processes

MDPI

Article

GCCSwin-UNet: Global Context and Cross-Shaped Windows Vision Transformer Network for Polyp Segmentation

Jianbo Zhu [1,2], Mingfeng Ge [2,*], Zhimin Chang [2] and Wenfei Dong [1,2]

[1] School of Intelligence and Information Engineering, Shandong University of Traditional Chinese Medicine, Jinan 250355, China; 2021110547@sdutcm.edu.cn (J.Z.)
[2] Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou 215163, China
* Correspondence: gemf@sibet.ac.cn

Abstract: Accurate polyp segmentation is of great importance for the diagnosis and treatment of colon cancer. Convolutional neural networks (CNNs) have made significant strides in the processing of medical images in recent years. The limited structure of convolutional operations prevents CNNs from learning adequately about global and long-range semantic information interactions, despite the remarkable performance they have attained. Therefore, the GCCSwin-UNet framework is suggested in this study. Specifically, the model utilizes an encoder–decoder structure, using the patch-embedding layer for feature downsampling and the CSwin Transformer block as the encoder for contextual feature extraction. To restore the feature map's spatial resolution during upsampling operations, a symmetric decoder and patch expansion layer are also created. In order to help the backbone module to do better feature learning, we also create a global context module (GCM) and a local position-enhanced module (LPEM). We conducted extensive experiments on the Kvasir-SEG and CVC-ClinicDB datasets, and compared them with existing methods. GCCSwin-UNet reached remarkable results with Dice and MIoU of 86.37% and 83.19% for Kvasir-SEG, respectively, and 91.26% and 84.65% for CVC-ClinicDB, respectively. Finally, quantitative analysis and statistical tests are applied to further demonstrate the validity and plausibility of our method.

Keywords: deep learning; colorectal cancer; colonoscopy images; vision transformer; medical image segmentation

check for updates

Citation: Zhu, J.; Ge, M.; Chang, Z.; Dong, W. GCCSwin-UNet: Global Context and Cross-Shaped Windows Vision Transformer Network for Polyp Segmentation. *Processes* **2023**, *11*, 1035. https://doi.org/10.3390/pr11041035

Academic Editor: Jie Zhang

Received: 14 March 2023
Revised: 21 March 2023
Accepted: 24 March 2023
Published: 29 March 2023

1. Introduction

Colorectal cancer (CRC) is the third most prevalent, and the deadliest, cancer worldwide [1]. Therefore, early detection and accurate diagnosis and treatment are key to effective treatment and reduced mortality [2,3]. As one of the most obvious precursors in CRC, accurate localization and segmentation play a key role in early diagnosis and treatment [2,4,5]. Currently, colonoscopy is the most common screening tool used in clinical diagnosis to detect abnormal polyps in the colon [6]. However, the high rate of misdiagnosis and high labour costs make colonoscopy ineffective in diagnosing early to mid-stage polyp lesions [6,7]. To effectively enhance patient outcomes, radiologists can improve diagnostic accuracy and efficiency through the use of computer-aided diagnostic procedures [8].

The different sizes and shapes of polyps present a significant obstacle for routine colonoscopies [9]. The following difficulties with using deep learning for clinical diagnosis were discovered by comparing polyp segmentation with traditional early segmentation algorithms. (1) Traditional convolutional neural networks (CNNs) segmentation algorithms only use global feature information from the last encoder block, which can lead to loss of local feature information in the intermediate layers. (2) Traditional global self-attention mechanisms are complex to compute, while local self-attention mechanisms can limit feature information interaction and do not allow for integrated global and local computation.

(3) Traditional segmentation algorithms focus only on the overall mass distribution of the lesion, ignoring edge and shadow trends, resulting in ambiguous segmentation results for effective diagnosis.

With the aforementioned issues in mind, this research sought to suggest enhanced strategies for improvement. In accordance with the literature [10,11], we contend that global background features help in the segmentation of large polyps, while local background information is crucial for the identification of small polyps. Therefore, the network model needs to have good feature information extraction capability. We used a CSwin Transformer [12] as the foundation of the encoder and decoder for feature extraction to address the first and second problems because we discovered that using a cross-shaped window self-attention mechanism not only reduces computational costs, but also offers powerful feature extraction capability. To prevent the loss of feature information, the GCM is likewise positioned at the top of the encoder, and its output is sent to the upsampling step. The LPEM, a module that directly projects position data onto the linear projection and analyses it as a channel for greater attention to areas such as boundaries, is the tool we use to solve the third challenge. Finally, all of the suggested modules are included in the GCCSwin-UNet polyp segmentation baseline network, which has improved generalization performance and improved detection accuracy. The following are this paper's significant contributions.

(1) We build a symmetric encoder–decoder architecture with a skip-connection structure based on the CSwin Transformer. In the encoder, feature extraction is performed using a cross-shaped window self-attention mechanism with the aim of better extraction of feature information; in the decoder, a patch expansion layer is used to achieve upsampling and feature dimensionality increase without using convolution or interpolation operations to facilitate better polyp segmentation.
(2) We design a global context module with the aim of capturing the feature information that is continuously lost during encoder downsampling and sequentially forwarding it to the corresponding decoder module, with a view to better weighing the global information.
(3) We design a local position-enhanced module that operates on the channel dimension intending to enhance the segmentation of boundary regions by stepping important position information in the feature map.
(4) To verify the segmentation performance of our GCCSwin-UNet, we conduct experiments on two public datasets. The results show that our proposed network not only performs best in polyp segmentation but also achieves state-of-the-art results in two public datasets.

2. Related Work

CNN-based methods: Early methods for polyp segmentation were mainly based on morphology and traditional machine learning classifier algorithms [13–15], which required clearly labelled lesion outlines and large polyp masses, and are not as effective at segmenting small polyps in the early stages and those with fuzzy borders. In recent years, with the development of deep CNNs, UNet [15] was proposed and widely used for medical image segmentation. Due to its good segmentation performance and simple and efficient structure, its translation to polyp segmentation tasks has also been a huge success [16]. Subsequently, various UNet-like methods have emerged, such as SegNet [17], SFANet [18], ResUNet [19], etc., all of which continue to improve polyp segmentation accuracy and efficiency.

Vision transformers: Transformer was initially suggested for machine translation jobs; it was then widely applied in the field of NLP and produced cutting-edge outcomes in a variety of difficulties [20]. Inspired by the success of Transformer, researchers innovatively designed Vision Transformers (ViT) [21], which achieved high accuracy and robustness in image recognition tasks. The major disadvantage of ViT in comparison to CNN-based techniques is the extremely tough pre-training phase that necessitates enormous datasets.

In recent years, researchers have attempted to address the numerous problems of ViT training and quantification and migrated to the field of medical image segmentation. For example, Swin-UNet [22], MedT [23] UTNet [24], etc., have made great advances in medical segmentation and real-time detection. Therefore, inspired by the cross-shaped window displacement mechanism, we try to use a CSwin Transformer as the backbone of feature extraction and add other auxiliary modules to enhance the accuracy of polyp segmentation.

Global Context and Local Position: Because of the feature map's continuous downsampling, which removes feature information, there is a bias in the localization of polyps. Researchers have devised spatial pyramid pooling [25] for feature compensation, and have demonstrated through extensive experiments that this method not only mitigates feature information loss but also improves the robustness of the model to the overall position and layout of the object by extracting spatial information of different sizes. The location information of the markers is disregarded during model training [26] as a result of the invariance of the self-attentive mechanism [27], which leads to hazy local detail segmentation of polyps. Researchers have designed position encoding techniques to be applied in Transformer to enhance the local detail information. Common local position encoding is absolute position encoding (APE) [20], relative position encoding (RPE) [28], and conditional position encoding (CPE) [29]. APE and RPE are usually defined as a series of learnable parameters or frequency functions that generate position encoding with a fixed feature input as the input, making it more difficult to handle inputs of different resolutions. CPE can generate position encoding for arbitrary CPE and can generate position codes for any input resolution, and the generated position codes are then added to the input features.

3. Method

In this section, we describe the architecture of the GCCSwin-UNet and the details of the constituent modules, including the CSwin Transformer block, the global context module (GCM), and the local position enhanced module (LPEM).

3.1. Overall Architecture

The architecture of the GCCSwin-UNet we designed is shown in Figure 1. It is modified from the U-Net and FPN [30]. The architecture uses a symmetrical segmentation system and is accompanied by a skip-connection, the basic unit of which is the CSwin Transformer block. For the encoder, to serialize the input processing, we segment the polyp lesion images into non-overlapping patches, all of 4 × 4 sizes. A patch-embedding layer is added to adjust the dimensionality of the feature map. The two are used in combination, with the CSwin Transformer block responsible for feature learning and the patch-embedding layer responsible for downsampling and dimensionality adjustment. For the decoder, it consists of a CSwin Transformer block and a patch extension layer.

The extracted contextual features are fused with the multi-scale feature output from the GCM by hopping connections to complement the loss of spatial information due to downsampling. In contrast to the patch-embedding layer, the patch extension layer is specifically designed to perform upsampling. The patch extension layer reshapes the feature map of adjacent dimensions into a large feature map with L times (L is the corresponding downsampling multiplier) the upsampling resolution, and this operation restores the feature map to the original input resolution (W × H). Finally, the pixel-level segmentation prediction is output through a linear mapping layer. It is worth noting that we use the CSwin Transformer block with a similar structure to the total of different multi-headed self-attentive mechanisms, but with two changes: (1) using the cross-shaped window displacement mechanism instead of the original attention mechanism, and (2) adding LPEM to each block to enhance local extraction.

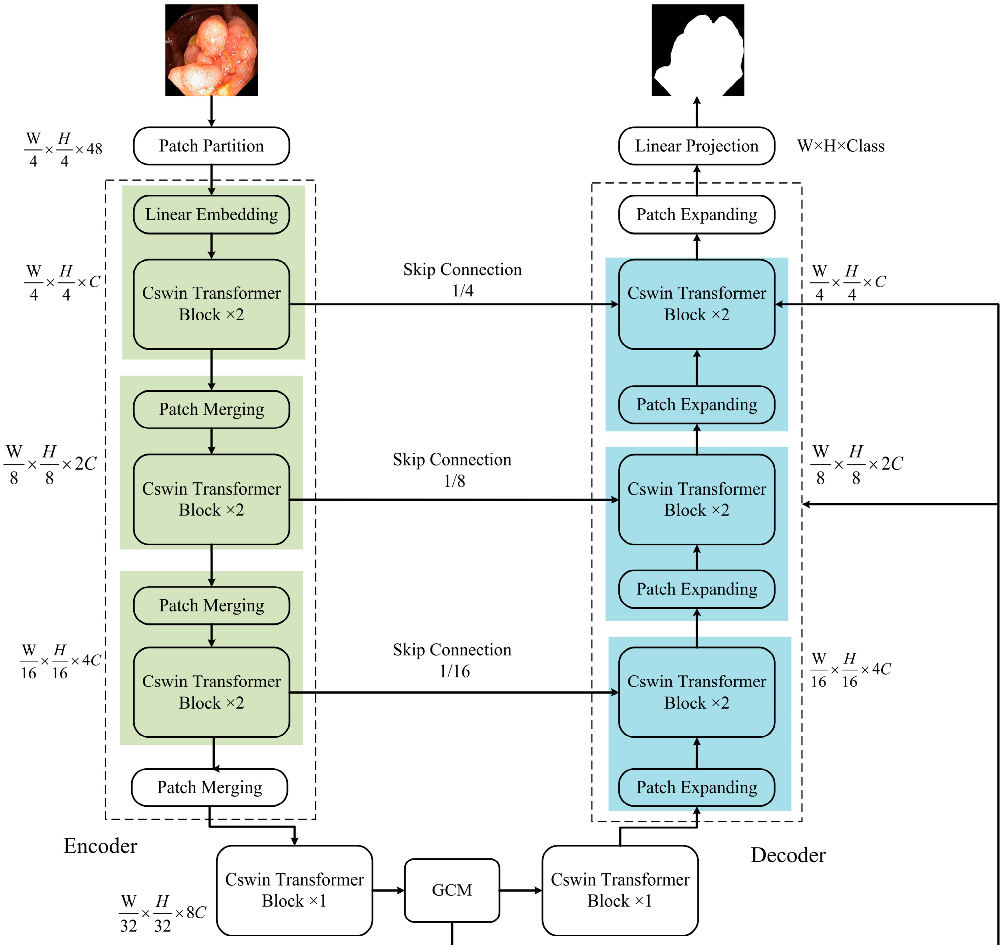
Patch Partition
Linear Projection
W×H×Class
Linear Embedding
Cswin Transformer Block ×2
Patch Merging
Skip Connection 1/4
Skip Connection 1/8
Skip Connection 1/16
Patch Expanding
Encoder
Decoder
Cswin Transformer Block ×1
GCM
$\frac{W}{4}\times\frac{H}{4}\times 48$
$\frac{W}{4}\times\frac{H}{4}\times C$
$\frac{W}{8}\times\frac{H}{8}\times 2C$
$\frac{W}{16}\times\frac{H}{16}\times 4C$
$\frac{W}{32}\times\frac{H}{32}\times 8C$

Figure 1. The overall architecture of our GCCSwin-UNet for polyp segmentation.

3.2. *CSwin Transformer Block*

The multi-headed self-attentive mechanism [20] and the multilayer perceptron (MLP) [31] are the encoder's main building blocks in the standard ViT model. These two components are connected in series by LayerNorm (LN) [32] layers and residual structures, ensuring the stability of the data distribution and the success of the deep network training. In Swin-UNet, a window-based multi-headed self-attentive mechanism and a shifted window-based multi-headed attention mechanism are employed in a two-layer nesting model to gather diverse location information and improve the feature interaction capacity. The comparative analysis shows that although the standard ViT has strong long-range context modeling capability, its multi-headed self-attentive computational complexity is too large, and the computational overheads and training period are huge for high-resolution medical images. Despite the fact that Swin-UNet is enhanced by the shifted window method, each Transformer block still has a high number of attention regions that must be computed more than once, and global feature extraction is accomplished by continually stacking sliding windows. Figure 2 shows the different structures of the three methods.

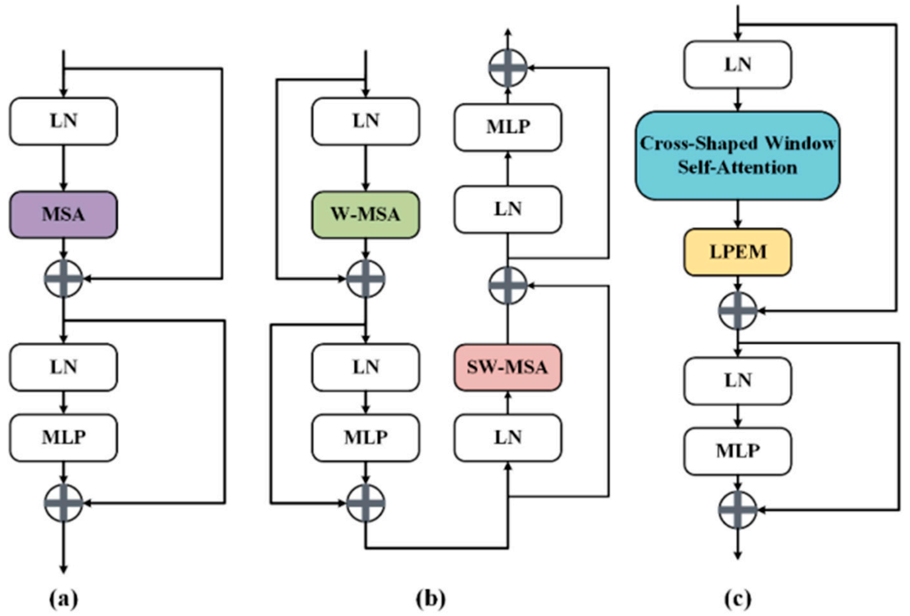

Figure 2. The architecture of three different Transformer structures: (**a**) represents the illustration of the original ViT; (**b**) represents the illustration of the Swin-Transformer; (**c**) the illustration of the CSwin Transformer block.

In order to expand the attention area and extract global feature information more efficiently, a cross-shaped window self-attentiveness mechanism is used in the CSwin Transformer. This module not only reduces the complexity of the computation, but also effectively enhances the information interaction between patches. Specifically, it forms a cross-shaped window by dividing the input features into equal-width stripes. The calculation of weights in the horizontal and vertical directions is performed by translating the sliding cross window. Figure 3 shows a schematic of the operation of the cross-shaped sliding window. Its calculation formula is as follows.

$$\begin{gathered} Y_k^i = Attention(X^i W_k^Q, X^i W_k^K, X^i W_k^V) \\ H - Attention_k(X) / V - Attention_k(X) = [Y_k^1, Y_k^2, \ldots, Y_k^M] \end{gathered} \tag{1}$$

$$\begin{gathered} CSwin - Attention(X) = Concat(head_1, \ldots, head_k) W^o \\ where \;\; head_k = \begin{cases} H - Attention\,_k(X) & k = 1, \ldots, \frac{K}{2} \\ V - Attention\,_k(X) & k = \frac{K}{2} + 1, \ldots, K \end{cases} \end{gathered} \tag{2}$$

where $\mathrm{M} = \frac{\mathrm{H}}{\mathrm{sw}}$, $X^i \in R^{(sw \times W) \times C}$ denotes the input eigenvalues, sw denotes the width of the cross-shaped window, and $W^{\circ} \in R^{C \times c}$ denotes the projection matrix for projecting the results of self-attentiveness to the target output dimension. Self-attentive weights are calculated for *H-Attention (X)* and *V-Attention (X)* in the horizontal and vertical directions, respectively. The computational complexity is as follows.

$$\Omega(CSwin) = HWC \times (4C + sw \times H + sw \times W) \tag{3}$$

Given that H and W will be greater than C in the early stages and smaller than C in the later stages, we choose a small sw in the early stages and a large sw in the latter stages for the high-resolution input. The ability to successfully increase the region of focus for each marker in the later phases is specifically provided by altering the sw.

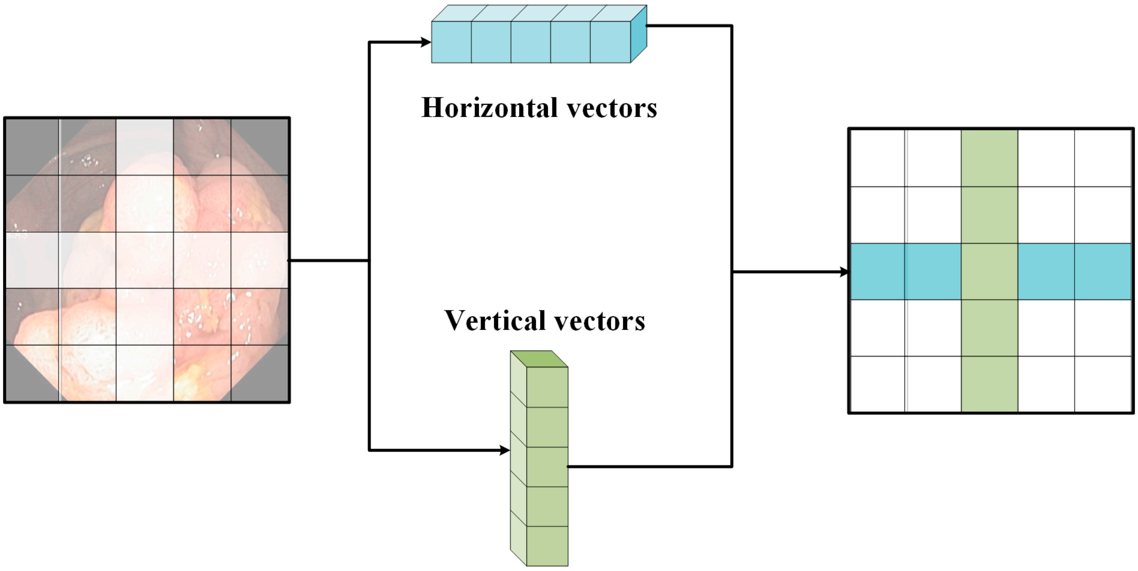

Figure 3. The illustration of Cross-Shaped Window Self-Attention mechanisms.

3.3. Global Context Module (GCM)

The GCM is designed for the problem of loss of feature information in encoder downsampling operations. The module improves the performance and robustness of multi-scale feature extraction by continuously collecting intermediate layer losses and refining them for transfer to the corresponding upsampling layers.

Figure 4 illustrates the structure of the GCM, which uses a multi-branch design for better extraction of information at different feature maps. Specifically, the module consists of a 1 × 1 Conv, three 3 × 3 Atrous Conv [33,34] with different rates and an adaptive level pooling branch [25]. With the feature information collected from the encoder, the GCM uses the above branches to perform extraction and channel concatenate the feature maps at different scales to obtain a global feature map, which is sequentially upsampled and assigned to the CSwin block of the corresponding decoder. The advantage of this mechanism is that it increases the perceptual field while minimising information loss, so that each convolutional output contains more feature information.

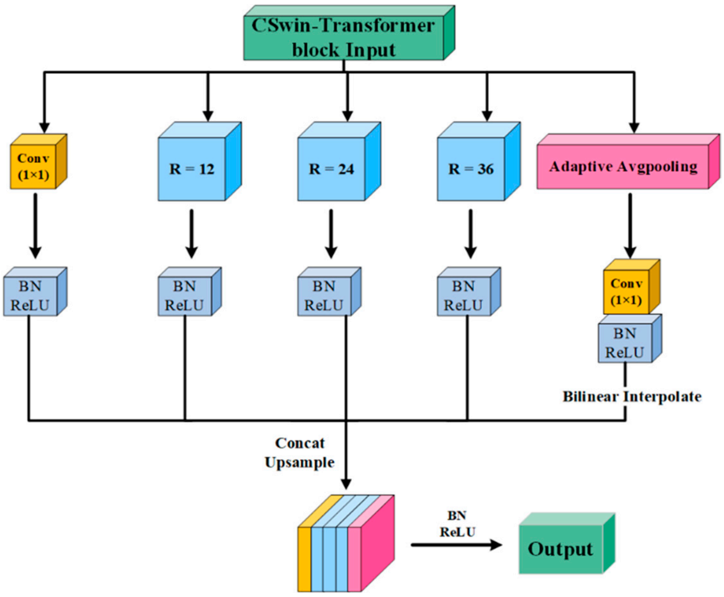

Figure 4. The overall architecture of the Global Context Module (GCM).

3.4. Local Position Enhanced Module (LPEM)

Although cross-shaped window self-attention effectively establishes a long-range dependency between patches, pixel-level features in the patches are ignored. However, this detailed information is particularly important for small polyp samples and edge segmentation. In addition, the fine-grained nature of the feature information can also be enhanced by this method, with coarse-grained features being more easily captured by large patches and fine-grained features being more effective for small patches. In Figure 5, we show some typical location enhancement mechanisms and compare them with our proposed local location enhancement mechanism. Specifically, APE and CPE add location information to the input tokens before the input transformer block, whereas RPE and our LPEM add location information to each transformer block. However, unlike RPE, which adds location information to the attention calculation, we consider a more direct way of imposing location information on the linear projection values. Additionally, we note that RPE introduces bias on a per-attentional-head basis, whereas our LPEM introduces bias on a per-channel basis, which is more intuitive for the positional embedding effect. The formula for its calculation is as follows.

$$\begin{aligned} z_i &= \sum_{j=1}^{n} a_{ij}v_j, a_{ij} = \exp(\frac{q_i^T k_j}{\sqrt{d}}) \\ z_i^k &= \sum_{j=1}^{n} (\alpha_{ij}^k + \beta_{ij}^k)v_j^k \end{aligned} \tag{4}$$

where q_i, k_i, v_i denote the queue, key and value obtained by performing a linear change and self-attentive mechanism on the input x_i. The position-enhanced bias of individual elements is obtained by Equation (4). z_i^k denotes the vector z_i calculation. Therefore, after the weighted bias calculation of the LPEM, the output of the CSwin Transformer block is defined as.

$$\begin{aligned} \hat{X}^l &= CSwin - Attention(LN(X^{l-1})) + X^{l-1} \\ X^l &= MLP(LN(\hat{X}^l)) + \hat{X}^l \end{aligned} \tag{5}$$

where X^l represents the output of the Transformer block of the current layer or the output of the corresponding previous convolutional layer.

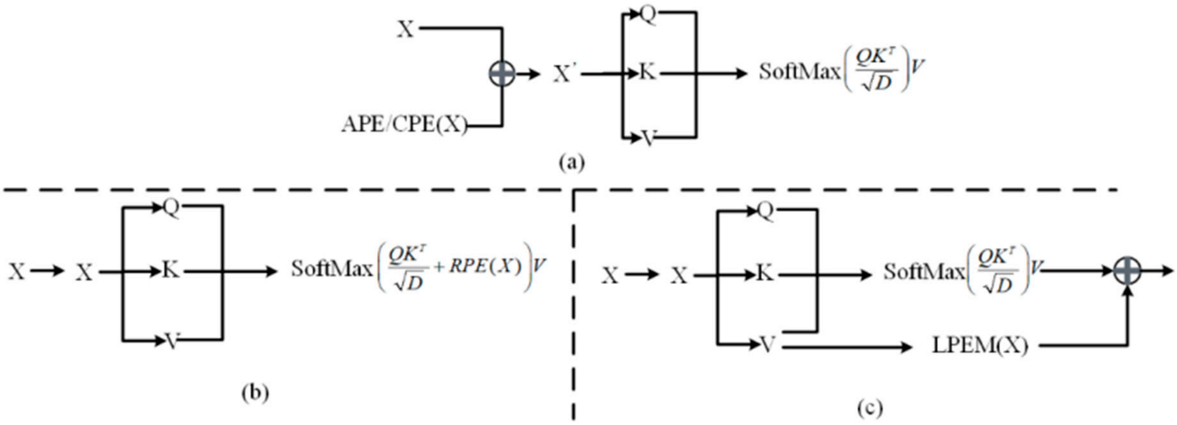

Figure 5. Comparison between different positional enhanced mechanisms: (**a**) APE and CPE (**b**) RPE (**c**) LPEM.

3.5. Mixed Loss Function

The polyp segmentation task is often thought of as a pixel-level classification problem. Each pixel in a high-resolution colonoscopic image will be classified as a polyp or non-polyp site. In general, we use a binary cross-entropy loss function to solve this problem, which is formulated as follows.

$$L_{BEC} = -\frac{1}{N}\sum_{n=1}^{N}\left(y_n \log\left(y'_n\right) + (1-y_n)\log\left(1-y'_n\right)\right) \tag{6}$$

However, in high-resolution colonoscopic images, the majority of pixels are background or non-focal areas, especially many small and discrete polyps. This means that there is a large difference in the proportion of pixels in these two categories, and if the binary cross-entropy loss function is used uniformly for parameter optimization, the final prediction results will move towards the non-focal direction, thus falling into a local optimum and leading to a decrease in the accuracy of the real polyp segmentation results. To solve this problem, we redesigned the loss function with the following formula.

$$L_{DICE} = 1 - \frac{1}{N}\left(\frac{2\sum_{n=1}^{N} y_n y'_n + o}{\sum_{n=1}^{N} y_n + \sum_{n=1}^{N} y'_n + o} + \frac{2\sum_{n=1}^{N}(1-y_n)(1-y'_n) + o}{\sum_{n=1}^{N}(1-y_n) + \sum_{n=1}^{N}(1-y'_n) + o}\right) \tag{7}$$

$$Loss = \lambda_1 L_{BCE}(y_n, {y_n}') + \lambda_2 L_{DICE}(y_n, {y_n}') \tag{8}$$

where L_{DICE} denotes Dice loss, y_n denotes the true value and y'_n denotes the actual output of the model. λ_1, λ_2 are the ratio coefficients.

4. Experiment

In this section, we first introduce the two public polyp datasets and common evaluation metrics, present the experimental results on two polyp datasets, and then interpret the effectiveness of the proposed modules through statistical analysis and visualization.

4.1. Datasets

According to the study of [11], we selected the currently popular datasets, Kvasir [35] and CVC-ClinicDB [36], for a fair comparison with other methods, and used reasonable data augmentation methods to optimize the dataset in order to further enhance the model.

CVC-ClinicDB: The CVC-ClinicDB dataset contains 612 images cut from 31 colonoscopy videos with an image size of 384 × 288 and polyps in different scales and color.

Kvasir: The Kvasir dataset contains 1000 images of different sizes, with polyps varying widely in shape, size, angle and texture.

4.2. Evaluation Metrics

We used three standard metrics to evaluate the segmentation performance of the model, including Dice, MIoU and accuracy. Accuracy represents the proportion of correctly segmented polyp pixels out of all detected sample results, and Dice and MIoU are used to measure the similarity of the network segmentation results to the correct result Mask. The formulae for their calculation are as follows.

$$Dice = \frac{2|X \cap Y|}{|X| + |Y|} = \frac{2TP}{2TP + FP + FN} \tag{9}$$

$$MIoU = \frac{1}{k}\sum_{i=0}^{k}\frac{X \cap Y}{X \cup Y} = \frac{1}{k}\sum_{i=0}^{k}\frac{TP}{TP + FP + FN} \tag{10}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{11}$$

where X is predicted images, Y is the ground-truth, TP is true positive, TN is true negative, FP is false positive and FN is false negative, and k = 2 denotes the weighted category.

4.3. Training Strategies

In the training phase, we used data augmentation to expand the training set, including random horizontal and vertical flipping, rotation and scaling [25]. The processing and training strategy of the literature [11] was followed, and to ensure that the final results are convincing, comparative experiments were conducted using five-fold cross-validation for the two data sets mentioned above. Specifically, 80% of this data was used for model training, 10% for parameter optimisation and validation, and 10% for results testing. In addition during training, the LambdaLR function was designed and implemented to control adaptive changes in the learning rate to prevent overfitting of the model, which is calculated as follows.

$$lr = init\_lr \times (1 - \frac{epoch}{nEpoch})^{power} \tag{12}$$

where *init_lr* = 1×10^{-3}, *nEpoch* = 300, and *power* = 0.9. and pre-trained weights were pre-trained using a CSwin Transformer [12].

4.4. Ablation Experiments

We used various distinct GCCSwin-UNet variations for ablation tests in order to evaluate the efficacy of the suggested approach. The results are displayed in Table 1.

Table 1. Ablation Experiments on Kvasir-SEG.

Setting	***Dice%***	***MIoU%***	***Acc%***
Baseline	79.50	78.33	81.68
Baseline + CSwin	82.45	80.14	83.21
Baseline + GCM	80.78	78.92	82.56
Baseline + LPEM	81.39	79.27	82.75
Baseline + CSwin + LPEM	83.86	82.06	84.54
Baseline + CSwin + GCM	82.93	81.28	84.32
Baseline + GCM + LPEM + CSwin	86.37	83.19	85.94

First, using only the Transformer-UNet of the base self-attention yielded Dice and MIoU of 79.50% and 78.33%, respectively.

The ablation experiment was conducted for the CSwin Transformer block, and considering the changing relationship between input HW and C for high resolution, we were able to flexibly expand the attention area of each marker effectively in the later stages by adjusting the sw. Four stages of the sw were set to 1, 2, 8 and 8 based on previous practical experience. After adding the CSwin, the Transformer Dice and MIoU improved significantly, by 2.9% and 1.8%, respectively. It can be seen that feature extraction via a cross-shaped sliding window can substantially improve the segmentation performance of the polyp region.

The ablation experiments were carried out for GCM and LPEM, which are auxiliary function modules with relatively small-effect improvement for single use. The LPEM is used for spatial detail enhancement, with Dice and MIoU improved by 1.9% and 0.9%, respectively; the GCM is used for network downsampling for feature compensation, and if the moduleitself is relatively poor in feature extraction, the effect of this module will also be greatly weakened, with Dice and MIoU improved by 1.3% and 0.6%, respectively.

When the auxiliary modules GCM and LPEM are used in conjunction with the CSwin Transformer, the segmentation performance can be further improved by enhancing feature extraction while considering channel and detail supplementation. When used with GCM, Dice and MIoU are improved by 3.4% and 2.9%, respectively; when used with LPEM, their Dice and MIoU are improved by 4.3% and 3.7%, respectively. Finally, the best results were achieved when the three were used in unison, with tuning and pre-training operations, with Dice and MIoU of 86.37% and 83.19%, respectively, and a corresponding increase in accuracy of 85.94%.

To verify the effectiveness of the proposed network in our experiments, we validated the model predictions using statistical methods (including Bland–Altman, Pearson and Kappa consistency tests). The results demonstrate that $P_{pearson} = 0.873$, $P_{Kappa} = 0.829$, and the 95% confidence interval is (0.833, 0.906), indicating that the predicted images are strongly correlated with the ground-truth. In summary, there was no significant difference between GCCSwin-UNet and manual segmentation (ρ > 0.05).

4.5. Comparative Experiments

(a) Quantitative analysis

Table 2 shows the results of the quantitative analysis of GCCSwin-UNet against other models, including PraNet [11], UNet [15], SFANet [18], ResUNet [19], Swin-Unet [22] and PNS-Net [37], where PraNet, UNet, SFANet and ResUNet are traditional CNNs methods and Swin-Unet and PNS-Net are Transformer-based methods. During the experiment, all models used the same data augmentation strategy and parameter settings to ensure the fairness of the comparison experiment. The results show that GCCSwin-UNet had the best accuracy (Dice = 86.37, MIoU = 83.19 on Kvasir-SEG, Dice = 91.26, MIoU = 84.65 on CVC-ClinicDB) in both experimental datasets compared to other methods.

Table 2. The comparison of other state-of-the-art networks with our method.

Dataset	Method	*Dice%*	*MIoU%*	*Acc%*
Kvasir-SEG	U-Net [15]	81.80	74.60	82.17
	Residual U-Net [19]	79.10	76.38	73.12
	SFANet [18]	72.35	61.15	——
	PraNet [11]	83.80	81.20	**86.14**
	PNS-Net [37]	84.00	79.50	83.56
	Swin-Unet [22]	82.31	81.62	85.61
	GCCSwin-UNet(ours)	**86.37**	**83.19**	85.94
CVC-ClinicDB	U-Net [15]	87.62	75.50	87.36
	Residual U-Net [19]	86.73	76.17	87.48
	SFANet [18]	70.05	60.75	——
	PraNet [11]	89.82	83.20	91.72
	PNS-Net [37]	87.30	80.00	90.39
	Swin-Unet [22]	88.96	80.71	91.57
	GCCSwin-UNet(ours)	**91.26**	**84.65**	**92.13**

Bold indicates the best result; '——' denotes that the corresponding value is not reported.

In summary, the GCCSwin-UNet model has a stronger ability to interact with semantic information globally and over long distances, and is better than other methods for detail extraction, resulting in better segmentation results.

(b) Qualitative analysis

Figure 6 shows the polyp segmentation results of the GCCSwin-UNet method and some conventional methods on the Kvasir-SEG test set. Our model is able to accurately locate and segment polyps in a variety of challenging situations, such as different sizes, different regions, the presence of noise such as blood mucosa, different textures and different numbers of polyps.

Finally, we screened some colonoscopic images with lesion features for testing. The results are shown in Figure 7. When faced with noise such as blood mucosa, its segmentation results are accurate and clear with intact edges, proving its superiority.

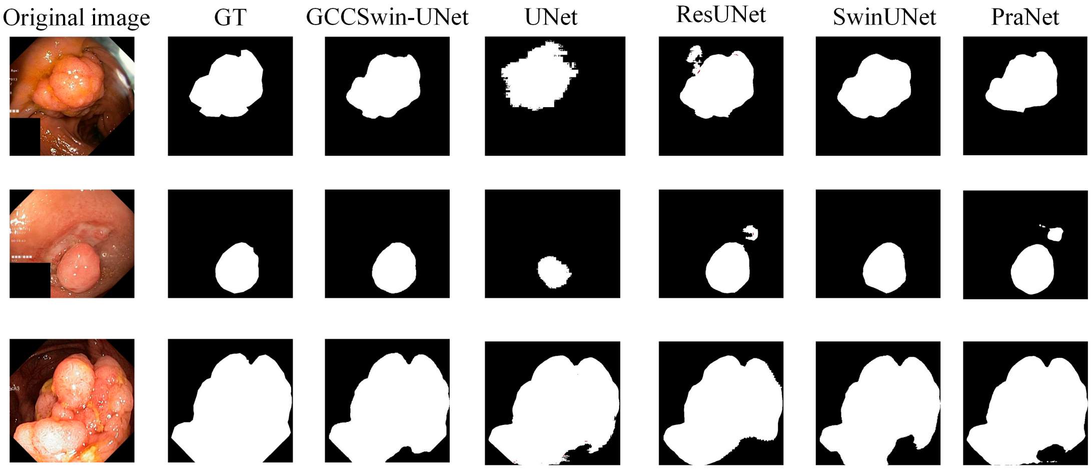

Figure 6. Visual comparison of polyp region segmentation from state-of-the-art methods.

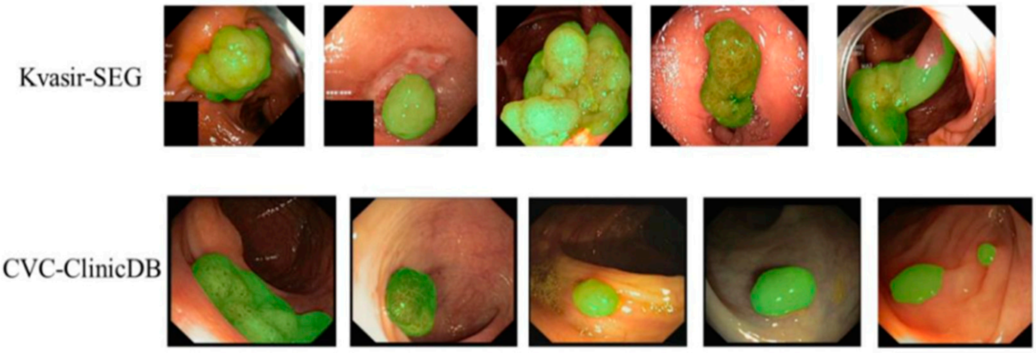

Figure 7. The visualization results of GCCSwin-UNet.

4.6. Discussion

Quantitative analysis showed that the Transformer-based segmentation structure performs better than the traditional CNNs approach. In contrast to Swin-UNet and PNS-Net, which also uses the Transformer block as the backbone structure, the actual segmentation results for polyp lesion images are significantly better. Quantitative experimental results show that conventional methods suffer from poor segmentation and blurred and incomplete edge segmentation for large polyps, and blurred boundaries and poor shape prediction and error noise for small polyps. In contrast, the GCCSwin-UNet segmentation is more precise, covering the lesion area more comprehensively and with well-defined edges. In summary, the GCCSwin-UNet model has a stronger ability to interact with semantic information globally and over long distances, and is better than other methods for detail extraction, resulting in better segmentation results.

The GCCSwin-UNet has demonstrated that it is highly competitive with traditional methods in terms of result accuracy and segmentation effectiveness through numerous comparisons and ablation experiments. However, there are still some uncertainties associated with the model. First, although performance tests have been conducted on publicly available datasets, the models have not yet been applied to clinical validation. Therefore, more realistic and valid clinical data need to be collected for generalisation experiments. Second, GCCSwin-UNet requires long training periods using high performance computing resources, which are not conducive to a lightweight clinical environment.

In the future we will further optimize model performance and improve generalizability in two ways. From the model structure perspective, redundant pruning operations through model quantization and distillation techniques and integration of HD hardware devices will improve clinical utility; from the clinical application perspective, we will consider the complex and variable state of intestinal polyps, such as polyps that are submerged from bruised secretions or obscured by overlays, to further promote its clinical value.

5. Conclusions

This paper presents a framework called GCCSwin-UNet for polyp segmentation based on Vision Transformer. Unlike in traditional CNNs approaches, we incorporate the Transformer idea into the encoder–decoder structure and use the CSwin-Transformer block for representation learning, which not only enhances the information interaction between patches, but also reduces the computational complexity. The auxiliary modules GCM and LPEM are designed. GCM fuses multi-scale feature information at the encoder end to compensate for the loss of global information during downsampling and improve the accuracy of polyp localisation; LPEM acts directly on the channel dimension to focus on the target detail location during feature extraction and thus improve the edge segmentation of the polyp region. In the experimental section analysis, the quantitative and qualitative comparison experiments and the statistical tests of our method are conducted. The results show that our method achieves the best performance (Dice = 86.37, MIoU = 83.19 on Kvasir-SEG, Dice = 91.26, MIoU = 84.65 on CVC-ClinicDB) and that it is statistically significant ($\rho > 0.05$). Finally, we visualize the segmentation results to demonstrate the effectiveness of our proposed method. We hope that this study will provide an inspiration for future clinical polyp segmentation research as well as to explore ever more powerful segmentation models.

Author Contributions: Methodology, conceptualization, software, validation, writing of the original draft, J.Z.; supervision, reviewing, project administration, funding acquisition, M.G.; writing, reviewing, and editing, Z.C.; writing, reviewing, and editing, W.D. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the National Key R&D Program of China (No. 2021YFB3602200).

Data Availability Statement: The data were obtained from the Kvasir-SEG (Jha, Debesh, et al. "Kvasir-seg: A segmented polyp dataset." International Conference on Multimedia Modeling. Springer, Cham, 2020.) and CVC-ClinicDB (Bernal, Jorge, et al. "WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. "Computerized medical imaging and graphics 43 (2015): 99-111.)

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influenced the work reported in this study.

References

1. Siegel, R.L.; Miller, K.D.; Goding Sauer, A.; Fedewa, S.A.; Butterly, L.F.; Anderson, J.C.; Jemal, A. Colorectal cancer statistics, 2020. *CA A Cancer J. Clin.* **2020**, *70*, 145–164. [CrossRef] [PubMed]
2. Barua, I.; Vinsard, D.G.; Jodal, H.C.; Løberg, M.; Kalager, M.; Holme, Ø.; Mori, Y. Artificial intelligence for polyp detection during colonoscopy: A systematic review and meta-analysis. *Endoscopy* **2021**, *53*, 277–284. [CrossRef] [PubMed]
3. Ciardiello, F.; Ciardiello, D.; Martini, G.; Napolitano, S.; Tabernero, J.; Cervantes, A. Clinical management of metastatic colorectal cancer in the era of precision medicine. *CA A Cancer J. Clin.* **2022**, *72*, 372–401. [CrossRef] [PubMed]
4. Tian, Y.; Pu, L.Z.C.T.; Liu, Y.; Maicas, G.; Verjans, J.W.; Burt, A.D.; Carneiro, G. Detecting, localising and classifying polyps from colonoscopy videos using deep learning. *arXiv* **2021**, arXiv:2101.03285.
5. Biller, L.H.; Schrag, D. Diagnosis and treatment of metastatic colorectal cancer: A review. *JAMA* **2021**, *325*, 669–685. [CrossRef]
6. Jha, D.; Ali, S.; Tomar, N.K.; Johansen, H.D.; Johansen, D.; Rittscher, J.; Halvorsen, P. Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. *IEEE Access* **2021**, *9*, 40496–40510. [CrossRef]
7. Le, A.; Salifu, M.O.; McFarlane, I.M. Artificial Intelligence in Colorectal Polyp Detection and Characterization. *Int. J. Clin. Res. Trials* **2021**, *6*, 157. [CrossRef]
8. Brown, J.R.G.; Mansour, N.M.; Wang, P.; Chuchuca, M.A.; Minchenberg, S.B.; Chandnani, M.; Berzin, T.M. Deep learning computer-aided polyp detection reduces adenoma miss rate: A United States multi-center randomized tandem colonoscopy study (CADeT-CS trial). *Clin. Gastroenterol. Hepatol.* **2022**, *20*, 1499–1507. [CrossRef]
9. Turner, J.K.; Wright, M.; Morgan, M.; Williams, G.T.; Dolwani, S. A prospective study of the accuracy and concordance between in-situ and postfixation measurements of colorectal polyp size and their potential impact upon surveillance. *Eur. J. Gastroenterol. Hepatol.* **2013**, *25*, 562–567. [CrossRef]

10. Zhang, R.; Li, G.; Li, Z.; Cui, S.; Qian, D.; Yu, Y. Adaptive context selection for polyp segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; Springer: Cham, Switzerland, 2020.
11. Fan, D.P.; Ji, G.P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. Pranet: Parallel reverse attention network for polyp segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; Springer: Cham, Switzerland, 2020.
12. Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Guo, B. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022.
13. Hwang, S.; Oh, J.; Tavanapong, W.; Wong, J.; De Groen, P.C. Polyp detection in colonoscopy video using elliptical shape feature. In Proceedings of the 2007 IEEE International Conference on Image Processing, San Antonio, TX, USA, 16–19 September 2007; IEEE: Piscataway, NJ, USA, 2007; Volume 2.
14. Gross, S.; Kennel, M.; Stehle, T.; Wulff, J.; Tischendorf, J.; Trautwein, C.; Aach, T. Polyp segmentation in NBI colonoscopy. In *Bildverarbeitung für die Medizin 2009*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 252–256.
15. Du, N.; Wang, X.; Guo, J.; Xu, M. Attraction propagation: A user-friendly interactive approach for polyp segmentation in colonoscopy images. *PLoS ONE* **2016**, *11*, e0155371. [CrossRef]
16. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015.
17. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
18. Fang, Y.; Chen, C.; Yuan, Y.; Tong, K.Y. Selective feature aggregation network with area-boundary constraints for polyp segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019; Springer: Cham, Switzerland, 2019.
19. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual u-net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
21. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
22. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv* **2021**, arXiv:2105.05537.
23. Valanarasu, J.M.J.; Oza, P.; Hacihaliloglu, I.; Patel, V.M. Medical transformer: Gated axial-attention for medical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; Springer: Cham, Switzerland, 2021.
24. Gao, Y.; Zhou, M.; Metaxas, D.N. UTNet: A hybrid transformer architecture for medical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; Springer: Cham, Switzerland, 2021.
25. Zhu, J.; Ge, M.; Chang, Z.; Dong, W. CRCNet: Global-local context and multi-modality cross attention for polyp segmentation. *Biomed. Signal Process. Control* **2023**, *83*, 104593. [CrossRef]
26. Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-attention with relative position representations. *arXiv* **2018**, arXiv:1803.02155.
27. Ho, J.; Kalchbrenner, N.; Weissenborn, D.; Salimans, T. Axial attention in multidimensional transformers. *arXiv* **2019**, arXiv:1912.12180.
28. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.
29. Chu, X.; Tian, Z.; Zhang, B.; Wang, X.; Wei, X.; Xia, H.; Shen, C. Conditional positional encodings for vision transformers. *arXiv* **2021**, arXiv:2102.10882.
30. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
31. Tolstikhin, I.O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Dosovitskiy, A. Mlp-mixer: An all-mlp architecture for vision. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 24261–24272.
32. Xu, J.; Sun, X.; Zhang, Z.; Zhao, G.; Lin, J. Understanding and improving layer normalization. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 4381–4391.
33. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
34. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.

35. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Halvorsen, P.; de Lange, T.; Johansen, D.; Johansen, H.D. Kvasir-seg: A segmented polyp dataset. In Proceedings of the International Conference on Multimedia Modeling, Daejeon, Republic of Korea, 5–8 January 2020; Springer: Cham, Switzerland, 2020.
36. 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.* **2015**, *43*, 99–111. [CrossRef] [PubMed]
37. Ji, G.P.; Chou, Y.C.; Fan, D.P.; Chen, G.; Fu, H.; Jha, D.; Shao, L. Progressively normalized self-attention network for video polyp segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; Springer: Cham, Switzerland, 2021.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.