these models employ a U-Net inspired encoder-decoder style structure. CNN-TN hybrid models are now commonly used as these can help overcome the limitations of using pure TN models. The main such limitation of pure TN models is that the lack of engineered feature extraction processes (when compared to FCNs) means that pure TNs typically require large amounts of training data. This is problematic for polyp segmentation due to the small sizes and limited availability of polyp image databases. Examples of high performing CNN-TN hybrid models are summarized below.

Polyp2Seg [33] uses the Pyramid Vision Transformer Version 2 (PVTv2) [9] as an encoder for multi-scale feature extraction. For each encoder stage, extracted features are passed into Compression Modules (CMs) to reduce the channel dimensions to a consistent size. The compressed features are then passed into Feature Aggregation Modules (FAMs) to directly combine lower-level and higher-level features. A Multi-Context Attention Module (MCAM) is also applied on the lowest-level feature maps to enhance the capture of low-level information, such as polyp texture and color. ESFPNet uses the Mix Transformer (MiT) [25] as an encoder. For each encoder stage, extracted features are passed into an Efficient Stage-wise Feature Pyramid (ESFP) decoder. The decoder generates linear predictions for each output stage and then linearly fuses the processed features together. Intermediate processed features are also concatenated with the previous decoder layers intermediate processed features. This allows the model to progressively integrate global features from later layers with local features from earlier layers to construct comprehensive feature maps. The FCN-Transformer [7] employs the unique approach of having a Transformer Branch (TB) and a Fully Convolutional Branch (FCB) which run in parallel. The TB uses the PVTv2 as an encoder which passes features to an enhanced Progressive Locality Decoder (PLD). The PLD features advanced local emphasis (LE) and stepwise feature aggregation (SFA) modules. The FCB is composed of modern residual blocks and encourages the extraction of features required for processing outputs of the TB into full-size (i.e. matching ground truth resolution) segmentation predictions.

This paper proposes extensive modifications to the FCN-Transformer to further improve polyp semantic segmentation performance.

III. FCB-SwinV2 Transformer Model Design

A. Overview of existing SwinV2 Models

The SwinV2 Transformer [10] was developed to tackle issues with training stability and resolution gaps between pre-training and fine-tuning that arose with the original Swin Transformer [34]. In the original Swin Transformer, training instability was caused by large activation output amplitude discrepancies in different layers of the network. To solve this problem the authors of the SwinV2 Transformer developed an approach called 'residual post normalization'. In this approach, the output of residual blocks is normalized before merging into the main branch of the network. This was shown by the SwinV2 creators to cause much milder activation amplitudes than in the original pre-normalization configuration. The authors of the SwinV2 Transformer also replaced the original dot product attention mechanism due to the finding that learnt attention maps of some blocks were frequently dominated by only a few pixel pairs. The dot product attention mechanism was therefore replaced by a mechanism which computes the attention logit of a pixel pair using a scaled cosine function. Since the cosine function is naturally normalized, this helps the network become more insensitive to the amplitude of activations.

The SwinV2 Transformer network first partitions an RGB input image into non-overlapping patches. Each patch is a concatenation of the RGB pixel values which are subsequently passed to a linear embedding layer which projects them to have an arbitrary channel dimension. Patch merging layers then concatenate neighboring patches before passing them through a linear layer. This reduces the number of patches by a factor of 2 and increases the channel dimension by a factor of 2. The output of the patch merging layer is then passed through several successive SwinV2 transformer blocks. The successive SwinV2 transformer blocks are shown in Figure 1.

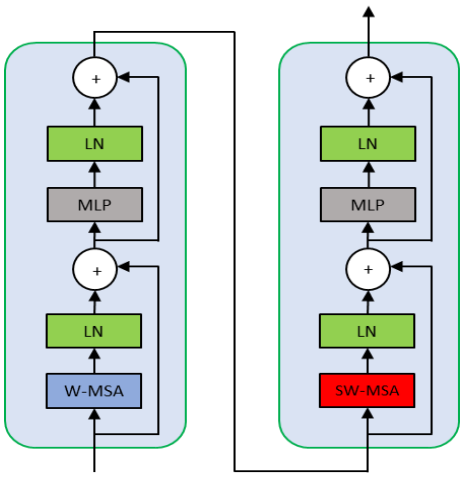


Figure 1: Two successive SwinV2 Transformer Blocks [13]. The residual post normalization configuration ensures layer normalization is conducted after attention layers and MLP layers.

These blocks consist of either window based multi-head self-attention (W-MSA) (first block) or shifted window partition multi-head self-attention (SW-MSA) (second block) layers, followed by an MLP with GELU [35] activation layer and stages of layer normalization (LN). Residual connections are also present within the block.

B. FCB-SwinV2 Transformer Architecture

The overall architecture of the FCB-SwinV2 Transformer is shown in Figure 2.

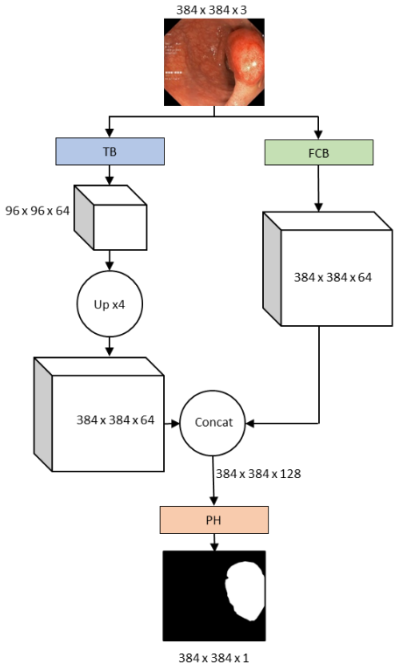


Figure 2: Overall FCB-SwinV2 Transformer architecture.

The TB of the original FCN-Transformer is replaced in this work by a SwinV2 U-Net architecture. SwinV2 models apply shifted window-based self-attention mechanisms which excel in capturing complex hierarchical structures. It was hypothesized that this would help extract relevant information across various polyp morphologies to improve segmentation performance and generalizability to data outside of the training distribution. Empirical evidence of improved segmentation performance is reflected by the ADE20K [36] segmentation benchmark. SwinV2 models are capable of achieving 59.9%mIoU [10] whilst PVTv2 models achieve 48.7% [9]. This substantial improvement in a general segmentation challenge strongly indicates that SwinV2 can enhance polyp segmentation, both in terms of accuracy and reliability.

The SwinV2 U-Net style architecture used in this work was based on a model [37] which used a Swin encoder [34] [38] with decoder blocks composed of 'Spatial and Channel Squeeze and Excitation' (SCSE) modules [39] [40] and

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3376228

standard convolution modules. The SCSE module is composed of a 'Spatial Squeeze and Channel Excitation' (SSCE) module and a 'Channel Squeeze and Spatial Excitation' (CSSE) module. The SSCE module takes an input tensor and reduces the spatial dimension using global average pooling. The resulting tensor is passed through convolutional and activation layers before performing element-wise multiplication with the original input tensor. This results in an output tensor with adaptively re-weighted channel values. The CSSE module takes the input tensor and reduces the channel dimension using convolution. The resulting tensor is passed through an activation layer before element-wise multiplication with the original input tensor. This results in an output tensor with adaptively reweighted spatial features. The SCSE module combines the outputs of the CSSE and SSCE modules using element-wise summation, therefore maximizing information propagation through the network at a pixel and channel level simultaneously. Element-wise summation is chosen over concatenation in order to avoid increasing tensor dimensions and therefore model complexity. The structure of the decoder block and SCSE module is shown in Figure 3.

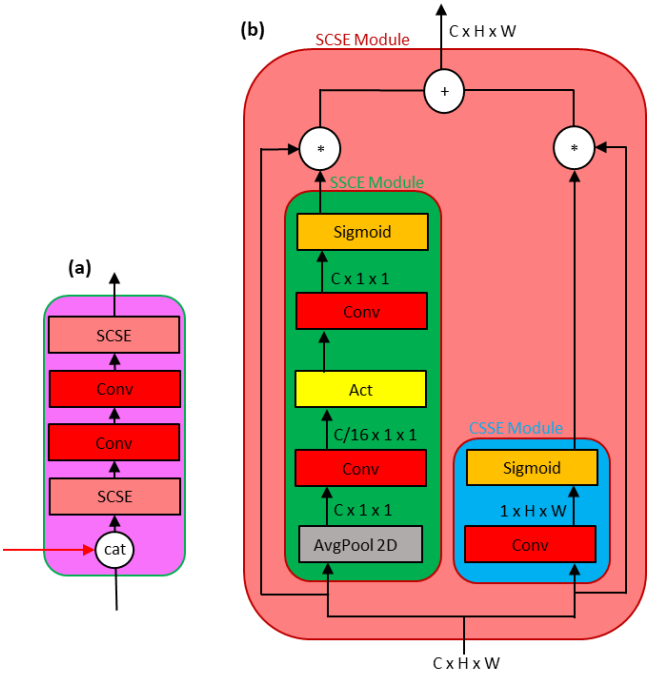


Figure 3: (a) The decoder block [35] uses channel wise concatenation to combine previous decoder layer output with encoder skip connection output. (b) The structure of the SCSE module which combines the output of the SSCE and CSSE modules.

The resolution of input images into the SwinV2 encoder model (and overall FCB-SwinV2 Transformer model) used in this work is 384x384 due to the availability of ImageNet [41] pre-trained SwinV2 encoder models available within the PyTorch Image Model library [38]. The SwinV2 U-Net architecture used as the TB is displayed in Figure 4.

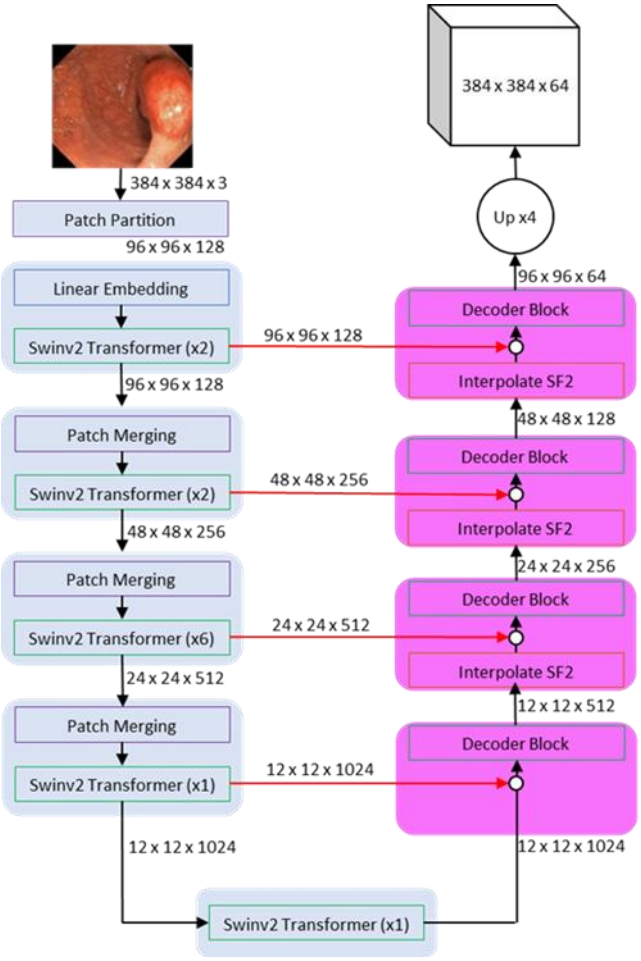


Figure 4: SwinV2-UNET [10] architecture used as the TB of the FCB-SwinV2 Transformer. The encoder stages reduce the spatial dimensions of feature maps while increasing the number of channel dimensions. Skip connections are used to pass feature maps generated by each stage of the encoder to decoder stages.

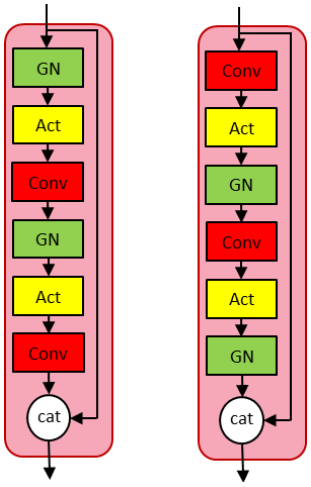


Figure 5: Changes made to the residual block (RB). Original RB used by the FCN-Transformer (left) vs the RB used by the FCB-SwinV2 Transformer (right) which features residual post normalization. The overall structure of the FCB is detailed fully in [7].

Figure 5 shows the minor changes made to the FCB when compared to the FCB of the original FCN-Transformer model. These modifications included an increased number of channel dimensions (to match the number of channel dimensions output from the TB) and a change in the order of group normalization (GN), convolution (Conv) and activation (Act) layers in the FCB residual block (RB) inspired by the residual post normalization approach of the SwinV2 Transformer.

IV. Experiments

A. Dataset Selection and Partitioning

The Kvasir-SEG [11], CVC-ClinicDB [12] and ETIS-LaribPolypDB [13] datasets have been used in this work to evaluate the performance of the FCB-SwinV2 Transformer. These datasets have been chosen because they are open access at the time of writing (which is not true for all popular datasets reported within literature) and because they are commonly used in colonoscopy segmentation literature to evaluate model performance and hence allow for comparative analysis between models. The Kvasir-SEG dataset consists of 1000 images of polyps and ground truth binary segmentation masks of varying resolutions. The CVC-ClinicDB dataset consists of 612 images of polyps and ground truth binary segmentation masks of standard resolution 384x288. The ETIS-LaribPolypDB dataset consists of 196 images of polyps and ground truth binary segmentation masks of resolution 1225x966. However, after examining recent literature on deep learning polyp segmentation models which use these datasets for evaluation, two important issues have been identified.

Firstly, many authors evaluating polyp segmentation models use popular dataset splitting functions (such as the train_test_split function from the scikit-learn python module) to create random training/validation/test data partitions (typically using an 80%/10%/10% ratio). However, for the relatively small image datasets used to evaluate colonoscopy segmentation models, minor differences in how the data is partitioned could cause noticeable performance changes. Recently this has been demonstrated for models being evaluated on the Kvasir-SEG dataset where performance changes greater than 1% were shown to occur for different data partitions [42]. This is significant because the current highest performing models are now typically separated by less than 1% performance differences. Random number seeds are often used to control the data partition, but the exact methodologies used to create data loaders and the specific random number seeds chosen are often not defined in enough detail. Differences in computer platforms and hidden random number state settings also exacerbate this problem.

Secondly, many authors using colonoscopy datasets which are composed of images from multiple video sequences (such as the CVC-ClinicDB [12] and ETIS-LaribPolypDB [13] datasets) create random training/validation/test data partitions but do not provide evidence to suggest they have taken steps to avoid data leakage and hence images of the same polyp could be present in the different data partitions. Data leakage is possible in the CVC-ClinicDB dataset as the 612 images are from video frames that have been taken from 29 video sequences. Data leakage is also possible in the ETIS-LaribPolypDB dataset as the 196 images are from video frames that have been taken from 34 video sequences. It is highly likely that frames from the same video sequences (and hence same polyps) are present across the training, validation and test data partitions that are evaluated and reported in literature when random splits have been used.

To provide comparative assessment and due to noticeable performance changes resulting from different data partitions, we evaluate our model against other state-of-the-art methods on the Kvasir-SEG, CVC-ClinicDB and ETIS-LaribPolypDB datasets using the same data partitions as those used in [8] to train and evaluate the DUCK-Net model. This is possible because the DUCK-Net authors provide information detailing exactly which images have been used for training, validation, and testing. To further demonstrate the issue of noticeable performance changes due to different data partitions we also include results obtained for random data partitions of the Kvasir-SEG, CVC-ClinicDB and ETIS-LaribPolypDB datasets. All data partitions used have training/validation/test data partitions with an 80%/10%/10% split. Files containing partition information are provided in the GitHub code repository of this work.

Due to the issue of data leakage being possible between partitions in the CVC-ClinicDB and ETIS-LaribPolypDB datasets, additional results are reported for the CVC-ClinicDB dataset for five training/validation/test data partitions with approximate 80%/10%/10% ratios where no data leakage occurs. This has been done to demonstrate the impact of data leakage on model evaluation. The data partitions with no data leakage were created based on a random selection of videos rather than images, preventing the same polyp being represented in the training/validation/test subsets. The video sequences used for validation and testing for the five data partitions with no data leakage are displayed in Table 1.

Table 1: Information on the CVC-ClinicDB dataset video sequences used to create 5 data partitions with no data leakage.

Partition Number	Validation Sequences	Validation Ratio	Testing Sequences	Testing Ratio
1	1, 2, 3	10.95%	4, 5, 6	9.64%
2	7, 8, 9	11.93%	10, 11, 12	8.66%
3	13, 14, 15	10.62%	16, 17, 18	10.78%
4	19, 20, 21	10.46%	22, 23, 24	9.15%
5	25, 26	7.03%	27, 28, 29	10.78%

Generalizability tests are also conducted in this work. The model trained on the Kvasir-SEG dataset (using the same 80%/10%/10% training/validation/test as DUCK-Net and random data partitions) is evaluated on the full CVC-ClinicDB dataset and the model trained on the CVC-ClinicDB dataset (using the same 80%/10%/10% training/validation/test as DUCK-Net and random data partitions) is evaluated on the full Kvasir-SEG dataset. It should be noted that CVC-ClinicDB and Kvasir-SEG were collected at different clinical sites and countries which further ensures dataset independence as different devices and acquisition protocols would have been used. Therefore, generalizability tests give an indication of how the models perform with respect to a somewhat different data distribution and help alleviate the identified issues as they greatly reduce the impact of data partition changes and eliminate data from the training partition leaking into the test partition. Such tests are also more representative of real-world scenarios where models are used on data sampled from distributions which may be different from the distribution of the training data.

B. Evaluation Metrics

The performance of the FCB-SwinV2 Transformer is assessed using Dice coefficient, Intersection over Union (IoU), precision, and recall metrics. These metrics are computed using the following formulas:

$$DICE = \frac{2TP}{2TP+FP+FN} \quad (1)$$

$$IoU = \frac{TP}{TP+FP+FN} \quad (2)$$

$$PRECISION = \frac{TP}{TP+FP} \quad (3)$$

$$RECALL = \frac{TP}{TP+FN} \quad (4)$$

Here, TP (True Positive) refers to correctly predicted segmentation pixels, FP (False Positive) represents incorrectly predicted pixels labeled as polyps, and FN (False Negative) denotes incorrectly predicted pixels labeled as non-polyp.

Dice and IoU metrics are positively correlated. The Dice score offers a reliable estimate of the average segmentation performance across a test partition, while the IoU score serves to penalize individual instances of poor segmentation more substantially within the test partition. Each image in a test partition is input into the FCB-SwinV2 Transformer model to generate a binary segmentation prediction, which is then compared to the corresponding ground truth binary segmentation map to compute Dice, IoU, precision, and recall scores. Upon processing all images within a test partition, the mean Dice (mDice), mean IoU (mIoU), mean precision (mPrec), and mean recall (mRec) scores for the test partition are calculated.

However, after examining recent literature on deep learning polyp segmentation models an important issue with how metrics are reported has been identified.

While many models use standard metrics such as the Dice coefficient for model evaluation, there is variance in how these metrics are reported across literature. This variance stems from the methods used to calculate average metrics for datasets. The most frequently used averaging procedures are 'image-wise averaging', 'batch-wise averaging', and 'dataset-wise averaging'. For image-wise averaging, each individual prediction map is compared with its corresponding ground truth map to produce a metric score for each image. These individual metric scores are then summed up and divided by the total number of images in the test set to yield an average score. For batch-wise averaging, prediction maps and their respective ground truth maps within a batch are merged to form larger composite prediction and ground truth maps. The metric score is then computed based on these aggregated maps. For dataset-wise averaging, all prediction maps in the test set are consolidated into a single composite prediction map, and similarly, all ground truth maps are combined. The metric score is determined by comparing this holistic prediction against the complete composite ground truth map (this is sometimes referred to as the 'global score'). For many colonoscopy datasets there is a large variation in the pixel-based size of polyps and corresponding ground truth segmentation maps which can cause discrepancies between each averaging method. A simple demonstration of this based on pixel percentage correctly classified is given in Figure 6.

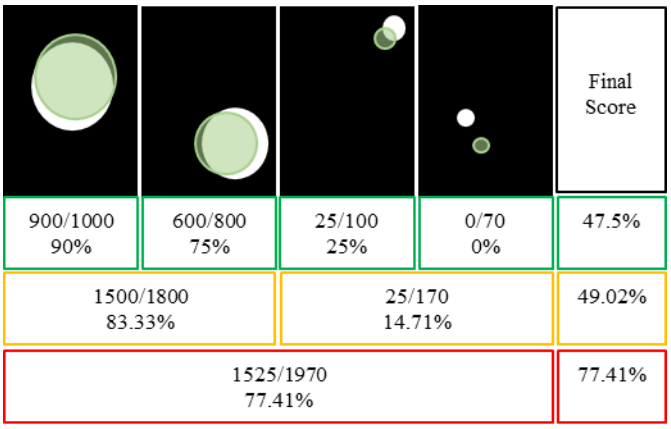


Figure 6: Demonstration of the discrepancies between image-wise averaging (top/green), batch-wise averaging using batch size of 2 (middle/orange) and dataset-wise averaging (bottom/red). The light green circles represent model predictions whilst the white circles represent ground truth maps.

Many models evaluated within literature do not explicitly state which averaging method has been used when reporting final scores which can lead to problems when trying to fairly compare model performance. A solution would be for all authors to use image-wise averaging as this is likely to produce the most conservative scores as it captures true individual segmentation performance for both large and small polyps.

In this work it has therefore been necessary to use different averaging methods to fairly compare model results reported in literature and demonstrate the impact of the different averaging techniques. To compare against results reported for the DUCK-Net model using the DUCK-Net data partitions, dataset-wise averaging was used to evaluate performance.

Efforts were also made to replicate the evaluation procedure of the DUCK-Net model in this work which also allowed image-wise averaged results to be reported for the DUCK-Net model. When using random data partitions to compare against models reported in literature (excluding the DUCK-Net model) image-wise averaging was used as this has been found to provide the most conservative estimates of model performance. For the FCB SwinV2 Transformer design ablation studies and to evaluate model performance when ensuring no-data leakage occurs batch-wise averaging was used.

An example of how mDice is calculated using dataset-wise averaging for a dataset with N images is given below:

$$TP_{total} = \sum_{i=1}^{N} TP_i \tag{5}$$

$$FP_{total} = \sum_{i=1}^{N} FP_i \tag{6}$$

$$TN_{total} = \sum_{i=1}^{N} FN_i \tag{7}$$

$$mDice_{datasetwise} = \frac{2TP_{total}}{2TP_{total}+FP_{total}+FN_{total}} \tag{8}$$

An example of how mDice is calculated using image-wise averaging for a dataset with N images is given below:

$$Dice_i = \frac{2TP_i}{2TP_i+FP_i+FN_i} \tag{9}$$

$$mDice_{imagewise} = \frac{1}{N}\sum_{1}^{N} Dice_i \tag{10}$$

Here TP_i, FP_i, FN_i and $Dice_i$ represent the number of true positive pixels, false positive pixels, false negative pixels, and dice score for the i^{th} image respectively.

C. Computational Implementation Details

The FCB SwinV2 Transformer model was implemented using PyTorch. The model was trained and evaluated on images of resolution 384x384 and predicted binary segmentation maps of resolution 384x384. This resolution was used due to the availability of ImageNet [41] pre-trained SwinV2 transformer models. For each of the datasets the model was trained for 200 epochs with the loss function consisting of the sum of the Binary Cross Entropy (BCE) loss and dice loss. All training was completed using a single NVIDIA 3090 GPU which necessitated a batch size of 2. Total training time ranged between approximately 3 hours (ETIS-LaribPolypDB) and 12 hours (Kvasir-SEG). When running in inference mode the model was capable of processing approximately 20 images per second. The AdamW [43] optimizer was used with an initial learning rate of 1e-5. The learning rate was reduced by a factor of 0.6 when the training loss did not improve over 10 epochs. Model weights were saved each time the validation dice score surpassed the previous maximum score. To generate the predicted segmentation map, pixel values were assigned a value of 1 if the model's output exceeded a threshold value of 0.5, and 0 if it fell below this threshold. Experimental results demonstrated that minor adjustments to the threshold value, either increasing or decreasing it, led to modest improvements in performance, depending on the specific dataset utilized. However, the threshold value was kept at 0.5 so comparison against other models within the literature was fair.

Table 2: Summary of training options/parameters used when training the FCB SwinV2 Transformer model.

Training Option/Parameter	Selected Option/Value
Input Resolution	384x384
Predicted Map Resolution	384x384
Optimizer	AdamW
Segmentation Threshold	0.5
Epochs	200
Initial Learning Rate	1e-5
Learning Rate Patience (epochs)	10
Learning Rate Reduction	0.6

The data augmentations used in this work closely follows those employed by the authors of the original FCN-Transformer [7]. Geometrical data augmentations applied to the training images and masks included: vertical and horizontal flips with a probability of 0.5; transposing with a probability of 0.5, scaling with a magnitude sampled uniformly from [0.5, 1.5]; shearing with an angle sampled uniformly from [-22.5º, 22.5º]; and affine transformations with rotations. Note that horizontal and vertical translations are sampled uniformly from [-48,48] with rotation angles being sampled uniformly from [-180º, 180º]. Color data augmentations were applied to the training images only and

included: color jitter with brightness factor sampled uniformly from [0.6, 1.6], contrast factor sampled uniformly from [0.8, 1.2], saturation factor sampled uniformly from [0.75, 1.25] and hue factor sampled uniformly from [0.99, 1.01]; and normalization of RGB images between the interval [-1, 1].

D. Estimation of Model Uncertainty

Uncertainty estimation in deep learning models can be broadly categorized into two types: epistemic and aleatoric. Epistemic uncertainty arises from the model's lack of knowledge due to limited training data and it can be potentially reduced with more data or changes in model architecture. In contrast, aleatoric uncertainty originates from the inherent noise in the data and remains irreducible even with additional data or model refinements. For polyp segmentation comprehensive uncertainty analysis would be ideal. This could be achieved by employing methods like varying random number seeds to induce small alterations in the model (e.g., weight initialization, dropout) and utilizing diverse data partitions (e.g., K-fold cross-validation). However, the constraints of computational resources often render comprehensive uncertainty evaluations infeasible. As such, there is a clear need for methods that efficiently approximate uncertainties.
Epistemic uncertainty is particularly relevant for polyp segmentation, especially given the likelihood of models encountering out-of-distribution data when used in real-world clinical scenarios. Monte Carlo (MC) dropout presents a viable solution for approximating epistemic uncertainty. Dropout is a widely used regularization technique in deep learning models which involves the random omission of neurons during training to prevent model over-reliance on specific neurons and reduce overfitting [44]. MC dropout [45] [46] takes advantage of the standard dropout mechanism by enabling dropout during inference. This allows slightly different versions of the model to make predictions on images within the test set. By recording performance metrics across multiple inference runs with MC dropout, we can assess the variance in model outputs hence allowing approximate estimations of epistemic uncertainty. However, it should be noted that when using MC dropout, the entire capacity of the model is not utilized and predictions can often be less accurate when compared to single deterministic run values where the whole capacity of the network is available. In this work 100 MC dropout model runs on selected test sets have been conducted to provide uncertainty estimates.

E. FCB-SwinV2 Transformer Ablation Study

In order to investigate other design architectures and help provide some insight into how modifications impact performance, additional FCB-SwinV2 Transformer model architectures were also developed and tested. The additional FCB-SwinV2 Transformer models were evaluated against the Kvasir-SEG [11] dataset using a random data partition and metrics have been calculated using batch-wise averaging.

The first additional model tested used Convolutional Block Attention Modules (CBAMs) [40] [47] within decoder blocks in the TB instead of SCSE modules. CBAMs are similar to SCSE modules as they aim to produce refined feature maps to maximize information propagation through the network. CBAM contains two sequential component modules called the Channel Attention Module (CAM) and the Spatial Attention Module (SAM). The structure of the CBAM is shown in Figure 7.

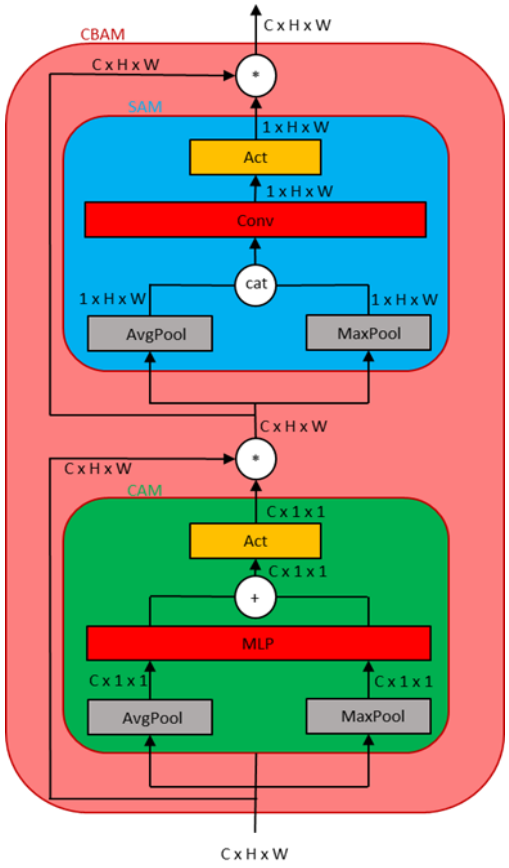


Figure 7: Structure of the CBAM which combines sequentially combines the output of the CAM and SAM. The mechanisms used are like that of the SCSE module.

The CAM first produces two tensors with reduced spatial dimensions from an input feature map by using parallel average pooling and max pooling layers. The two tensors are then passed through a shared MLP layer before being combined using element-wise summation. The resulting tensor is passed through an activation layer and it is then combined with the original feature map using element-wise multiplication. The new feature map is then passed to the SAM which produces two tensors with reduced channel dimensions using parallel average pooling and max pooling layers. The two tensors are then combined using concatenation before being passed through a convolution

layer. The resulting tensor is passed through an activation layer and it is then combined with the feature map produced by the CAM using element-wise multiplication.

The second additional model tested had a TB that was based very closely on the original, fully-attention based Swin U-Net [34]. The only modifications were related to the replacement of the original Swin Transformer blocks with SwinV2 Transformer blocks.

The final additional model tested used the original RB (pre-normalization approach) of the FCN-Transformer in the FCB (see Figure 5).

V. Results and Evaluation

A. Results of FCB-SwinV2 Transformer Ablation Study

Results for the additional investigated FCB-SwinV2 Transformer model architectures evaluated against the Kvasir-SEG [11] dataset using a random data partition are displayed in Table 3.

Table 3: FCB-SwinV2 Transformer model variation tests on the Kvasir-SEG dataset using the same random data partition. Scores reported here were calculated using batch-wise averaging.

	mDice	mIoU	mPrec	mRec
FCB-SwinV2 Transformer (Selected architecture)	**94.30**	**89.82**	93.77	**95.73**
CBAM Decoder in TB	94.10	89.44	93.73	95.26
Fully Attention based SwinV2 U-Net TB	93.25	88.87	93.56	94.90
Original RB of FCN-Transformer	93.84	89.08	**95.51**	93.15

The Fully Attention based SwinV2 U-Net TB model was found to perform worse than the selected FCB-SwinV2 Transformer design (using SCSE modules in the TB decoder) and the FCB-SwinV2 Transformer using CBAMs in the TB decoder. This provides evidence that convolutional based decoders making use of squeeze-excite style mechanisms perform better than when using attention only mechanisms for the small data set sizes used in this study. The drop in performance when using the original RB of FCN-Transformer provides evidence that the residual post normalization approach within the modified RB of the FCB helps to improve segmentation performance.

Since the additional models did not perform as well as the FCB-SwinV2 Transformer architecture design (described in Section III), they are not investigated further in this study.

B. DUCK-Net Data Partitioning Results

The performance of the FCB-SwinV2 Transformer using both image-wise averaging and dataset-wise averaging is assessed using DUCK-Net data partitions across the Kvasir-SEG, CVC-ClinicDB, and ETIS-LaribPolypDB datasets, with results presented in Tables 3-8. Comparative analysis was possible owing to the DUCK-Net authors providing folders containing the distinct training, validation, and test partitions employed for each respective dataset. Where applicable, comparisons are made against the DUCK-Net-34 and FCN-Transformer model results, both from prior works **[8]** and those derived from executing the readily available DUCK-Net-34 code base and pre-trained models, with no re-training or fine-tuning, in the present study (this work). This was necessary to generate results using image-wise averaging during evaluation not included in the original paper [11].

Table 4 and Table 5 demonstrate the model's performance on the Kvasir-SEG DUCK-Net data partition, employing dataset-wise and image-wise averaging, respectively. Both tables illustrate comparisons against DUCK-Net-34 results reported in [8] and those obtained in the current study.

Table 4: Comparison of model performance using dataset-wise averaging on the Kvasir-SEG dataset against the 34 filter DUCK-Net model and FCN-Transformer model using the DUCK-Net data partition.

	mDice	mIoU	mPrec	mRec
FCN-Transformer	92.20	85.54	92.38	92.03
DUCK-Net-34	95.02	90.51	96.28	93.79
DUCK-Net-34 (This work)	94.71	89.93	95.54	93.87
FCB-SwinV2 Transformer	**95.77**	**91.88**	**96.78**	**94.78**

Table 5: Comparison of model performance using image-wise averaging on the Kvasir-SEG dataset against the 34 filter DUCK-Net model using the DUCK-Net data partition.

	mDice	mIoU	mPrec	mRec
DUCK-Net-34 (This work)	93.91	89.38	94.81	94.22
FCB-SwinV2 Transformer	**94.88**	**90.82**	**95.61**	**94.90**

Tables 6 and 7 extend this analysis to the CVC-ClinicDB DUCK-Net data partition, adhering to the same averaging methods and comparative benchmarks.

Table 6: Comparison of model performance using dataset-wise averaging on the CVC-ClinicDB dataset against the 34 filter DUCK-Net model and FCN-Transformer model using the DUCK-Net data partition.

	mDice	mIoU	mPrec	mRec
FCN-Transformer	93.27	87.40	**97.28**	89.58
DUCK-Net-34	94.78	90.09	94.68	**94.89**
DUCK-Net-34 (This work)	94.64	89.82	94.42	94.86
FCB-SwinV2 Transformer	**94.89**	**90.28**	95.43	94.36

Table 7: Comparison of model performance using image-wise averaging on the CVC-ClinicDB dataset against the 34 filter DUCK-Net model using the DUCK-Net data partition.

	mDice	mIoU	mPrec	mRec
DUCK-Net-34 (This work)	92.65	87.44	**93.75**	92.64
FCB-SwinV2 Transformer	**93.32**	**88.09**	93.71	**93.38**

The model's performance is also examined using the ETIS-LaribPolypDB DUCK-Net data partition with results detailed in Tables 8 and 9.

Table 8: Comparison of model performance using dataset-wise averaging on the ETIS-LaribPolypDB dataset against the 34 filter DUCK-Net model and FCN-Transformer model using the DUCK-Net data partition.

	mDice	mIoU	mPrec	mRec
FCN-Transformer	91.63	84.55	96.33	87.36
DUCK-Net-34	93.54	87.88	93.09	94.00
DUCK-Net-34 (This work)	92.77	86.52	91.76	93.81
FCB-SwinV2 Transformer	**95.03**	**90.54**	**94.73**	**95.34**

Table 9: Comparison of model performance using image-wise averaging on the ETIS-LaribPolypDB against the 34 filter DUCK-Net model using the DUCK-Net data partition.

	mDice	mIoU	mPrec	mRec
DUCK-Net-34 (This work)	81.76	75.98	82.10	82.48
FCB-SwinV2 Transformer	**91.70**	**85.40**	**91.93**	**92.10**

The generalizability performance of the FCB-SwinV2 Transformer (using dataset-wise averaging to allow for comparisons against [8]) when trained on the Kvasir-SEG dataset using the DUCK-Net data partitions and evaluated on the CVC-ClinicDB dataset and when trained on the CVC-ClinicDB dataset using the DUCK-Net data partitions and evaluated on the CVC-ClinicDB dataset are displayed in Table 10 and Table 11 respectively. Note that the same data partitions are used (rather than re-training using all available data within a dataset) as this approach is used by other models within literature and hence allows fair comparison between models.

Table 10: Comparison of model generalizability performance using dataset-wise averaging when trained on the Kvasir-SEG dataset using the DUCK-Net data partitions and evaluated on the CVC-ClinicDB dataset against the 34 filter DUCK-Net model and FCN-Transformer model.

	mDice	mIoU	mPrec	mRec
FCN-Transformer	83.14	71.14	88.39	78.48
DUCK-Net-34	82.11	69.65	88.60	76.50
FCB-SwinV2 Transformer	**88.11**	**78.74**	**91.19**	**85.22**

Table 11: Comparison of model generalizability performance using dataset-wise averaging when trained on the CVC-ClinicDB dataset using the DUCK-Net data partitions and evaluated on the Kvasir-SEG dataset against the 34 filter DUCK-Net model and FCN-Transformer model.

	mDice	mIoU	mPrec	mRec
FCN-Transformer	88.00	78.58	**96.59**	80.82
DUCK-Net-34	82.51	70.23	77.40	**88.34**
FCB-SwinV2 Transformer	**88.64**	**79.59**	94.23	83.67

A visual comparison of mDice scores across all datasets and generalizability tests using dataset-wise averaging for the DUCK-Net data partitions used is presented in Figure 8. This visualization highlights that the FCB-SwinV2 Transformer consistently achieves the highest mDice scores.

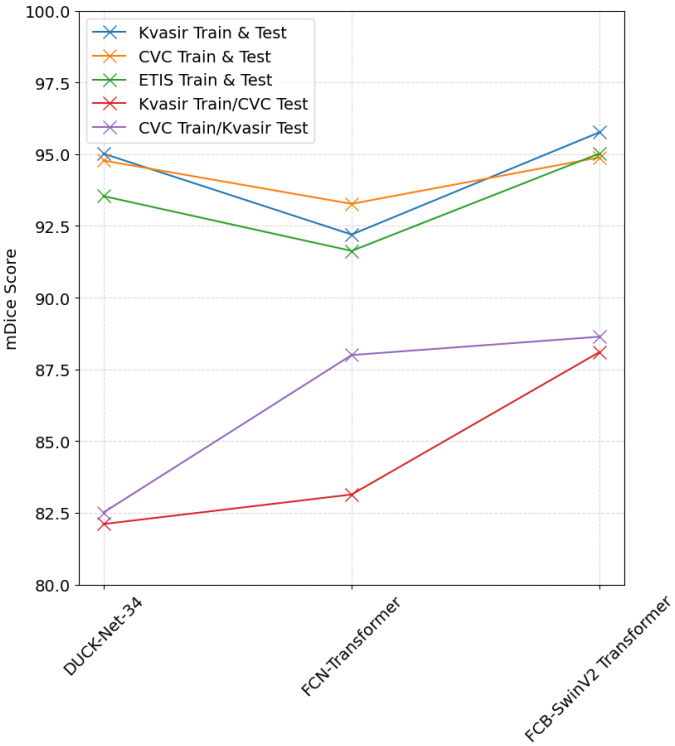


Figure 8: Visual comparison of mDice scores across all datasets and generalizability tests using dataset-wise averaging for the DUCK-Net data partitions.

As demonstrated through Tables 4-9, the FCB-SwinV2 Transformer consistently surpasses the DUCK-Net-34 and FCN-Transformer models with respect to mDice and mIoU scores across all datasets when using the DUCK-Net data partitions for both types of averaging techniques used.

As evidenced in Tables 10-11, The FCB-SwinV2 Transformer also outperforms the previous state-of-the-art models FCN-Transformer and DUCK-Net on mDice and mIoU scores for both generalizability tests. This is an important result as the impact of random seed variations on data partitioning are minimized and potential data leakage is

eliminated (due to the training and test data being from two separate datasets/distributions). This means that these generalizability results provide a reliable evaluation of true relative performance between models.

Whilst maintaining high mDice and mIoU scores, the FCB-SwinV2 Transformer also typically achieves the highest mRec scores. In a clinical setting where the primary goal would be to identify and map every potential polyp and ensure nothing is missed, having high mRec scores (whilst maintaining high mDice and mIoU scores) would be desirable.

As expected, the comparison between dataset-wise averaging and image-wise averaging demonstrates that image-wise averaging consistently produces more conservative results for both the 34 filter DUCK-Net and FCB-SwinV2 Transformer models. For the Kvasir-SEG dataset, the DUCK-Net model witnesses a marginal decrease in mDice and mIoU scores, namely 0.80% and 0.55% respectively, compared to a 0.89% and 1.06% reduction for the FCB-SwinV2 Transformer. A slightly larger impact is seen for the CVC-ClinicDB dataset with the DUCK-Net model witnessing a 1.99% decrease for the mDice score and 2.38% decrease for the mIoU score whilst for the FCB-SwinV2 Transformer there is a 1.57% decrease for the mDice score and 2.19% decrease for the mIoU score. For the ETIS-LaribPolypDB dataset, the impact of the different averaging techniques is much more pronounced with the DUCK-Net model experiencing a sharp 11.01% and 10.54% fall in mDice and mIoU scores respectively whilst there is a 3.33% and 5.14% downturn for the FCB-SwinV2 Transformer. It's postulated that this divergence could be caused by considerable variations in relative polyp sizes within the ETIS-LaribPolypDB, as visualized in Figure 9. Figure 6 also serves as a reminder of how variations in relative polyp sizes impact different averaging techniques.

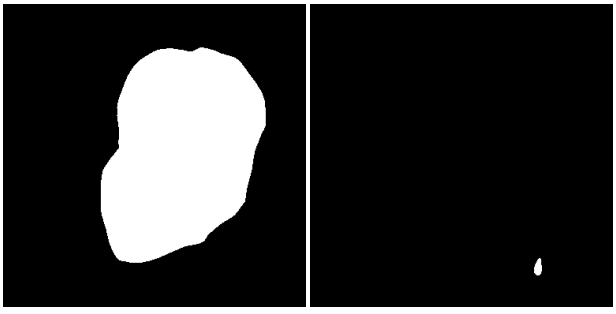


Figure 9: Demonstration of the large differences in relative polyp sizes contained within the ETIS-LaribPolypDB dataset [14].

C. Random Data Partitioning Results

To generate further comparisons of relative model performance, the FCB-SwinV2 Transformer is assessed using random data partitions across the Kvasir-SEG, CVC-ClinicDB, and ETIS-LaribPolypDB datasets with results presented in Tables 12-14. Results for the FCB-SwinV2 Transformer have been calculated using image-wise averaging to provide a more conservative estimate of performance. Where possible, comparisons are made against results reported in literature for recent high performing models. However, it is unknown which averaging techniques have been used when reporting mean scores for these other models.

The performance of the FCB-SwinV2 Transformer for the Kvasir-SEG [11] dataset is reported in Table 12 for a random data partition. Comparisons to FCN-Transformer model performance from the original paper [7] and from [42] with the advanced data augmentation technique named 'Spatially Exclusive Pasting' (SEP) are provided. Note that Results for the Meta-Polyp model reported in [48] (95.90mDice and 92.10mIoU) are not included in Table 12. This is because the Meta-Polyp authors report results for the Kvasir-SEG dataset after using a training dataset which merged 900 images from Kvasir-SEG with 550 images from CVC-ClinicDB.

Table 12: Comparison of model performance using image-wise averaging on the Kvasir-SEG dataset against using random data partitioning.

	mDice	mIoU	mPrec	mRec
PraNet [22]	90.11	84.03	90.34	92.72
MSRF-Net [23]	92.17	89.14	**96.66**	91.98
Polyp2Seg [33]	92.90	88.20	-	-
ESFPNet-L [30]	93.10	88.70	-	-
FCN-Transformer [7]	93.85	89.03	94.59	94.01
FCB-SwinV2 Transformer	94.04	89.49	93.95	**95.16**
FCN-Transformer + SEP [42]	**94.11**	**90.02**	-	-

The performance of the FCB-SwinV2 Transformer for the CVC-ClinicDB dataset is reported in Table 13 for a random data partition. Comparisons to FCN-Transformer model performance and other high performing models are included for comparison.

Table 13: Comparison of model performance using image-wise averaging on the CVC-ClinicDB dataset against using random data partitioning.

	mDice	mIoU	mPrec	mRec
Polyp2Seg [33]	92.90	88.10	-	-
PraNet [22]	93.58	88.67	93.70	93.88
MSRF-Net [23]	94.20	90.43	94.27	95.67
FCN-Transformer [7]	94.69	90.20	**95.25**	94.41
ESFPNet-L [30]	94.90	90.70	-	-
FCB-SwinV2 Transformer	**95.19**	**91.00**	94.79	**95.82**

The performance of the FCB-SwinV2 Transformer when trained and evaluated using the ETIS-LaribPolypDB. The results are displayed in Table 14.

Table 14: Comparison of model performance using image-wise averaging on the ETIS-LaribPolypDB using a random data partition.

	mDice	mIoU	mPrec	mRec
PraNet [22]	62.80	56.70	-	-
Polyp2Seg [33]	82.00	73.80	-	-
ESFPNet-L [30]	82.30	74.80	-	-
FCB-SwinV2 Transformer	**91.88**	**85.63**	92.97	91.95

The generalizability performance of the FCB-SwinV2 Transformer (using image-wise averaging to ensure conservative values) when trained on the Kvasir-SEG dataset using a random data partition and evaluated on the CVC-ClinicDB dataset and when trained on the CVC-ClinicDB dataset using a random data partition and evaluated on the Kvasir-SEG dataset are displayed in Table 15 and Table 16 respectively. Note again that the same data partitions are used (rather than re-training using all available data within a dataset) as this approach is used by other models within literature and hence allows fair comparison between models.

Table 15: Comparison of the model generalizability performance using image-wise averaging when trained on the Kvasir-SEG dataset using a random data partition and evaluated on the CVC-ClinicDB dataset.

	mDice	mIoU	mPrec	mRec
PraNet [22]	79.12	71.19	81.52	83.16
MSRF-Net [23]	79.21	64.98	70.00	**90.01**
FCN-Transformer [7]	87.35	80.38	**89.95**	88.76
FCB-SwinV2 Transformer	**87.77**	**80.78**	89.83	89.29

Table 16: Comparison of the model generalizability performance using image-wise averaging when trained on the CVC-ClinicDB dataset using a random data partition and evaluated on the Kvasir-SEG dataset against other high performing models.

	mDice	mIoU	mPrec	mRec
PraNet [22]	79.50	70.73	76.87	90.50
MSRF-Net [23]	75.75	63.37	83.14	71.97
FCN-Transformer [7]	88.48	82.14	93.54	87.54
FCB-SwinV2 Transformer	**89.35**	**83.34**	**94.26**	**88.15**

A visual comparison of mDice scores across all datasets and generalizability tests using image-wise averaging for the random data partitions is presented in Figure 10. This visualization highlights that the FCB-SwinV2 Transformer once again consistently achieves the highest mDice scores.

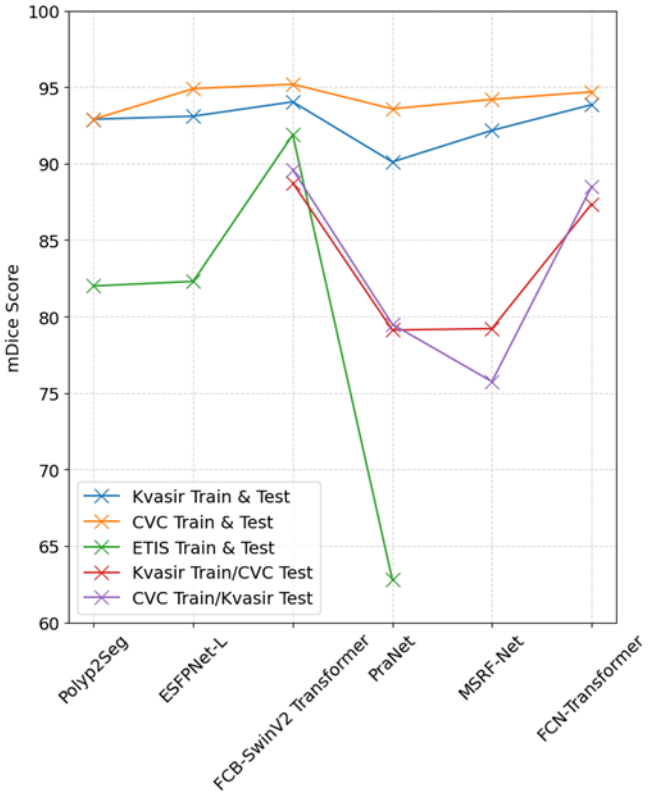


Figure 10: Visual comparison of mDice scores across all datasets and generalizability tests using image-wise averaging for random data partitioning.

The FCB-SwinV2 Transformer achieves the highest mDice and mIoU scores for both the CVC-ClinicDB and ETIS-LaribPolypDB datasets. For the Kvasir-SEG dataset, the FCB-SwinV2 Transformer achieves the second highest scores for the mDice and mIoU metrics behind only the FCN-Transformer trained using the advanced SEP data augmentation technique. When compared to baseline models (i.e. those only using standard data augmentation techniques) the FCB-SwinV2 Transformer achieves the highest scores for the mDice and mIoU metrics.

The FCB-SwinV2 Transformer also outperforms all previous high performing models on mDice and mIoU scores for both generalizability tests conducted using image-wise averaging. Once again, the importance of this result should be stressed as the effects of potential data leakage are eliminated. For future generalizability tests the impact of random seed variations could be reduced further by training using 90% or 95% of images within a dataset. Once again, the FCB-SwinV2 Transformer typically achieves the highest mRec scores (when still maintaining high mDice and mIoU scores) which could be desirable in clinical settings where the main aim is to identify and map every polyp present.

When comparing results for image-wise averaging for each dataset, model performance varies noticeably between the

DUCK-Net and random data partitions. This demonstrates the issue of noticeable performance changes due to differences in data partitioning. Since the FCB-SwinV2 Transformer outperforms other models across each dataset, using both the DUCK-Net and random data partitions, this provides further confidence that the model achieves state-of-the-art results.

D. CVC-ClinicDB No Data Leakage Partitioning Results

The results for the FCB-SwinV2 Transformer for the 5 training/validation/test data partitions with no data leakage are reported in Table 17 and displayed in Figure 11. The mean scores for the partitions with no data leakage were calculated using batch-wise averaging.

Table 17: FCB-SwinV2 Transformer model performance on CVC-ClinicDB dataset using training/validation/test data partitions with no data leakage.

Partition No.	mDice	mIoU	mPrec	mRec
1	83.16	73.26	87.81	81.95
2	85.61	78.03	91.23	81.74
3	92.50	86.35	90.91	94.96
4	93.65	88.26	95.10	92.63
5	81.86	71.17	91.04	78.23

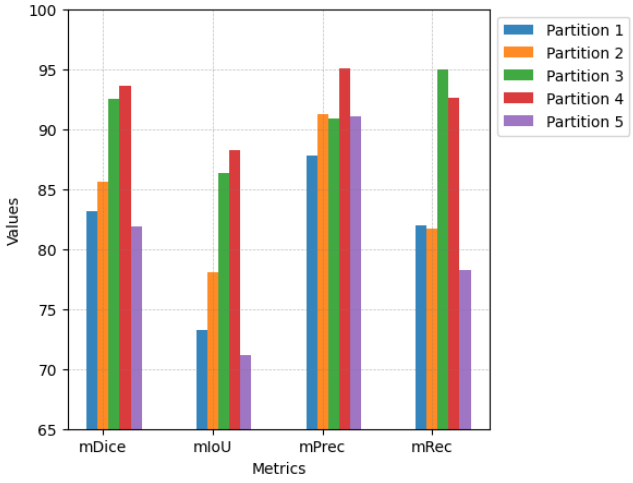


Figure 11: FCB-SwinV2 Transformer model performance on the CVC-ClinicDB dataset using training/validation/test data partitions with no data leakage.

When no data leakage occurs the performance of the FCB-SwinV2 Transformer model, whilst still strong, drops across all metrics for each of the five partitions. It is also expected that there would be a slight reduction in scores if image-wise averaging was used. For partitions 1 and 5 the performance drop is larger. This is likely due to the test images within partitions 1 and 5 being unusual when compared to the whole dataset. For example, partition 1 contains a video sequence where multiple small polyps are present which is unique when compared to other video sequences within the CVC-ClinicDB dataset. The significant reduction in scores demonstrates the impact of data leakage artificially increasing model performance for datasets like CVC-ClinicDB and ETIS-LaribPolypDB. It is extremely likely that this behavior would be replicated by any other deep learning model given that the training and testing partitions no longer contain images of the same polyps.

E. Qualitative Mask Comparisons

A visual comparison of predictions made by the FCB-SwinV2 Transformer for an image from the Kvasir-SEG dataset when trained on Kvasir-SEG and when trained using the CVC-ClinicDB dataset (i.e. generalizability test) are provided in Figure 12.

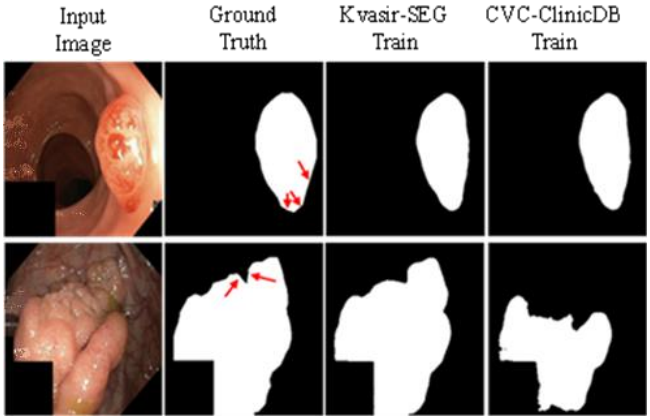


Figure 12: Comparisons of predictions made by the model for the Kvasir-SEG dataset when trained using the Kvasir-SEG and when trained using the CVC-ClinicDB dataset. Red arrows highlight sharp edges found within ground truth segmentation maps which do not appear to match polyp edges within the image.

Qualitative inspection of the predicted binary segmentation maps generated when trained using the CVC-ClinicDB dataset show that the model generalizes well for regular polyps but suffers a performance drop for large, irregular polyps within the Kvasir-SEG dataset (which are considerably different to any of the polyps within the CVC-ClinicDB dataset). Another interesting finding of the visual inspection is that some ground truth maps of the Kvasir-SEG dataset contain sharp edges (highlighted using red arrows within Figure 12) when the polyp contained within the input image appears to have smooth edges. The segmentation predictions made by the FCB-SwinV2 Transformer typically contain smoother edges. This may suggest that predictions made by the FCB-SwinV2 Transformer (and other recent state-of-the-art models) may be approaching the maximum achievable performance when trained and evaluated on available polyp segmentation datasets particularly when considering intra- and inter-observer variabilities in generating ground truth segmentation masks. This further highlights the importance of generalizability tests.

A comparison of predictions made by the FCB-SwinV2 Transformer for images from the test set of the Kvasir-SEG dataset when trained and evaluated using the DUCK-Net

partition are displayed alongside predictions made by the DUCK-Net [8] and FCN-Transformer [7] in Figure 13.

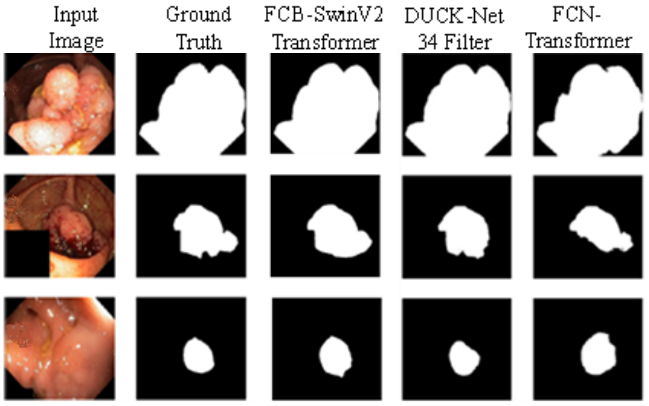


Figure 13: Predictions made by the FCB-SwinV2 Transformer for images from the test set of the Kvasir-SEG dataset when trained and evaluated using the DUCK-Net data partitions.

Qualitative inspection of the predicted binary segmentation maps generated shows that each model performs strongly. Across the three images presented the FCB-SwinV2 Transformer and DUCK-Net model produce marginally more accurate segmentation maps than the FCN-Transformer.

Visual comparisons of predictions made by the FCB-SwinV2 Transformer for a uniquely shaped polyp from the CVC-ClinicDB dataset are provided in Figure 14.

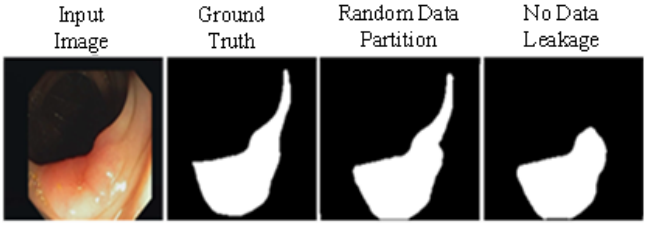


Figure 14: Comparisons of predictions made by the model for the CVC-ClinicDB dataset when trained using the random data partition and when trained using a data partition which ensured no data leakage. When using a random data partition, the model has been trained and evaluated on images of the same polyp from a video sequence resulting in artificially high performance.

These predictions are from when the model had been trained using the CVC-ClinicDB random data partition and the CVC-ClinicDB data partition which ensured no data leakage. Qualitative inspection of the predicted binary segmentation maps generated when using the random data partition shows that it matches the ground truth extremely well. When compared to the segmentation map generated ensuring no data leakage occurred (partition 4) we see a large drop in performance. This strongly demonstrates the artificially high performance when using a random data partition as, although the model has been trained and evaluated on different images, both training and test datasets contain images of the same polyp.

F. Evaluation of Model Uncertainty

The average results of the 100 MC dropout runs conducted for each dataset using the DUCK-Net data partitions and random data partitions are presented in Table 18 and Table 19 respectively.

Table 18: Mean and standard deviation (displayed using ± notation) of the 100 MC dropout conducted for each dataset using the DUCK-Net data partitions. Image-wise averaging has been used during evaluation.

Metric	Kvasir-SEG	CVC-ClinicDB	ETIS-LaribDB
mDice	93.46 ± 0.32	91.72 ± 0.45	84.43 ± 1.61
mIoU	89.00 ± 0.44	86.08 ± 0.54	76.18 ± 1.47
mPrec	94.40 ± 0.42	92.21 ± 0.56	81.83 ± 1.51
mRec	94.32 ± 0.23	92.62 ± 0.43	90.20 ± 2.07

Table 19: Mean and standard deviation (displayed using ± notation) of the 100 MC dropout runs conducted for each dataset using random data partitioning. Image-wise averaging has been used during evaluation.

Metric	Kvasir-SEG	CVC-ClinicDB	ETIS-LaribDB
mDice	92.73 ± 0.39	94.29 ± 0.35	80.82 ± 1.62
mIoU	87.59 ± 0.45	89.62 ± 0.48	71.53 ± 1.88
mPrec	92.92 ± 0.42	93.70 ± 0.46	76.19 ± 1.91
mRec	94.07 ± 0.32	95.43 ± 0.24	91.76 ± 1.48

The average mDice and average mIoU scores of the 100 MC Dropout runs conducted for each dataset using the DUCK-Net data partitions and random data partitions are visualized as boxplots in Figure 15 and Figure 17 respectively. The single deterministic run values for each dataset partition (see Tables 5, 12, 7, 13, 9 and 14) are also displayed using '+' markers. For both the DUCK-Net and random data partitions the Kvasir-SEG dataset MC Dropout results exhibited high performance. The average mDice and average mIoU scores for the DUCK-Net and random partitions represented a less than 2% drop in performance when compared to the single deterministic run values (dropout deactivated during evaluation) reported in Table 5 and Table 12 respectively. The standard deviations across all metrics are small (< 0.5%), suggesting that the model's performance is stable across different Kvasir-SEG runs.

For both the DUCK-Net and random data partitions, the CVC-ClinicDB dataset MC Dropout results again exhibited high performance. The average mDice and average mIoU scores for the DUCK-Net and random partitions represented a less than 2.5% drop in performance when compared to the single deterministic values reported in Table 7 and Table 13 respectively. With standard deviations still less than 0.6%, the model demonstrates stable performance across different CVC-ClinicDB runs. However, potential data leakage (due to

randomness in the data partitioning process) in the CVC-ClinicDB dataset warrants caution in interpreting these results.

The ETIS-LaribDB dataset shows a noticeable drop in performance for both the DUCK-Net and random data partitions compared to the other two datasets. Average metric scores fall by as much as 16% when compared to the single deterministic values reported in Table 9 and Table 14 respectively. The larger standard deviations (>1.5%) suggest a more variable performance across different runs. The reduced dataset size (196 images) likely contributes to this increased variability and reduced performance. It is possible that the smaller dataset size leads to the model overfitting and when dropout is applied the removal of over-specialized neurons significantly impacts performance.

The boxplots presented in Figure 15 and Figure 16 also provide a useful visualization of the impact of dataset partitioning on both the deterministic and MC Dropout results as for each dataset model performance varies noticeably between the DUCK-Net and random data partitions.

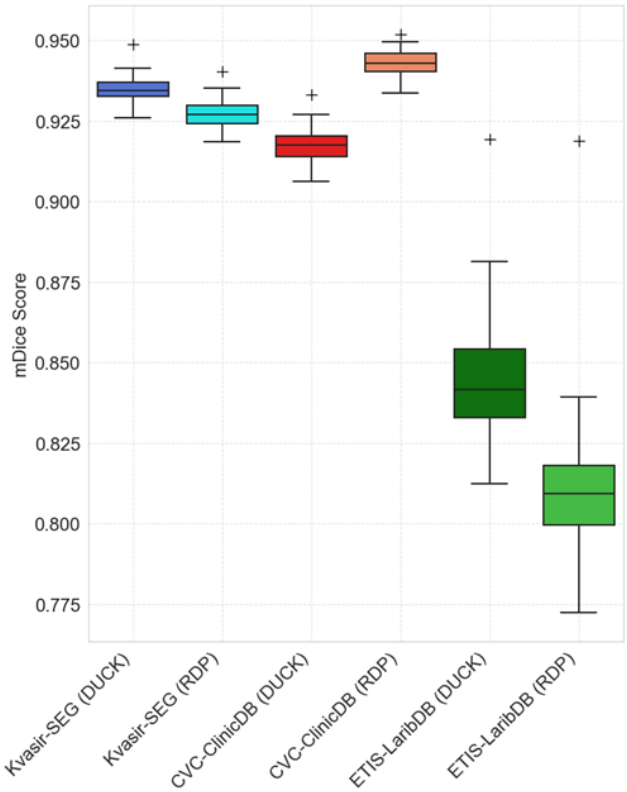


Figure 15: Boxplot visualization of mDice score statistics from the 100 MC dropout runs conducted for each dataset using the DUCK-Net and random data partitions.

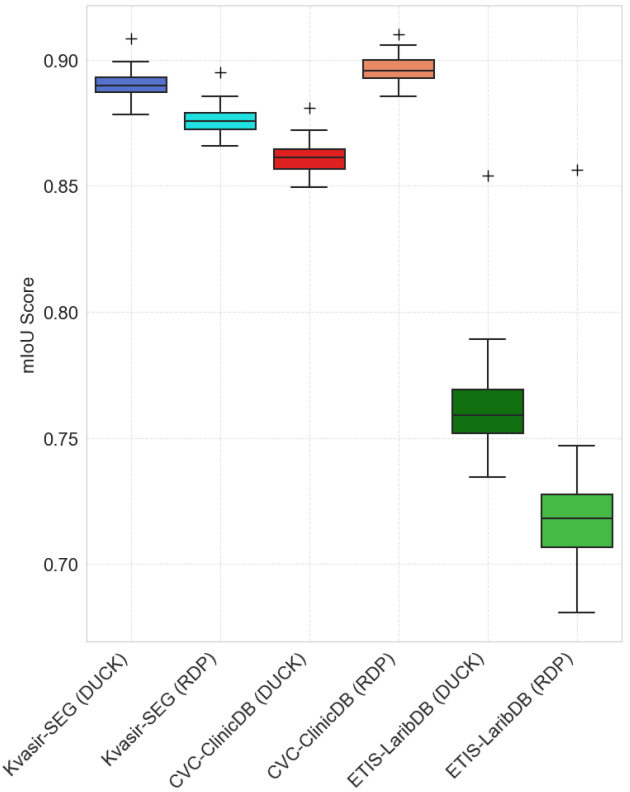


Figure 16: Boxplot visualization of mIoU score statistics from the 100 MC dropout runs conducted for each dataset using the DUCK-Net and random data partitions.

VI. Conclusion

This paper proposes a novel deep learning model for colon polyp segmentation called the FCB-SwinV2 Transformer. The performance of this model has been extensively investigated through rigorous quantitative and qualitative comparison against previous state-of-the-art models reported in literature. In addition, ablation studies and epistemic uncertainty analysis (estimated by MC dropout) were conducted to provide further insights into the performance of the FCB-SwinV2 Transformer. The FCB-SwinV2 Transformer achieved the highest mDice and mIoU scores in each of the respective test sets used when compared to baseline models reported in literature. In addition, generalizability tests which followed the same methodology reported in literature [7] [8] [22] [23], were conducted with results being compared against previous state-of-the-art models. These generalizability tests showed that the FCB-SwinV2 Transformer outperformed previous models on mDice and mIoU scores. These results demonstrate the state-of-the-art performance of the FCB-SwinV2 Transformer and its improved adaptability and applicability to data outside of the training distribution.

The critical importance of dataset partitioning and averaging methodologies used to calculate performance metrics have also been demonstrated experimentally. It has been observed

that data leakage can artificially increase model performance and that general data partitioning differences (even when there is no data leakage) can cause fluctuations in model performance. It has also been shown that when calculating performance metrics, image-wise averaging consistently yielded lower mean scores than dataset-wise averaging and is therefore believed to capture true individual polyp segmentation performance. The colonoscopy research community could therefore benefit from expert clinicians defining standardized reporting metrics and creating standardized K-Fold cross validation data splits on popular colonoscopy benchmarking datasets.

Some further incremental improvements to the FCB-SwinV2 Transformer and training process could also be made which may enhance model performance. Examples include: using more advanced data augmentation techniques; replacing the FCB branch with an ImageNet pre-trained FCN model or DUCK-Net based architecture [8], and larger ImageNet pre-trained SwinV2 encoders.

Acknowledgements

Data access statement: The study reported in this article has been supported by existing openly available datasets, namely Kvasir-SEG, CVC-ClinicDB, and ETIS-LaribPolypDB.

REFERENCES

[1] World Health Organization, "Colorectal cancer," July 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer#:~:text=Colon%20cancer%20is%20the%20second,and%20mortality%20rates%20were%20observed. [Accessed December 2023].

[2] A.M. Leufkens, M.G.H. van Oijen, F.P. Vleggaar & P.D. Siersema, "Factors influencing the miss rate of polyps in a back-to-back colonoscopy study," *Endoscopy*, vol. 44, no. 5, pp. 470-475, 2012.

[3] N.H. Kim, Y.S. Jung, W.S. Jeong, H.J. Yang, S.K. Park. K. Choi & D.I. Park, "Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies," *Intestinal Research*, vol. 15, no. 3, pp. 411-418, 2017.

[4] J. Lee, S.W. Park, Y.S. Kim, K.J. Lee, H.S, P.H. Song, W.J. Yoon & J.S. Moon, "Risk factors of missed colorectal lesions after colonoscopy," *Medicine,* vol. 96, no. 27, 2017.

[5] D.A. Corley, "Adenoma Detection Rate and Risk of Colorectal Cancer and Death," *New England Journal of Medicine,* vol. 370, pp. 1298-1306, 2014.

[6] M.J. Whitson, C.A. Bodian, J. Aisenberg & L.B. Cohen, "Is production pressure jeopardizing the quality of colonoscopy? A survey of U.S. endoscopists' practices and perceptions," *Gastrointestinal Endoscopy,* vol. 75, no. 3, pp. 641-648, 2012.

[7] E. Sanderson & B.J. Matuszewski, "FCN-Transformer Feature Fusion for Polyp Segmentation," in *Medical Image Understanding and Analysis (MIUA)*, 2022.

[8] RG. Dumitru, D. Peteleaza & C. Craciun, "Using DUCK-Net for polyp image segmentation," *Nature Scientific Reports,* vol. 13, 2023.

[9] W. Wang, E. Xie, X. Li, D-P. Fan, K. Song, D. Liang, T. Lu, P. Luo & LShao, "PVT v2: Improved baselines with Pyramid Vision Transformer," *Computational Visual Media,* vol. 8, no. 3, pp. 415-424, 2022.

[10] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, Li Dong, F. Wei & B. Guo, "Swin Transformer V2: Scaling Up Capacity and Resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[11] D. Jha, P.H. Smedsrud, M.A. Riegler, P. Halvorsen, T. de Lange, D. Johansen & H.D. Johansen, "Kvasir-SEG: A Segmented Polyp Dataset," in *International Conference on Multimedia Modeling*, 2020.

[12] J. Bernal, F.J. Sanchez, G. Fernandez-Esparrach, D. Gil, C. Rodríguez & F. Vilariño,, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics,* vol. 43, pp. 99-111, 2015.

[13] J. Bernal et al, "Comparative validation of polyp detection methods in video colonoscopy: Results from the MICCAI 2015 endoscopic vision challenge," *IEEE Trans. Med. Imaging*, vol. 36, pp. 1231-1249, 2017.

[14] O. Ronneberger, P. Fischer, T. Brox, N. Navab, J. Hornegger, W.M. Wells & A.J. Frangi, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI*, 2015.

[15] H.A. Qadir, Y. Shin, J. Solhusvik, J. Bergsland, L. Aabakken & I. Balasingham , "Polyp Detection and Segmentation using Mask R-CNN: Does a Deeper Feature Extractor CNN Always Perform Better?," in *13th International Symposium on Medical Information and Communication Technology (ISMICT)*, 2019.

[16] J. Debesh, M.A. Riegler, D. Johansen, P. Halvorsen & H.D. Johansen, "Double U-Net: A deep convolutional neural network for medical image segmentation," in *IEEE 33rd International Symposium on Computer-Based Medical Systems*, 2020.

[17] M. Hwang, D. Wang, X-X. Kong, Z. Wang, J. Li, W-C. Jiang, K-S. Hwang, K. Ding, "An automated detection system for colonoscopy images using a dual encoder-decoder model," *Computerized Medical Imagaing and Graphics*, vol. 84, 202.

[18] A.O. Ige, N.K. Tomar, F.O. Aranuwa, O. Oriola, A.O. Akingbesote, M.H. Noor, M. Mazzara & B.S. Aribisala, "ConvSegNet: Automated Polyp Segmentation From Colonoscopy Using Context Feature Refinement With Multiple Convolutional Kernel Sizes," *IEEE Access*, vol. 11, pp. 16142-16155, 2023.

[19] N.S. An, P.N. Lan, D.V. Hang, T.Q. Trung, N.T. Thuy & D.V. Sang, "BlazeNeo: Blazing Fast Polyp Segmentation and Neoplasm Detection," *IEEE Access*, vol. 10, pp. 43669-43684, 2022.

[20] D. Jha, P.H. Smedsrud, M.A. Riegler, D. Johansen, T. de Lange, P. Halvorsen & H.J. Dagenborg, "ResUNet++: An Advanced Architecture for Medical Image Segmentation," in *21st IEEE International Symposium on Multimedia*, 2019.

[21] Y. Guo, J. Bernal, & B. Matuszewski, "Polyp Segmentation with Fully Convolutional Deep Neural Networks - Extended Evaluation Study," *Journal of Imaging*, vol. 6, no. 7, 13 July 2020.

[22] D-P Fan, G-P Ji, T. Zhou, G. Chen, H. Fu, J. Shen & L. Shao , "PraNet: Parallel Reverse Attention Network for Polyp Segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020.

[23] A. Srivastava, D. Jha, S, Chanda, U. Pal, H.D. Johansen, D. Johansen, M.A. Riegler & S. Ali, "MSRF-Net: A Multi-Scale Residual Fusion Network for Biomedical Image Segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 5, pp. 2252-2263, 2022.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit & N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations*, Virtual, 2021.

[25] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez & P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," in *Proceedings of the Advances in Neural Information Processing Systems 34* , 2021.

[26] B. Dong, W. Wang, D-P. Fan, J. Li, H. Fu & L. Shao, "Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers," *CAAI Artificial Intelligence Research*, vol. 2, 2023.

[27] Q. Guo, X. Fang, L. Wang & E. Zhang, "Polyp Segmentation of Colonoscopy Images by Exploring the Uncertain Areas," *IEEE Access*, vol. 10, pp. 52971-52981, 2022.

[28] N.K. Tomar, D. Jha, S. Ali, H.D. Johansen, D. Johansen, M.A. Riegler & P. Halvorsen, "DDANet: Dual Decoder Attention Network for Automatic Polyp Segmentation," in *International Conference on Pattern Recognition: International Workshops and Challenges*, 2021.

[29] N. T. Duc, N. T. Oanh, N. T. Thuy, T. M. Triet & V. S. Dinh, "ColonFormer: An Efficient Transformer Based Method for Colon Polyp Segmentation," *IEEE Access*, vol. 10, pp. 80575-80586, 2022.

[30] Q. Chang, D. Ahmad, J. Toth, R. Bascom & W.E. Higgins, "ESFPNet: efficient deep learning architecture for real-time lesion segmentation in autofluorescence bronchoscopic video," *Medical Imaging 2023: Biomedical Applications in Molecular, Structural, and Functional Imaging;*, pp. Proceedings Volume 12468, Medical Imaging 2023: Biomedical Applications in Molecular, Structural, and Functional Imaging, 2023.

[31] Y. Huang, D. Tan, Y. Zhang, X. Li & K. Hu, "TransMixer: A Hybrid Transformer and CNN Architecture for Polyp Segmentation," in *IEEE International Conference on Bioinformatics and Biomedicine*, 2022 IEEE International Conference on Bioinformatics and Biomedicine.

[32] H. Lai, Y. Luo, G. Zhang, X. Shen, B. Li, & J. Lu,, "Toward accurate polyp segmentation with cascade boundary-guided attention.," *The Visual Computer*, vol. 39, no. 4, pp. 1453-1469, 2023.

[33] V. Mandujano-Cornejo & J. Montoya-Zegarra, "Polyp2Seg: Improved Polyp Segmentationwith Vision Transformer," in *Medical Image Understanding and Analysis (MIUA).*, 2022.

[34] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin & B. Guo, "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[35] D. Hendrycks & K. Gimpel, "Gaussian Error Linear Units (GELUs)," *arXiv*, 2016.

[36] "Semantic Understanding of Scenes Through the ADE20K Dataset," *International Journal of Computer Vision*, vol. 127, p. 302–321, 2019.

[37] M. Abe, "Swin V2 Unet/Upernet," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/code/abebe9849/swin-v2-unet-upernet. [Accessed January 2023].

[38] R. Wightman, "PyTorch Image Models," in *GitHub Repository*, 2019.

[39] A.G. Roy, N. Navab & C. Wachinger, "Concurrent Spatial and Channel ‘Squeeze & Excitation’ in Fully Convolutional Networks," in *International*

Conference on Medical Image Computing and Computer-Assisted Intervention, 2018.

[40] P. Iakubovskii, "Segmentation Models Pytorch," *GitHub Repository*, 2019.

[41] J. Deng, W. Dong, R. Socher, L. Li, K. Li & Li Fei-Fei,, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[42] L. Zhou, "Spatially Exclusive Pasting: A general data augmentation technique for the polyp segmentation," in *International Joint Conference on Neural Networks (IJCNN)*, 2023.

[43] I. Loshchilov & F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2019.

[44] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever & R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929-1958, 2014.

[45] Y. Gal & Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in *Proceedings of The 33rd International Conference on Machine Learning*, 2016.

[46] M. Combalia, F. Hueto, S. Puig, J. Malvehy & V. Vilaplana, "Uncertainty Estimation in Deep Neural Networks for Dermoscopic Image Classification," in *Computer Vision and Patern Recoginition*, 2020.

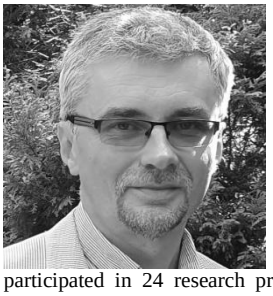
[47] S. Woo, J. Park, J-Y. Lee & IS Kweon, "CBAM: Convolutional Block Attention Module," in *European Conference on Computer Vision*, 2018.

[48] Q. Trinh, "Meta-Polyp: A Baseline for Efficient Polyp Segmentation," in *IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, Aquila, Italy, 2023.

KERR FITZGERALD received a BSc in Physics (Hons) from Durham University (UK) in 2013 where he was also awarded the 'Andy Brinkman Physics Prize'. From 2014 to 2020, he worked as a Fuel Performance Scientist at the UK National Nuclear Laboratory and was selected to serve as a UK representative for the Jules-Horowitz Reactor Fuel Working Group. Kerr is now a final-year Ph.D. candidate within the Computer Vision and Machine Learning (CVML) Group at the University of Central Lancashire. His research focuses on AI applications for medical image analysis, specializing in colonoscopy image classification and segmentation.

JORGE BERNAL received the Ph.D. degree in Computer Science from Universitat Autònoma de Barcelona in 2012. He is currently an Associate Professor at Computer Science Department at Universitat Autònoma de Barcelona and an Associated Researcher at Computer Vision Center, where he is the leader of the Image Sequence Evaluation lab. He has participated in 9 research projects, leading 3 of them, with funding secured from the Spanish Government and the Catalan Government. Currently he is the Principal Investigator of the Spanish Government Funded Project ALETHEIA 'Zero Forgetting in Neural Networks: Continual Learning with Anomaly Detection in Image Sequences' and the coordinator of the first Spanish network devoted to the development of AI systems for colonoscopy, also funded by the Spanish Government. He has published over 20 research papers, won several awards for his research, and supervised 3 PhDs to successful completion. His research interests include computer vision and AI (machine learning) within areas of healthcare technologies, segmentation, object characterization, and scene understanding.

AYMERIC HISTACE is currently a Professeur des Universités at ENSEA, French graduated School of Engineering (Bac+5). He is Deputy Director of the School, and Head of Research, Innovation and Partnerships. Aymeric does research in Computer Vision, Signal and Image Processing in interaction with Natural Science, Engineering, Medicine, and Information Science at ETIS lab (UMR 8051, ENSEA, UCP, CNRS). He is head of the CELL team which activities is mostly dedicated to Smart, Reliable, and Reconfigurable Embedded Systems for various domains. He has co-authored more than 150 papers (conferences and journals) in the domain of Computer Vision mainly. The current flagship projects he is working on are Smart Videocoloscopy, M2-SKAN (Non-Invasive Micro-Vascular Network Segmentation for early diagnostic of Type II diabetes), INSECTS (Innovation for Automatic Recognition of Blood-Sucking Insects), and HYPER-EYE dedicated to Machine Learning for non-conventional sensors (event-based camera).



BOGDAN J. MATUSZEWSKI (Member, IEEE) received the Ph.D. degree in electronic engineering from the Wrocław University of Science and Technology, Poland, in 1996. He is currently a Professor in Computer Vision, the Deputy Director of the Research Centre in Engineering, and the Head of the Computer Vision and Machine Learning (CVML) Group, University of Central Lancashire, Preston, U.K. He has participated in 24 research projects, leading 11 of them, with funding secured from the U.K. Research Councils, EU, and industry. Most recently, he has been the Principal Investigator of the Science and Technology Facilities Council CDN+ funded ''Machine Learning System for Decision Support and Computational Automation of Early Cancer Detection and Categorization in Colonoscopy (AIdDeCo)'' project. He has published over 150 research papers, won several awards for his research, and supervised 19 PhDs to successful completion. His research interests include computer vision, data science, and AI (machine learning) within areas of imaging, healthcare technologies, digital engineering, deformation modelling, segmentation, registration, object characterization, and 3D scene reconstruction and understanding.