FCN-Transformer Feature Fusion for Polyp Segmentation

Edward Sanderson(✉) and Bogdan J. Matuszewski

Computer Vision and Machine Learning (CVML) Group,
University of Central Lancashire, Preston, UK
{esanderson4,bmatuszewski1}@uclan.ac.uk
https://www.uclan.ac.uk/research/activity/cvml

Abstract. Colonoscopy is widely recognised as the gold standard procedure for the early detection of colorectal cancer (CRC). Segmentation is valuable for two significant clinical applications, namely lesion detection and classification, providing means to improve accuracy and robustness. The manual segmentation of polyps in colonoscopy images is time-consuming. As a result, the use of deep learning (DL) for automation of polyp segmentation has become important. However, DL-based solutions can be vulnerable to overfitting and the resulting inability to generalise to images captured by different colonoscopes. Recent transformer-based architectures for semantic segmentation both achieve higher performance and generalise better than alternatives, however typically predict a segmentation map of $\frac{h}{4} \times \frac{w}{4}$ spatial dimensions for a $h \times w$ input image. To this end, we propose a new architecture for full-size segmentation which leverages the strengths of a transformer in extracting the most important features for segmentation in a primary branch, while compensating for its limitations in full-size prediction with a secondary fully convolutional branch. The resulting features from both branches are then fused for final prediction of a $h \times w$ segmentation map. We demonstrate our method's state-of-the-art performance with respect to the mDice, mIoU, mPrecision, and mRecall metrics, on both the Kvasir-SEG and CVC-ClinicDB dataset benchmarks. Additionally, we train the model on each of these datasets and evaluate on the other to demonstrate its superior generalisation performance.

Code available: https://github.com/CVML-UCLan/FCBFormer.

Keywords: Polyp segmentation · Medical image processing · Deep learning

1 Introduction

Colorectal cancer (CRC) is a leading cause of cancer mortality worldwide; e.g., in the United States, it is the third largest cause of cancer deaths, with 52,500 CRC deaths predicted in 2022 [27]. In Europe, it is the second largest cause of cancer deaths, with 156,000 deaths in 27 EU countries reported in 2020 [7].

G. Yang et al. (Eds.): MIUA 2022, LNCS 13413, pp. 892–907, 2022.
https://doi.org/10.1007/978-3-031-12053-4_65

Colon cancer survival rate depends strongly on an early detection. It is commonly accepted that most colorectal cancers evolve from adenomatous polyps [26]. Colonoscopy is the gold standard for colon screening as it can facilitate detection and treatment during the same procedure, e.g., by using the resect-and-discard and diagnose-and-disregard approaches. However, colonoscopy has some limitations; e.g., It has been reported that between 17%–28% of colon polyps are missed during colonoscopy screening procedures [18,20]. Importantly, it has been assessed that improvement of polyp detection rates by 1% reduces the risk of CRC by approximately 3% [4]. It is therefore vital to improve polyp detectability. Equally, correct classification of detected polyps is limited by variability of polyp appearance and subjectivity of the assessment. Lesion detection and classification are two tasks for which intelligent systems can play key roles in improving the effectiveness of the CRC screening and robust segmentation tools are important in facilitating these tasks.

To improve on the segmentation of polyps in colonoscopy images, a range of deep learning (DL) -based solutions [8,13,14,17,19,22,28,30,32,37] have been proposed. Such solutions are designed to automatically predict segmentation maps for colonoscopy images, in order to provide assistance to clinicians performing colonoscopy procedures. These solutions have traditionally used fully convolutional networks (FCNs) [1,9,10,13–15,17,25,28,39]. However, transformer-based architectures [24,32–34,36] have recently become popular for semantic segmentation and shown superior performance over FCN-based alternatives. This is likely a result of the ability of transformers to efficiently extract features on the basis of a global receptive field from the first layers of the model through global attention. This is especially true in generalisability tests, where a model is trained on one dataset and evaluated on another dataset in order to test its robustness to images from a somewhat different distribution to that considered during training. Some studies have also combined FCNs and transformers/attention mechanisms [3,8,19,22,30,37] in order to combine their strengths in a single architecture for medical image segmentation, however these hybrid architectures do not outperform the highest performing FCN-based and transformer-based models in this task, notably MSRF-Net [28] (FCN) and SSFormer [32] (transformer). One significant limitation of most the highlighted transformer-based architectures is however that the predicted segmentation maps of these models are typically of a lower resolution than the input images, i.e. are not full-size. This is due to these models operating on tokens which correspond to patches of the input image rather than pixels.

In this paper, we propose a new architecture for polyp segmentation in colonoscopy images which combines FCNs and transformers to achieve state-of-the-art results. The architecture, named the Fully Convolutional Branch-TransFormer (FCBFormer) (Fig. 1a), uses two parallel branches which both start from a $h \times w$ input image: a fully convolutional branch (FCB) which returns full-size ($h \times w$) feature maps; and a transformer branch (TB) which returns reduced-size ($\frac{h}{4} \times \frac{w}{4}$) feature maps. The output tensors of TB are then upsampled to full-size, concatenated with the output tensors of FCB along the channel

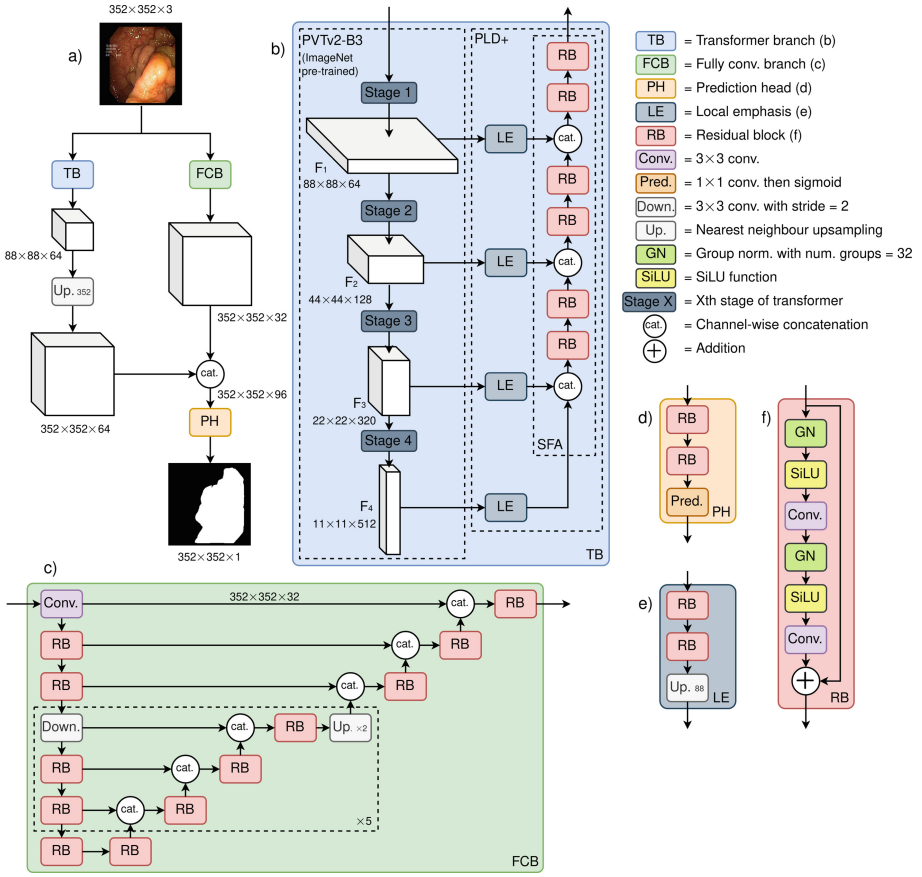

Fig. 1. The architectures of a) FCBFormer, b) the transformer branch (TB), c) the fully convolutional branch (FCB), d) the prediction head (PH), e) the improved local emphasis (LE) module, f) the residual block (RB).

dimension, before a prediction head (PH) processes the concatenated tensors into a full-size segmentation map for the input image. Through the use of the ImageNet [5] pre-trained pyramid vision transformer v2 (PVTv2) [34] as an image encoder, we encourage the model to extract the most important features for segmentation in TB. We then randomly initialise FCB to encourage extraction of the features required for processing outputs of TB into full-size segmentation maps. TB largely follows the structure of the recent SSFormer [32] which predicts segmentation maps of $\frac{h}{4} \times \frac{w}{4}$ spatial dimensions, and which achieved the current state-of-the-art performance on polyp segmentation at reduced-size. However, we update the SSFormer architecture with a new progressive locality decoder (PLD) which features improved local emphasis (LE) and stepwise feature aggregation (SFA). FCB then takes the form of an advanced FCN architecture, composed of a modern variant of residual blocks (RBs) that include group normalisation

[35] layers, SiLU [12] activation functions, and convolutional layers, with a residual connection [11,29]; in addition to dense U-Net style skip connections [25]. PH is then composed of RBs and a final pixel-wise prediction layer which uses convolution with 1×1 kernels. On this basis, we achieve state-of-the-art performance with respect to the mDice, mIoU, mPrecision, and mRecall metrics on the Kvasir-SEG [16] and CVC-ClinicDB [2] datasets, and on generalisability tests where we train the model on one Kvasir-SEG and evaluate it on CVC-ClinicDB, and vice-versa.

The main novel contributions of this work are therefore:

1. The introduction of a simple yet effective approach for FCNs and transformers in a single architecture for dense prediction which, in contrast to previous work on this, demonstrates advantages over these individual model types through state-of-the-art performance in polyp segmentation.
2. The improvement of the progressive locality decoder (PLD) introduced with SSFormer [32] for decoding features extracted by a transformer encoder through residual blocks (RBs) composed of group normalisation [35], SiLU activation functions [35], convolutional layers, and residual connections [11].

The rest of this paper is structured as follows: we first define the design of FCBFormer and its components in Sect. 2; we then outline our experiments in terms of the implementation of methods, the means of evaluation, and our results, in Sect. 3; and in Sect. 4 we give our conclusion.

2 FCBFormer

2.1 Transformer Branch (TB)

The transformer branch (TB) (Fig. 1b) is highly influenced by the current state-of-the-art architecture for reduced-size polyp segmentation, the SSFormer [32]. Our implementation of SSFormer, as used in our experiments, is illustrated in Fig. 2. This architecture uses an ImageNet [5] pre-trained pyramid vision transformer v2 (PVTv2) [34] as an image encoder, which returns a feature pyramid with 4 levels that is then taken as the input for the progressive locality decoder (PLD). In PLD, each level of the pyramid is processed individually by a local emphasis (LE) module, in order to address the weaknesses of transformer-based models in representing local features in the feature representation, before fusing the locally emphasised levels of the feature pyramid through stepwise feature aggregation (SFA). Finally, the fused multi-scale features are used to predict the segmentation map for the input image.

PLD takes the tensors returned by the encoder, with a number of channels defined by PVTv2, and changes the number of channels in the first convolutional layer in each LE block to 64. Each subsequent layer, except channel-wise concatenation and the prediction layer, then returns the same number of channels (64).

The rest of this subsection will specify the design of TB in the proposed FCB-Former and how this varies from this definition of SSFormer. The improvements resulting from our changes are then demonstrated in the experimental section of this paper.

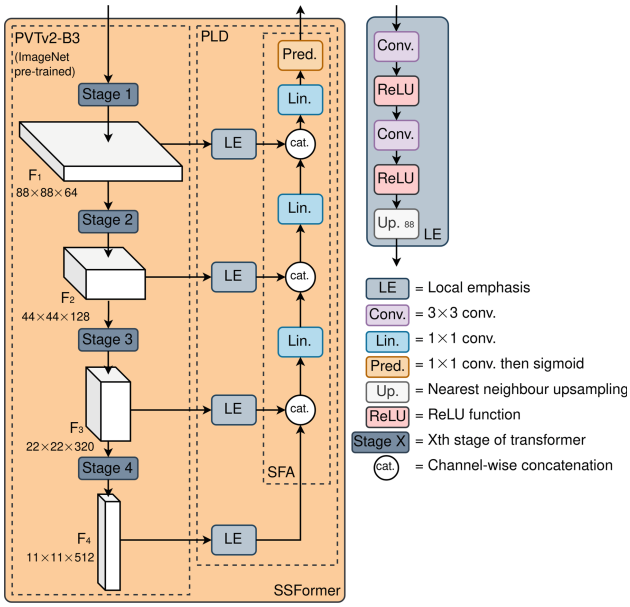

Fig. 2. The architecture of our implementation of SSFormer.

Transformer Encoder. As in SSFormer, we used the PVTv2 [34] for the image encoder in TB, pre-trained on ImageNet [5]. The variant of PVTv2 used is the B3 variant, which has 45.2M parameters. This model demonstrates exceptional feature extraction capabilities for dense prediction owing to its pyramid feature representation, contrasting with more traditional vision transformers which maintain the size of the spatial dimensions throughout the network, e.g. [6,24,31]. Additionally, the model embeds the position of patches through zero padding and overlapping patch embedding via strided convolution, as opposed to adding explicit position embeddings to tokens, and for efficiency uses linear spatial reduction attention. On this element we do not deviate from the design of SSFormer.

Improved Progressive Locality Decoder (PLD+). We improve on the progressive locality decoder (PLD) introduced with SSFormer using the architecture shown in Fig. 1b (PLD+), where we use residual blocks (RBs) (Fig. 1f) to

overcome identified limitations of the SSFormer's LE and SFA. These RBs take inspiration from the components of modern convolutional neural networks which have seen boosts in performance due to the incorporation of group normalisation [35], SiLU activation functions [12], and residual connections [11]. We identified SSFormer's LE and SFA as being limited due to a lack of such modern elements, and a relatively low number of layers. As such, we modified these elements in FCBFormer to form the components of PLD+. The improvements resulting from these changes are shown through ablation tests in the experimental section of this paper.

As in SSFormer, the number of channels returned by the first convolutional layer in the LE blocks 64. Every subsequent layer, except channel-wise concatenation, then returns the same number of channels (64).

2.2 Fully Convolutional Branch (FCB)

We define the fully convolutional branch (FCB) (Fig. 1c) as a composition of residual blocks (RBs), strided convolutional layers for downsampling, nearest neighbour interpolation for upsampling, and dense U-Net style skip connections. This design allows for the extraction of highly fused multi-scale features at full-size, which when fused with the important but coarse features extracted by the transformer branch (TB) allows for inference of full-size segmentation maps in the prediction head (PH).

Through the encoder of FCB, we increase the number of channels returned by each layer by a factor of 2 in the first convolutional layer of the first RB following the second and fourth downsampling layers. Through the decoder of FCB, we then decrease the number of channels returned by each layer by a factor of 2 in the first convolutional layer in the first RB after the second and fourth upsampling layers.

2.3 Prediction Head (PH)

The prediction head (PH) (Fig. 1d) takes a full-size tensor resulted from concatenating the up-sampled transformer branch (TB) output and the output from the fully convolutional branch (FCB). The PH predicts the segmentation map from important but coarse features extracted by TB by fusing them with the fine-grained features extracted by FCB. This approach for the combination of FCNs and transformers for dense prediction to the best of our knowledge has not been used before. As shown by our experiments, this approach is highly effective in polyp segmentation and indicates that FCNs and transformers operating in parallel prior to the fusion of features and pixel-wise prediction on the fused features is a powerful basis for dense prediction. Each layer of PH returns 64 channels, except the prediction layer which returns a single channel.

3 Experiments

To evaluate the performance of FCBFormer in polyp segmentation, we considered 2 popular open datasets, Kvasir-SEG [16][1] and CVC-ClinicDB [2][2], and trained our models using the implementation detailed in Sect. 3.1. These datasets provide 1000/612 (Kvasir-SEG/CVC-ClinicDB) ground truth input-target pairs in total, with the samples in Kvasir-SEG varying in the size of the spatial dimensions while all samples in CVC-ClinicDB are of 288×384 spatial dimensions. All images across both datasets contain polyps of varying morphology. These datasets have been used extensively in the development of polyp segmentation models, and as such provide strong benchmarks for this assessment.

3.1 Implementation Details

We trained FCBFormer to predict binary segmentation maps of $h \times w$ spatial dimensions for RGB images resized to $h \times w$ spatial dimensions, where we set $h, w = 352$ following the convention set by [8,32,37]. We used PyTorch, and due to the aliasing issues with resizing images in such frameworks which have recently been brought to light [23], we used anti-aliasing in our resizing of the images. Both the images and segmentation maps were initially loaded in with a value range of $[0, 1]$. We then used a random train/validation/test split of 80%/10%/10% following the convention set by [8,15,17,28,32], and randomly augmented the training input-target pairs as they were loaded in during each epoch using: 1) a Gaussian blur with a 25×25 kernel with a standard deviation uniformly sampled from $[0.001, 2]$; 2) colour jitter with a brightness factor uniformly sampled from $[0.6, 1.4]$, a contrast factor uniformly sampled from $[0.5, 1.5]$, a saturation factor uniformly sampled from $[0.75, 1.25]$, and a hue factor uniformly sampled from $[0.99, 1.01]$; 3) horizontal and vertical flips each with a probability of 0.5; and 4) affine transforms with rotations of an angle sampled uniformly from $[-180°, 180°]$, horizontal and vertical translations each of a magnitude sampled uniformly from $[-44, 44]$, scaling of a magnitude sampled uniformly from $[0.5, 1.5]$ and shearing of an angle sampled uniformly from $[-22.5°, 22°]$. Out of these augmentations, 1) and 2) were applied only to the image, while the rest of the augmentations were applied consistently to both the image and the corresponding segmentation map. Following augmentation, the image RGB values were normalised to an interval of $[-1, 1]$. We note that performance was achieved by resizing the segmentation maps used for training with bilinear interpolation without binarisation, however the values of the segmentation maps in the validation and test sets were binarised after resizing.

We then trained FCBFormer on the training set for each considered polyp segmentation dataset for 200 epochs using a batch size of 16 and the AdamW optimiser [21] with an initial learning rate of 1e−4. The learning rate was then reduced by a factor of 2 when the performance (mDice) on the validation set

[1] Available: https://datasets.simula.no/kvasir-seg/.
[2] Available: https://polyp.grand-challenge.org/CVCClinicDB/.

did not improve over 10 epochs until reaching a minimum of 1e−6, and saved the model after each epoch if the performance (mDice) on the validation set improved. The loss function used was the sum of the binary cross entropy (BCE) loss and the Dice loss.

For comparison against alternative architectures, we also trained and evaluated a selection of well-established and state-of-the-art examples, which also predict full-size segmentation maps, on the same basis as FCBFormer, including: U-Net [25], ResUNet [38], ResUNet++ [17], PraNet [8], and MSRF-Net [28]. This did not include SSFormer, as an official codebase has yet to be made available and the model by itself does not predict full-size segmentation maps. However, we considered our own implementation of SSFormer in an ablation study presented at the end of this section. To ensure these models were trained and evaluated in a consistent manner while ensuring training and inference was conducted as the authors intended, we used the official codebase[3] provided for each, where possible[4] and modified this only to ensure that the models were trained and evaluated using data of 352×352 spatial dimensions and that the same train/validation/test splits were used.

Some of the codebases for the existing models implement the respective model in TensorFlow/Keras, as opposed to PyTorch as is the case for FCBFormer. After observing slight variation in the results returned by the implementations of the considered metrics in these frameworks for the same inputs, we took steps to ensure a fair and balanced assessment. We therefore predicted the segmentation maps for each assessment within each respective codebase, after training, and saved the predictions. In a separate session using only Scikit-image, we then loaded in the targets for each assessment from source, resized to 352×352 using bilinear interpolation, and binarised the result. The binary predictions were then loaded in, and we used the implementations of the metrics in Scikit-learn to obtain our results. Note that this was done for all models in each assessment.

3.2 Evaluation

We present some example predictions for each model in Fig. 3. From this, it can be seen how FCBFormer predicts segmentation maps which are generally more consistent with the target than the segmentation maps computed by the existing models, and which demonstrate robustness to challenging morphology, highlighted by cases where the existing models are unable to represent the boundary well. This particular strength in segmenting polyps for which the boundary is less apparent is likely a result of the successful combination of the strengths of transformers and FCNs in FCBFormer, leading to the main structures of polyps being dealt with by the transformer branch (TB), while the fully convolutional

[3] ResUNet++ code available: https://github.com/DebeshJha/ResUNetPlusPlus.
PraNet code available: https://github.com/DengPingFan/PraNet.
MSRF-Net code available: https://github.com/NoviceMAn-prog/MSRF-Net.

[4] For U-Net and ResUNet, we used the implementations built into the ResUnet++ codebase (available: https://github.com/DebeshJha/ResUNetPlusPlus).

branch (FCB) serves to ensure a reliable full-size boundary around this main structure. We demonstrate this in Fig. 4, where we show the features extracted by TB and FCB, and the predictions, for examples from the Kvasir-SEG [16] test set. The predictions are shown for the model with FCB, as defined, as well as for the model without FCB, where we concatenate the output of TB channel-wise with a tensor of 0's in place of the output of FCB. This reveals how the prediction head (PH) performs with and without the information provided by FCB, and in turn the role of FCB in assisting with the prediction. The most apparent function is that FCB highlights the edges of polyps, as well as the edges of features that may cause occlusions of polyps, such as other objects in the scene or the perimeter of the colonoscope view. This can then be seen to help provide a well-defined boundary, particularly when a polyp is near or partly occluded by such features.

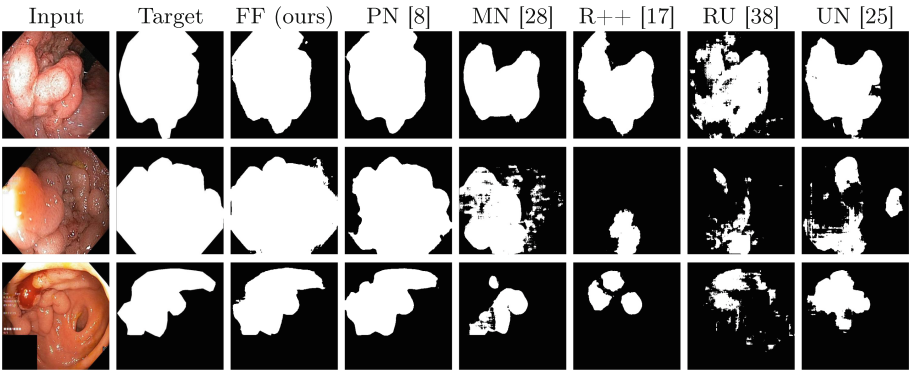

Fig. 3. Example inputs and targets from the Kvasir-SEG test set [16] and the predictions for FCBFormer and the considered existing architectures. FF is FCBFormer, PN is PraNet, MN is MSRF-Net, R++ is ResUNet++, RU is ResUNet, and UN is U-Net. Each model used for this was the variant trained on the Kvasir-SEG training set.

Primary Evaluation. For each dataset, we evaluated the performance of the models with respect to the mDice, mIoU, mPrecision, and mRecall metrics, where m indicates an average of the metric value over the test set. The results from these primary assessments are shown in Table 1, which show that FCB-Former outperformed the existing models with respect to all metrics.

We note that for some of the previously proposed methods, we obtain worse results than has been reported in the original papers, particularly MSRF-Net [28]. This is potentially due to some of the implementations being optimised for spatial dimensions of size 256×256, as opposed to 352×352 as has been used here. This is supported by our retraining and evaluation of MSRF-Net [28] with 256×256 input-targets, where we obtained similar results to those reported in the original paper. We therefore present the results originally reported by the authors of each model in Table 2. Despite the potential differences in the experimental

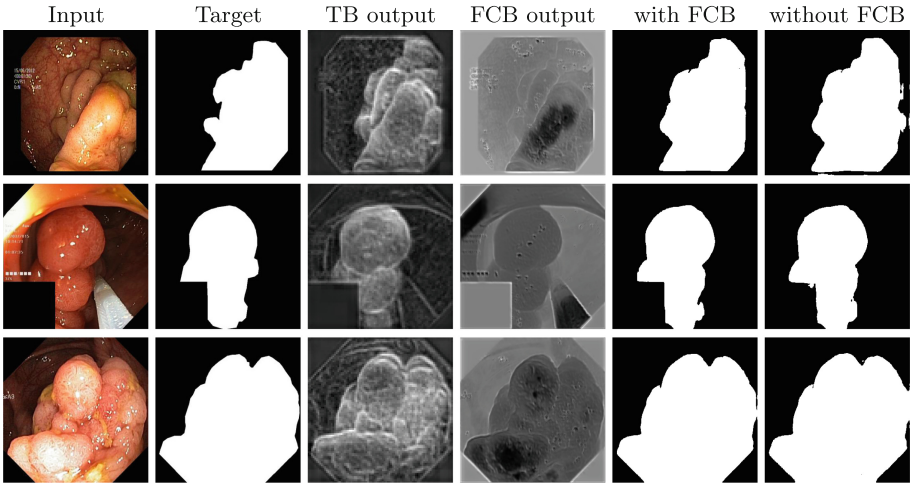

Fig. 4. Visualisation of the features returned by TB and FCB (channel-wise average), and the with/without FCB predictions for examples from the Kvasir-SEG [16] test set.

set up, it can be seen that FCBFormer consistently outperforms other models with respect to the observed mDice, one of the most important metrics out of those considered, and also outperforms other models with respect to mRecall on the Kvasir-SEG dataset [16], and mPrecision on the CVC-ClinicDB dataset [2]. FCBFormer can also be seen to perform competitively with respect to the mIoU.

Table 1. Results from our primary assessment.

Dataset	Kvasir-SEG [16]				CVC-ClinicDB [2]			
Metric	mDice	mIoU	mPrec.	mRec.	mDice	mIoU	mPrec.	mRec.
U-Net [25]	0.7821	0.8141	0.7241	0.8450	0.8464	0.7730	0.8496	0.8796
ResUNet [38]	0.5133	0.3792	0.5937	0.5968	0.5221	0.4120	0.6151	0.5895
ResUNet++ [17]	0.8074	0.7231	0.8991	0.7874	0.5211	0.4126	0.5633	0.5693
MSRF-Net [28]	0.8586	0.7906	0.8933	0.8774	0.9198	0.8729	0.9222	0.9308
PraNet [8]	0.9011	0.8403	0.9034	0.9272	0.9358	0.8867	0.9370	0.93888
FCBFormer (ours)	**0.9385**	**0.8903**	**0.9459**	**0.9401**	**0.9469**	**0.9020**	**0.9525**	**0.9441**

Generalisability Tests. We also performed generalisability tests following the convention set by [28,32]. Using the same set of metrics, we evaluated the models trained on the Kvasir-SEG/CVC-ClinicDB training set on predictions for the full CVC-ClinicDB/Kvasir-SEG dataset. Such tests reveal how models perform with respect to a different distribution to that considered during training.

Table 2. Results originally reported for existing models. Note that U-Net and ResUNet were not originally tested on polyp segmentation, and as such we present the results obtained by the authors of ResUNet++ [17] for these models. For ease of comparison, we include the results we obtained for FCBFormer in our primary assessment.

Dataset	Kvasir-SEG [16]				CVC-ClinicDB [2]			
Metric	mDice	mIoU	mPrec.	mRec.	mDice	mIoU	mPrec.	mRec.
U-Net [25]	0.7147	0.4334	0.9222	0.6306	0.6419	0.4711	0.6868	0.6756
ResUNet [38]	0.5144	0.4364	0.7292	0.5041	0.4510	0.4570	0.5614	0.5775
ResUNet++ [17]	0.8133	0.7927	0.7064	0.8774	0.7955	0.7962	0.8785	0.7022
MSRF-Net [28]	0.9217	**0.8914**	**0.9666**	0.9198	0.9420	**0.9043**	0.9427	**0.9567**
PraNet [8]	0.898	0.840	–	–	0.899	0.849	–	–
FCBFormer (ours)	**0.9385**	0.8903	0.9459	**0.9401**	**0.9469**	0.9020	**0.9525**	0.9441

The results for the generalisability tests are given in Table 3, where it can be seen that FCBFormer exhibits particular strength in dealing with images from a somewhat different distribution to those used for training, significantly outperforming the existing models with respect to most metrics. This is likely a result of the same strengths highlighted in the discussion of Fig. 3.

Table 3. Results from our generalisability tests.

Training data	Kvasir-SEG [16]				CVC-ClinicDB [2]			
Test data	CVC-ClinicDB [2]				Kvasir-SEG [16]			
Metric	mDice	mIoU	mPrec.	mRec.	mDice	mIoU	mPrec.	mRec.
U-Net [25]	0.5940	0.5081	0.6937	0.6184	0.5292	0.4036	0.4613	0.8481
ResUNet [38]	0.3359	0.2425	0.5048	0.3307	0.3344	0.2222	0.2618	0.8164
ResUNet++ [17]	0.5638	0.4750	0.7175	0.5908	0.3077	0.2048	0.3340	0.4778
MSRF-Net [28]	0.6238	0.5419	0.6621	0.7051	0.7296	0.6415	0.8162	0.7421
PraNet [8]	0.7912	0.7119	0.8152	0.8316	0.7950	0.7073	0.7687	**0.9050**
FCBFormer (ours)	**0.8735**	**0.8038**	**0.8995**	**0.8876**	**0.8848**	**0.8214**	**0.9354**	0.8754

As in our primary assessment, we also present results reported elsewhere. Similar generalisability tests were undertaken by the authors of MSRF-Net [28], leading to the results presented in Table 4. Again, we observe that FCBFormer outperforms other models with respect to most metrics.

Ablation Study. We also performed an ablation study, where we started from our implementation of SSFormer given in Fig. 2, since an official codebase has yet to be made available, and stepped towards FCBFormer. We refer to our implementation of SSFormer as SSFormer-I. This model was trained to predict segmentation maps of $\frac{h}{4} \times \frac{w}{4}$ spatial dimensions, and its performance in predicting

Table 4. Results from the generalisability tests conducted by the authors of MSRF-Net [28]. Note, ResUNet [38] was not included in these tests. For ease of comparison, we include the results we obtained for FCBFormer in our generalisability tests.

Training data	Kvasir-SEG [16]				CVC-ClinicDB [2]			
Test data	CVC-ClinicDB [2]				Kvasir-SEG [16]			
Metric	mDice	mIoU	mPrec.	mRec.	mDice	mIoU	mPrec.	mRec.
U-Net [25]	0.7172	0.6133	0.7986	0.7255	0.6222	0.4588	0.8133	0.5129
ResUNet++ [17]	0.5560	0.4542	0.6775	0.5795	0.5147	0.4082	0.7181	0.4860
MSRF-Net [28]	0.7921	0.6498	0.7000	**0.9001**	0.7575	0.6337	0.8314	0.7197
PraNet [8]	0.7225	0.6328	0.7888	0.7531	0.7293	0.6262	0.7623	0.8007
FCBFormer (ours)	**0.8735**	**0.8038**	**0.8995**	0.8876	**0.8848**	**0.8214**	**0.9354**	**0.8754**

full-size segmentation maps was then assessed by upsampling the predictions to $h \times w$ using bilinear interpolation then binarisation. We then removed the original prediction layer and used the resulting architecture as the transformer branch (TB) in FCBFormer, to reveal the benefits of our fully convolutional branch (FCB) and prediction head (PH) for full-size segmentation in isolation of the improved progressive locality decoder (PLD+), and we refer to this model as SSFormer-I+FCB. The additional performance of FCBFormer over SSFormer-I+FCB then reveals the benefits of PLD+. Note that SSFormer-I and SSFormer-I+FCB were both trained and evaluated on the same basis as FCBFormer and the other considered existing state-of-the-art architectures.

The results from this ablation study are given in Tables 5 and 6, which indicate that: 1) there are significant benefits of FCB, as demonstrated by SSFormer-I+FCB outperforming SSFormer-I with respect to most metrics; and 2) there are generally benefits of PLD+, demonstrated by FCBFormer outperforming SSFormer-I+FCB on both experiments in the primary assessment and 1 out of 2 of the generalisability tests, with respect to most metrics.

Table 5. Results from the primary assessment in the ablation study. For ease of comparison, we include the results we obtained for FCBFormer in our primary assessment.

Dataset	Kvasir-SEG [16]				CVC-ClinicDB [2]			
Metric	mDice	mIoU	mPrec.	mRec.	mDice	mIoU	mPrec.	mRec.
SSFormer-I	0.9196	0.8616	0.9316	0.9226	0.9318	0.8777	0.9409	0.9295
SSFormer-I+FCB	0.9337	0.8850	0.9330	**0.9482**	0.9410	0.8904	**0.9556**	0.9307
FCBFormer	**0.9385**	**0.8903**	**0.9459**	0.9401	**0.9469**	**0.9020**	0.9525	**0.9441**

Table 6. Results from the generalisability test in the ablation study. For ease of comparison, we include the results we obtained for FCBFormer in our generalisability tests.

Training data	Kvasir-SEG [16]				CVC-ClinicDB [2]			
Test data	CVC-ClinicDB [2]				Kvasir-SEG [16]			
Metric	mDice	mIoU	mPrec.	mRec.	mDice	mIoU	mPrec.	mRec.
SSFormer-I	0.8611	0.7813	0.8904	0.8702	0.8691	0.7986	0.9178	0.8631
SSFormer-I+FCB	**0.8754**	0.8059	**0.8935**	**0.8963**	0.8704	0.7993	0.9280	0.8557
FCBFormer	0.8735	**0.8038**	0.8995	0.8876	**0.8848**	**0.8214**	**0.9354**	**0.8755**

4 Conclusion

In this paper, we introduced the FCBFormer, a novel architecture for the segmentation of polyps in colonoscopy images which successfully combines the strengths of transformers and fully convolutional networks (FCNs) in dense prediction. Through our experiments, we demonstrated the models state-of-the-art performance in this task and how it outperforms existing models with respect to several popular metrics, and highlighted its particular strengths in generalisability and in dealing with polyps of challenging morphology. This work therefore represents another advancement in the automated processing of colonoscopy images, which should aid in the necessary improvement of lesion detection rates and classification.

Additionally, this work has interesting implications for the understanding of neural network architectures for dense prediction. The method combines the strengths of transformers and FCNs, by running a model of each type in parallel and concatenating the outputs for processing by a prediction head (PH). To the best of our knowledge, this method has not been used before, and its strengths indicate that there is still a great deal to understand about these different architecture types and the basis on which they can be combined for optimal performance. Further work should therefore explore this in more depth, by evaluating variants of the model and performing further ablation studies. We will also consider further investigation of dataset augmentation for this task, where we expect the random augmentation of segmentation masks to aid in overcoming variability in the targets produced by different annotators.

Acknowledgements. This work was supported by the Science and Technology Facilities Council grant number ST/S005404/1.

Discretionary time allocation on DiRAC Tursa HPC was also used for methods development.

References

1. Ali, S., et al.: Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. Med. Image Anal. **70**, 102002 (2021)
2. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Comput. Med. Imaging Graph. **43**, 99–111 (2015)
3. Chen, J., et al.: TransuNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
4. Corley, D.A., et al.: Adenoma detection rate and risk of colorectal cancer and death. N. Engl. J. Med. **370**(14), 1298–1306 (2014)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
6. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. In: ICLR (2021)
7. Dyba, T., et al.: The European cancer burden in 2020: incidence and mortality estimates for 40 countries and 25 major cancers. Eur. J. Cancer **157**, 308–347 (2021)
8. Fan, D.-P., et al.: PraNet: parallel reverse attention network for polyp segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 263–273. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_26
9. Guo, Y.B., Matuszewski, B.: Giana polyp segmentation with fully convolutional dilation neural networks. In: Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, pp. 632–641. SCITEPRESS-Science and Technology Publications (2019)
10. Guo, Y., Bernal, J., J Matuszewski, B.: Polyp segmentation with fully convolutional deep neural networks-extended evaluation study. J. Imaging **6**(7), 69 (2020)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
12. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
13. Huang, C.H., Wu, H.Y., Lin, Y.L.: HardNet-MSEG: a simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps. arXiv preprint arXiv:2101.07172 (2021)
14. Jha, D., et al.: Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. IEEE Access **9**, 40496–40510 (2021)
15. Jha, D., Riegler, M.A., Johansen, D., Halvorsen, P., Johansen, H.D.: Doubleu-net: a deep convolutional neural network for medical image segmentation. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), pp. 558–564. IEEE (2020)
16. Jha, D., et al.: Kvasir-SEG: a segmented polyp dataset. In: Ro, Y.M., et al. (eds.) MMM 2020. LNCS, vol. 11962, pp. 451–462. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_37
17. Jha, D., et al.: Resunet++: an advanced architecture for medical image segmentation. In: 2019 IEEE International Symposium on Multimedia (ISM), pp. 225–2255. IEEE (2019)

18. Kim, N.H., et al.: Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies. Intestinal Res. **15**(3), 411 (2017)
19. Kim, T., Lee, H., Kim, D.: UacaNet: Uncertainty augmented context attention for polyp segmentation. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 2167–2175 (2021)
20. Lee, J., et al.: Risk factors of missed colorectal lesions after colonoscopy. Medicine **96**(27) (2017)
21. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
22. Lou, A., Guan, S., Ko, H., Loew, M.H.: CaraNet: context axial reverse attention network for segmentation of small medical objects. In: Medical Imaging 2022: Image Processing, vol. 12032, pp. 81–92. SPIE (2022)
23. Parmar, G., Zhang, R., Zhu, J.Y.: On aliased resizing and surprising subtleties in GAN evaluation. In: CVPR (2022)
24. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12179–12188 (2021)
25. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
26. Salmo, E., Haboubi, N.: Adenoma and malignant colorectal polyp: pathological considerations and clinical applications. Gastroenterology **7**(1), 92–102 (2018)
27. Siegel, R.L., Miller, K.D., Fuchs, H.E., Jemal, A.: Cancer statistics, 2022. CA Cancer J. Clin. (2022)
28. Srivastava, A., et al.: MSRF-net: a multi-scale residual fusion network for biomedical image segmentation. IEEE J. Biomed. Health Inform. (2021)
29. Srivastava, R.K., Greff, K., Schmidhuber, J.: Highway networks. arXiv preprint arXiv:1505.00387 (2015)
30. Tomar, N.K., et al.: DDANet: dual decoder attention network for automatic polyp segmentation. In: Del Bimbo, A., et al. (eds.) ICPR 2021. LNCS, vol. 12668, pp. 307–314. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-68793-9_23
31. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H.: Training data-efficient image transformers & distillation through attention. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 10347–10357. PMLR, 18–24 July 2021. https://proceedings.mlr.press/v139/touvron21a.html
32. Wang, J., Huang, Q., Tang, F., Meng, J., Su, J., Song, S.: Stepwise feature fusion: local guides global. arXiv preprint arXiv:2203.03635 (2022)
33. Wang, W., et al.: Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568–578 (2021)
34. Wang, W., et al.: Pvtv 2: improved baselines with pyramid vision transformer. Comput. Vis. Media **8**(3), 1–10 (2022)
35. Wu, Y., He, K.: Group normalization. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
36. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SEGFormer: simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems, vol. 34 (2021)

37. Zhang, Y., Liu, H., Hu, Q.: TransFuse: fusing transformers and CNNs for medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 14–24. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_2
38. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual U-net. IEEE Geosci. Remote Sens. Lett. **15**(5), 749–753 (2018)
39. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1