

Article

HEA-Net: Attention and MLP Hybrid Encoder Architecture for Medical Image Segmentation

Lijing An, Liejun Wang *  and Yongming Li 

College of Information Science and Engineering, Xinjiang University, Urumqi 830000, China

* Correspondence: wljxu@xju.edu.cn; Tel.: +86-139-9981-6618

Abstract: The model, Transformer, is known to rely on a self-attention mechanism to model distant dependencies, which focuses on modeling the dependencies of the global elements. However, its sensitivity to the local details of the foreground information is not significant. Local detail features help to identify the blurred boundaries in medical images more accurately. In order to make up for the defects of Transformer and capture more abundant local information, this paper proposes an attention and MLP hybrid-encoder architecture combining the Efficient Attention Module (EAM) with a Dual-channel Shift MLP module (DS-MLP), called HEA-Net. Specifically, we effectively connect the convolution block with Transformer through EAM to enhance the foreground and suppress the invalid background information in medical images. Meanwhile, DS-MLP further enhances the foreground information via channel and spatial shift operations. Extensive experiments on public datasets confirm the excellent performance of our proposed HEA-Net. In particular, on the GlaS and MoNuSeg datasets, the Dice reached 90.56% and 80.80%, respectively, and the IoU reached 83.62% and 68.26%, respectively.

Keywords: attention; MLP; Transformer

Citation: An, L.; Wang, L.; Li, Y. HEA-Net: Attention and MLP Hybrid Encoder Architecture for Medical Image Segmentation. *Sensors* **2022**, *22*, 7024. <https://doi.org/10.3390/s22187024>

Academic Editor: Sheryl Berlin Brahnam

Received: 8 August 2022

Accepted: 12 September 2022

Published: 16 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Medical image segmentation is a fundamental and critical task in computer-aided diagnosis, which aims to accurately identify the target organ, tissue, or lesion area at the pixel level. As opposed to images of natural scenes, medical images are often complicated in texture and limited by imaging technology and equipment. Although the research [1,2] on artificial intelligence-based enhanced magnetic resonance imaging has made contributions to medical image processing, medical images are still characterized by noise interference, blurring, and are difficult to judge. Moreover, medical image labeling intensively depends on the cognition and experience of the medical experts, which results in significant labeling errors and causes there to be few annotated data records available for training. From the characteristics mentioned above of blurred edges, less training data, and inevitable labeling errors in medical images, the auxiliary diagnosis system based on traditional image segmentation algorithms is not enough to meet the requirements of clinical applications.

In recent years, CNN has gradually become a mainstream image-processing method. Subsequently, researchers proposed a full convolutional neural network (FCN) [3] and U-Net [4] for image segmentation, with the latter being more widely used. Due to its remarkable performance in image segmentation, U-Net is used in the segmentation of the eye [5], heart [6,7], liver, brain [8], skin [9], prostate [10], breast, and other organs. Relying on U-Net, researchers have made great strides in practical application and network performance, such as ANU-Net [11], DIU-Net [12], AttUNet [13], MRUNet [14], HDC-Net [15], and FAC-Net [16]. Although U-Net [4] and its derived models have made great achievements in the field of medical image segmentation, they all face the inevitable problem similar to CNN models: a lack of long-term global correlation modeling capability. The main reason is that CNN simply extracts local information and cannot effectively measure global correlation.

Much recent work has attempted to address this problem by using Transformer. Transformer is a model based on self-attention (SA). SA's ability to model the dependencies between all of the input elements means that Transformer can handle global long-term dependencies. Some recent work has yielded satisfactory results, such as that of Chen et al., who first proposed the Transformer architecture TransUNet [17] for medical image segmentation. It uses CNN to extract features and inputs the extracted features into Transformer for long-range dependency computation. Zhang et al. proposed TransFuse [18] based on ViT [19], which integrates features extracted from Transformer and CNN. Valanarasu et al. proposed MedT [20] based on Axis-attention [21], which proved that Transformer could be applied to small datasets. Since it is difficult to distinguish the boundary of the foreground area of medical images, the local detail information is also important to the segmentation result. However, Transformer focuses on extracting global information while weakening local information, so it also has some shortcomings in the task of medical image segmentation. How to properly strengthen foreground information, suppress useless background information, and how to better attain the joint modeling of local and global features have become the focus of our research.

To solve the above problems, we designed a new attention mechanism, EAM, and proposed DS-MLP. We used EAM to effectively connect the convolution block to Transformer. EAM attention enhances the model's perception of the boundary of the sample foreground region and weakens the background information, thus providing richer foreground information for the global modeling of Transformer. We further propose DS-MLP to strengthen the acquired features and obtain more significant coding ability for our model.

We summarize the contributions of this paper as follows:

- (1) We propose an Efficient Attention Module (EAM) that enhances the foreground and suppresses invalid background information in medical images by performing feature refinement at the channel and spatial level;
- (2) We propose a Dual-channel Shifted MLP module (DS-MLP), which can further enhance foreground information via channel and spatial shift operations.

We briefly review the related network models in Section 2. The third part introduces the general framework of our network and the details of associated modules. The fourth part explicates the datasets and indicators used. The fifth part presents our experimental results and experimental analysis. The sixth part is the conclusion.

2. Related Work

CNN uses convolution as a primary operator to collect the local features and filter images in a sliding window. Typical networks in image segmentation are encoder-decoder-based models, such as R2U-net [22], U-Net [4], DenseUNet [23], UNet++ [24], ResUNet++ [25], etc. They have gradually dominated the field with their excellent performance. However, the local features collected by convolution alone are not enough to distinguish the foreground information from background information in medical images.

In recent years, the attention mechanism has been proved to be a potential means of enhancing deep CNN. Since then, researchers have proposed a variety of attention modules to strengthen the capability of a convolutional neural network. For example, SENet [26] can adaptively adjust the feature response values of each channel and model the internal dependence between the channels. BAM [27] infers attention maps along two different paths (channels and spaces) and embeds them in each bottleneck block. CBAM [28] provides a solution to embed channel and spatial attention submodules sequentially. To avoid the neglect of cross-dimensional interaction, TAM [29] considered dimensional correlation by rotating the feature mapping. However, existing attention modules are composed of multiple layers of convolution or pooling, which inevitably increases the complexity of the module.

Natural Language Processing (NLP) was the first field to apply the Transformer architecture. Researchers applied Transformer to computer vision joining the popularity of Transformer with NLP applications. Most of the networks used for medical image

segmentation are based on hybrid architectures of Transformer and CNN. For example, Gao et al. proposed UTNet [30], which applied self-attention modules to the encoder and decoder to obtain long-term dependence relations of different scales at the lowest cost. The UCTransNet [31] proposed by Wang et al. utilized the CTrans module as a skip connection to solve the problem of the semantic gap. There are also architectures based on pure Transformer, such as Swin-Unet [32] proposed by Cao et al., where each module of the network is built based on the Swin Transformer block [33].

CNN can only simply extract local information, but cannot effectively measure global relevance. Transformer mainly models the global context of an image, which has limited awareness of local features. Since it is difficult to distinguish the boundary of the foreground area of a medical image, the local detail information is also important to the segmentation result. Based on this, we propose HEA-Net. In our approach, we make use of the attention mechanism to effectively connect the convolution blocks with Transformer to form a powerful, joint local and global model.

3. Methods

The overall architecture of the network is shown in Figure 1. The network consists of four parts: encoder, skip connection, bottleneck block, and decoder. The encoder uses EAM and convolution layers for feature extraction. As shown in Figure 1, EAM enhances the foreground and suppresses the invalid background information in medical images by performing feature refinement at the channel and spatial levels. In order to provide the decoder with more refined foreground information, DS-MLP further enhances the foreground information via the channel and spatial shift operations and embeds it into the bottleneck block. The skip connections and decoders are aligned with UCTransNet [31].

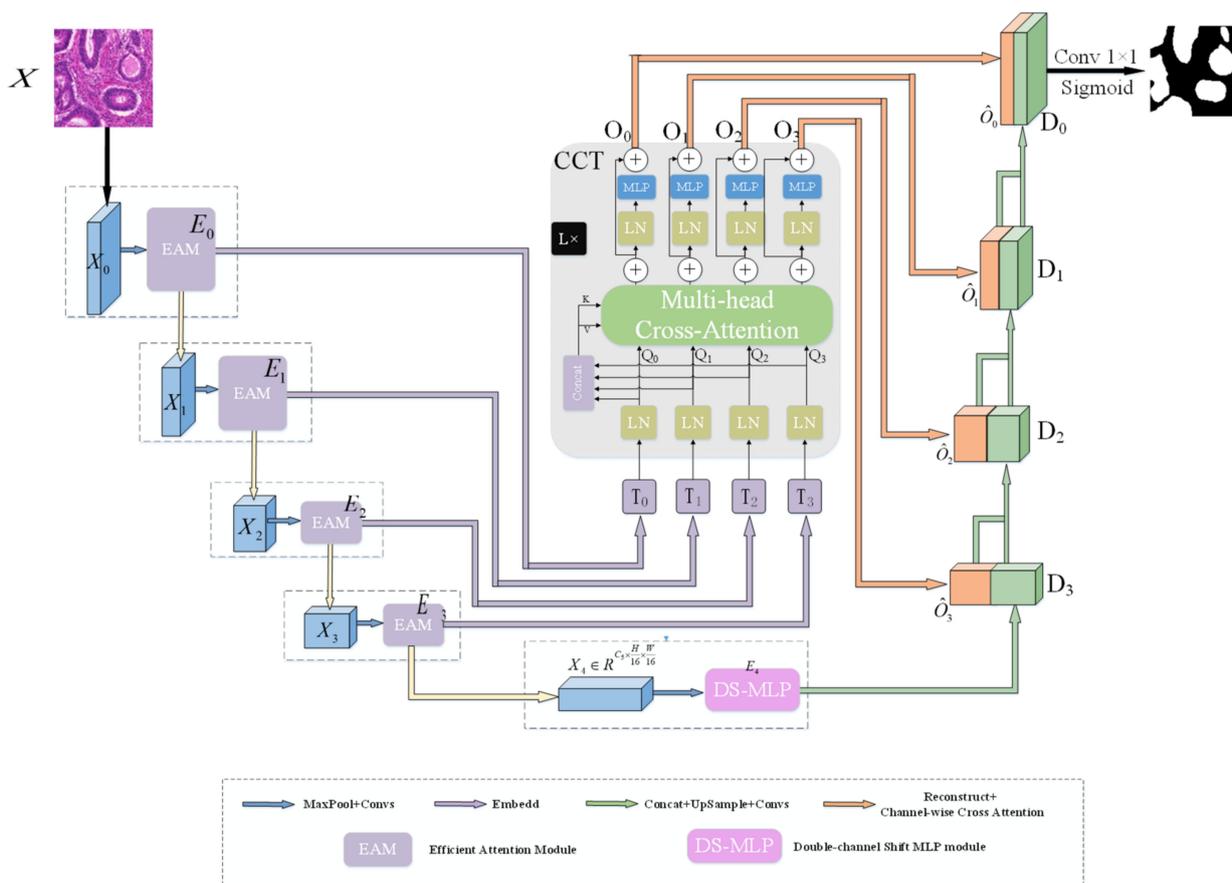


Figure 1. The architecture diagram of HEA-Net.

3.1. Efficient Attention Module (EAM)

Medical research in recent years has shown that the early treatment of most diseases largely depends on the quality of lesion segmentation. However, the exceptionally high similarity between the diseased tissue and the surrounding tissue makes the edges blurred and not easily recognized by the naked eye. This section introduces the EAM module, which can adequately detect and segment the diseased tissue margins and nuclear boundaries.

As shown in Figure 2, the EAM module inputs feature maps into the two branches to estimate the weights of channel dimension and spatial dimension, respectively. In terms of channel dimension, the scale factor (γ) in batch normalization (BN) [34] assigns a unique weight to each channel in the input feature map to extract different feature information from each channel. The weights are calculated as follows [35]:

$$\omega_\gamma = \frac{\gamma_i}{\sum_{j=0} \gamma_j}, \quad (1)$$

where γ is the scale factor; i and j are indexes over the channel dimension. Therefore, the calculation formula for the channel dimension is as follows:

$$\omega_C = \text{sigmoid}(\omega_\gamma \otimes \text{BN}(\omega)), \quad (2)$$

where \otimes denotes the element-wise multiplication; BN denotes batch normalization; and $\omega \in R^{C \times H \times W}$ represents the input image.

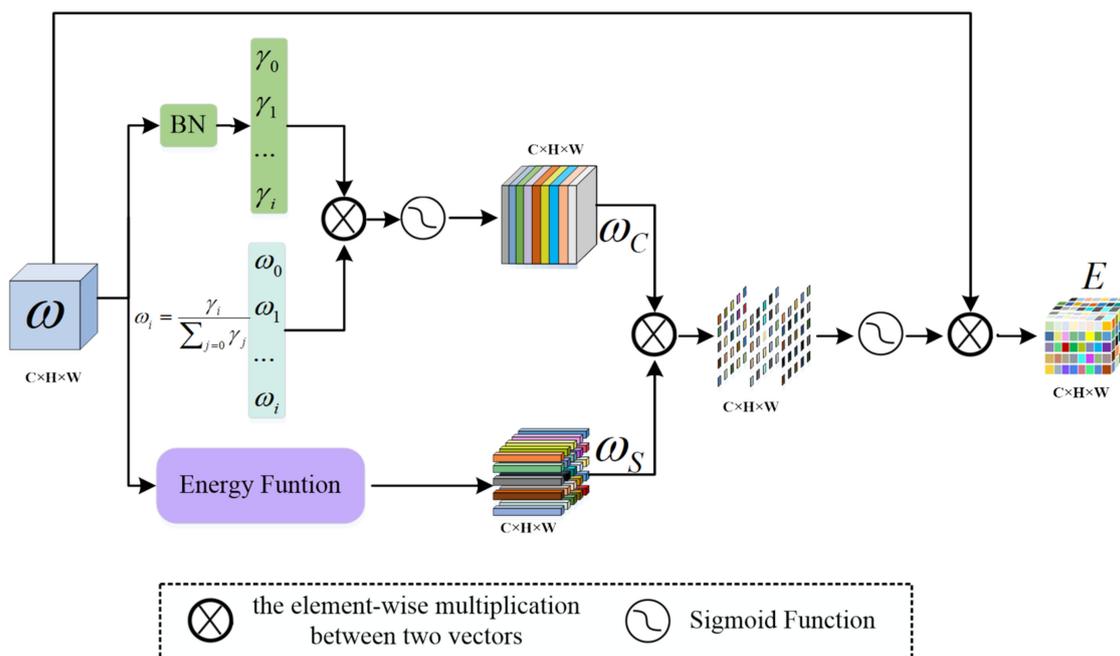


Figure 2. Efficient Attention Module.

In the spatial dimension, the minimum energy function [36] within a single channel is used to estimate the weight of each neuron within the spatial dimension of the image to extract the edge information of the lesion area and nucleus. The minimum energy function [36] was proposed by Yang et al. By summarizing some of the neuroscience findings, they found that the neurons with a large amount of information can show different firing patterns from the surrounding neurons. They define an energy function for each neuron by measuring the linear separability between one target neuron and other neurons. Assuming that all of the pixels on a single channel follow the same distribution, the mean

and variance of all neurons can be calculated and reused for all neurons on that channel [37]. Finally, a minimum energy function is obtained:

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t - \hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda}, \tag{3}$$

where $\hat{\mu} = \frac{1}{M} \sum_{i=1}^M x_i$ and $\hat{\sigma}^2 = \frac{1}{M} \sum_{i=1}^M (x_i - \hat{\mu})^2$; $M = H \times W$ is the number of neurons on that channel; t is the target neuron; and λ is a hyperparameter. The importance of each neuron can be obtained by $\frac{1}{e_t^*}$. In the experiment, λ is set to 0.0001. Therefore, the calculation of the spatial dimension is as follows:

$$\omega_S = \frac{1}{e_t^*} \tag{4}$$

To obtain the importance of each neuron in the channel dimension and the spatial dimension simultaneously, we combined the weight of the channel dimension with the weight of the spatial dimension by multiplying the elements. The output of the EAM module can be expressed as:

$$E = \text{sigmoid}(\omega_C \otimes \omega_S) \otimes \omega, \tag{5}$$

where \otimes denotes the element-wise multiplication and $\omega \in R^{C \times H \times W}$ represents the input image.

3.2. Dual-Channel Shift MLP Module (DS-MLP)

The spatial shift MLP [38] module adopts spatial shift operations to realize the communication between patches. However, the effect of feature enhancement using only spatial shift operation is weak, and Global Maximum Pooling (GMP) can collect more salient features. The DS-MLP module implements further feature enhancement using space shift operations and channel shift operations. The complete structure of the DS-MLP module is shown in Figure 3, and the number of channels and resolution of its output and input feature map remain unchanged.

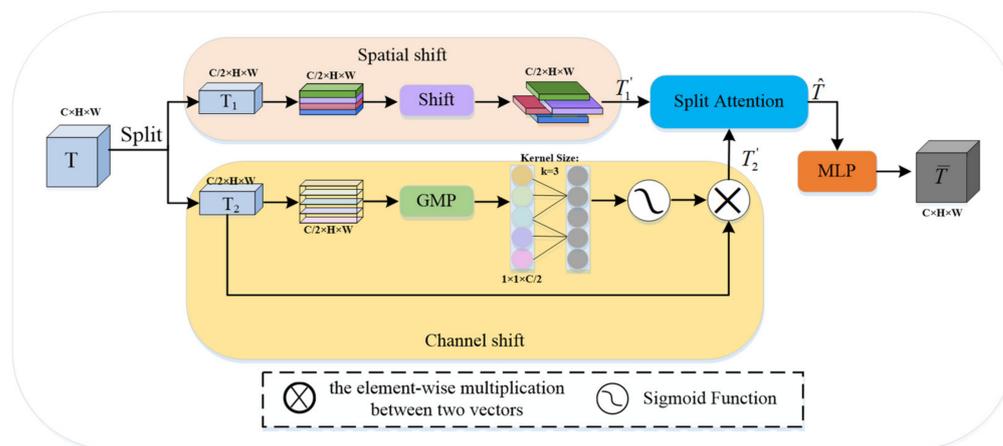


Figure 3. Dual-channel Shift MLP module.

Given an input feature map $T \in R^{W \times H \times C}$, T is divided into two parts along the channel dimension, represented by T_1 and T_2 , respectively, $T_1, T_2 \in R^{W \times H \times \frac{C}{2}}$:

$$T_1 = T \left[:, :, 1 : \frac{C}{2} \right], T_2 = T \left[:, :, \frac{C}{2} + 1 : C \right], \tag{6}$$

where C indicates the number of channels.

We carried out two different shift operations for T_1 and T_2 to achieve better feature enhancement, namely $SS(\cdot)$ and $CS(\cdot)$. $SS(\cdot)$ performs the same spatial shift operation as [38]. That is, T_1 is equally divided into four parts along the channel dimension and moved in four directions:

$$T_{11}[1 : w, :, 1 : c/4] \leftarrow T_{11}[0 : w - 1, :, 1 : c/4], \quad (7)$$

$$T_{12}[0 : w - 1, :, c/4 + 1 : c/2] \leftarrow T_{12}[1 : w, :, c/4 + 1 : c/2], \quad (8)$$

$$T_{13}[:, 1 : h, c/2 : 3c/4] \leftarrow T_{13}[:, 0 : h - 1, c/2 : 3c/4], \quad (9)$$

$$T_{14}[:, 0 : h - 1, 3c/4 : c] \leftarrow T_{14}[:, 1 : h, 3c/4 : c], \quad (10)$$

where w is the width of the feature map; h is the height of the feature map; and c is the number of channels. $CS(\cdot)$ performs the channel shift operation on T_2 . Specifically, we used *GMP* to extract dominant features without dimensionality reduction, then used one-dimensional convolution to carry out channel shift operation to realize cross-channel information interaction, and finally, the final channel shift descriptor is generated by *sigmoid* function and residual connection; the formula is as follows:

$$T_2' = \text{sigmoid}(\text{conv}(\text{GMP}(T_2))) \otimes T_2, \quad (11)$$

where *conv* denotes a one-dimensional convolution of kernel size 3; *GMP* denotes global max pooling; and \otimes denotes the element-wise multiplication. Then, the two feature maps, T_1 and T_2 , obtained through different operations are fused using the Split Attention (SA) module [39] to achieve feature enhancement.

In the SA module, T_1' , T_2' are reshaped into matrices $\{T_k\}_{k=1}^2$. We used $[t_1, t_2, \dots, t_S]$ to represent the S feature maps with a size of $N \times C$, where N is the number of patches and C is the number of channels. The average value is taken:

$$\alpha = \sum_{s=1}^S 1t_s, \quad (12)$$

where $1 \in R^n$ is all the n -dimensional row vectors of 1s, $\alpha \in R^C$. Generated by a series of MLPs and activation functions:

$$\hat{\alpha} = \text{MLP2}(\text{GELU}(\text{MLP1}(\alpha))), \quad (13)$$

where $\hat{\alpha} \in R^C$. Further processing along one dimension with softmax function the following is obtained:

$$\bar{\alpha} = \text{softmax}(\hat{\alpha}), \quad (14)$$

where $\bar{\alpha} \in R^C$. The workflow of the SA module is shown in Figure 4, and the output of the SA module can be expressed as:

$$\hat{T} = t' \otimes t, \quad (15)$$

where \otimes denotes the element-wise multiplication; t is the output of the reshaping operation; t' is the output of the un-squeezing operation.

Finally, the output of the DS-MLP module can be expressed as:

$$\bar{T} = \text{MLP}(\hat{T}), \quad (16)$$

where \hat{T} is the output of the SA module.

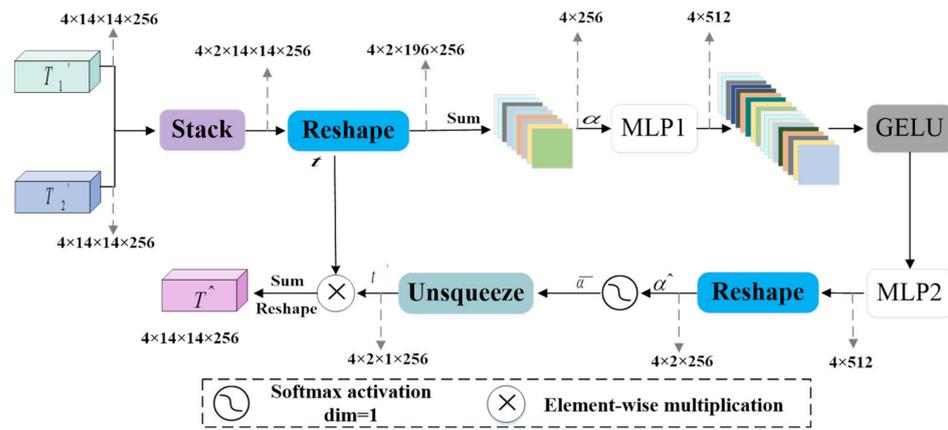


Figure 4. Split Attention module.

3.3. Channel-Wise Cross Fusion Transformer (CCT)

In the skip connection part, this paper uses the CCT module in UTransNet (Figure 1) [31]. This module can not only fuse features from different scales in the encoder but also exchange information across channels. The output of the CCT module is:

$$MCA_i = (CA_i^1 + CA_i^2 + \dots + CA_i^N) / N, \tag{17}$$

$$O_i = MCA_i + MLP(Q_i + MCA_i), \tag{18}$$

where CA_i is a cross-attention (CA); N is the number of heads; O_i is the output of the CCT module; $Q_i = T_i W_{Q_i}$, $W_{Q_i} \in \mathbb{R}^{C_i \times d}$; d is the patch numbers; and C_i are the channel dimensions of the four EAM modules. Layer normalization (LN) is omitted in Formula (18), and the L th layer CCT module can be constructed by repeating the formula (18). In the experiment, N and L are both set to 4.

3.4. Decoder

In the decoder, we use the Channel-wise Cross Attention (CCA) module [31] to fuse the semantic features between the output of the CCT module and the upsampling. The CCA module can filter the features generated by the Transformer and reduce the ambiguity of the decoder features. The detailed structure of CCA is shown in Figure 5.

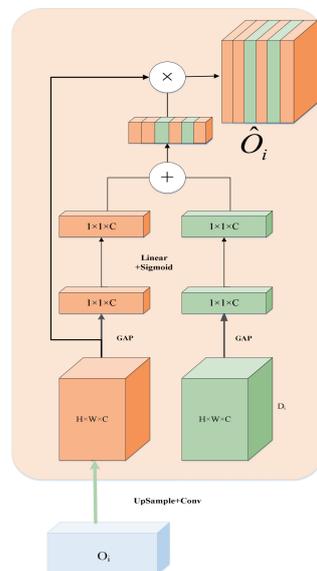


Figure 5. Channel-wise Cross Attention module.

4. Experiments

4.1. Datasets

GlaS [40] is a glandular segmentation dataset from colon tissue images. There were 165 images in this dataset, including 85 in the training set and 80 in the test set. These images are sections of colorectal adenocarcinoma. The slides were processed on different occasions in the laboratory, and each slide belonged to a different patient.

MoNuSeg [41] is a multi-organ nuclear segmentation dataset. The training set of this dataset has 30 images with 21,623 independent nuclei annotated. These images contain seven organs: stomach, prostate, liver, colon, bladder, kidney, and breast. The test set of this dataset has 14 images containing the seven organs of the brain, breast, lung, prostate, bladder, colon, and kidney. The lung and brain tissue images are unique to the test set, making testing more challenging.

4.2. Evaluation Metrics

The experiment uses the Dice coefficient (Dice) and intersection ratio (IoU) to evaluate network performance. These two indicators are related to four values, namely TP, FP, TN, and FN. TP indicates that the predicted result is the same as the actual result, and both are positive, meaning that the prediction is correct. FP implies that the prediction result is positive, but the actual value is negative, indicating that the model misjudges a negative value as a positive value and the model predicts incorrectly. TN demonstrates that the prediction result is the same as the actual, and both values are negative, indicating that the model predicts correctly. FN demonstrated that the prediction result is negative, but the real value is positive, indicating that the model misjudges a positive value as a negative value, and the model predicts incorrectly. The calculation formulas of the Dice and IoU indicators are as follows:

$$Dice = \frac{2TP}{2TP + FP + FN} \quad (19)$$

$$IoU = \frac{TP}{TP + FP + FN} \quad (20)$$

4.3. Implementation Details

We used flipped (vertical and horizontal) and rotated (random) forms of online data augmentation to avoid overfitting. Our proposed HEA-Net is not trained with any pre-trained weights. We set the input resolution size of GlaS to 224×224 , the input resolution size of MoNuSeg to 512×512 , and the batch size to 4. To achieve fast convergence when training the model, we used the Adam optimizer with an initial learning rate set at 0.001. The loss function in which we trained the network employed a combined cross-entropy loss and dice loss. All of our baselines were trained with the same settings and loss function.

5. Results

5.1. Comparative Experiment

The experimental results on the two datasets are shown in Tables 1 and 2. As shown in Table 1, the UNet-based approach still has a good performance, with UNet++ and MRUNet outperforming MedT on GlaS. However, our approach is significantly superior to both the UNet-based and the Transformer-based methods.

In the GlaS dataset, Dice and IoU reached 90.56% and 83.62%, respectively, obtaining the best Dice and IoU. We show the qualitative comparison results in Figure 6. It can be seen that: (1) UNet is more likely to cause over-segmentation or under-segmentation (for example, line 1, line 2, and line 4). (2) In UCTransNet, these phenomena are improved with Transformer. This shows that the Transformer-based hybrid model has strong global context modeling ability. However, due to Transformer's weak local modeling ability, it can be seen from Swin-UNet's forecast segmentation diagram that its segmentation contour is relatively rough. (3) Compared with other models, the segmentation effect of HEA-Net is

better. Its segmentation profile is also smoother. It can accurately identify the edge of the lesion tissue (for example, the first and fourth lines).

Table 1. Comparative experimental results on GlaS datasets. Network results with “*” are derived from [31]. The superscript “1” indicates the best experimental result.

Method	Dice (%)	IoU (%)
U-Net * (2015)	86.34	76.81
UNet++ * (2018)	87.07	78.10
AttUNet * (2018)	86.98	77.53
MRUNet * (2020)	87.72	79.39
UTNet (2021)	83.05	72.32
MedT * (2021)	86.68	77.50
Swin-UNet * (2021)	88.25	79.86
UCTransNet (2021)	89.54	82.26
HEA-Net (Ours)	90.56 ¹	83.62 ¹

Table 2. Comparative experimental results on MoNuSeg datasets. Network results with “*” are derived from [31]. The superscript “1” indicates the best experimental result.

Method	Dice (%)	IoU (%)
U-Net * (2015)	73.97	59.42
UNet++ * (2018)	75.28	60.89
AttUNet * (2018)	76.20	62.64
MRUNet * (2020)	77.54	63.80
UTNet (2021)	78.35	64.74
MedT * (2021)	79.24	65.73
Swin-UNet * (2021)	78.49	64.72
UCTransNet (2021)	79.51	66.15
HEA-Net (Ours)	80.80 ¹	68.26 ¹

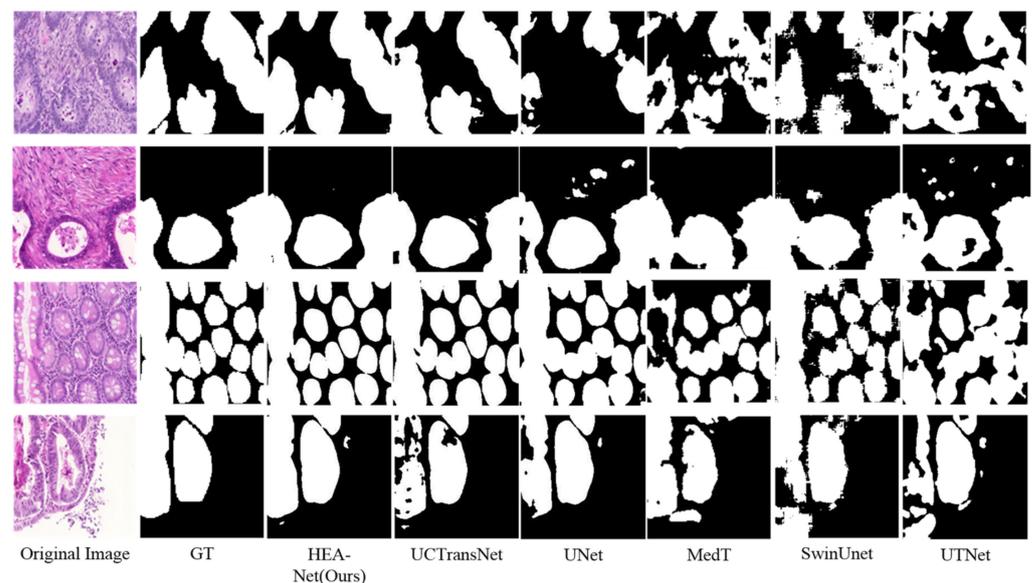


Figure 6. Qualitative results on GlaS dataset.

In the MoNuSeg dataset, Dice and IoU reached 80.80% and 68.26%, respectively. As shown in Figures 7 and 8, we offer the network segmentation results based on UNet and Transformer. It can be seen from the figure that: (1) UNet still has some cases of misclassifying background pixels as foreground pixels (for example, the first line of Figure 7). (2) The Transformer hybrid model ameliorates some of the incorrect predictions. However, Transformer’s poor sensitivity to local information is at a disadvantage for nuclear segmentation

(see lines 1 and 3 in Figure 8). (3) Compared with other models, HEA-Net can capture the nucleus better.

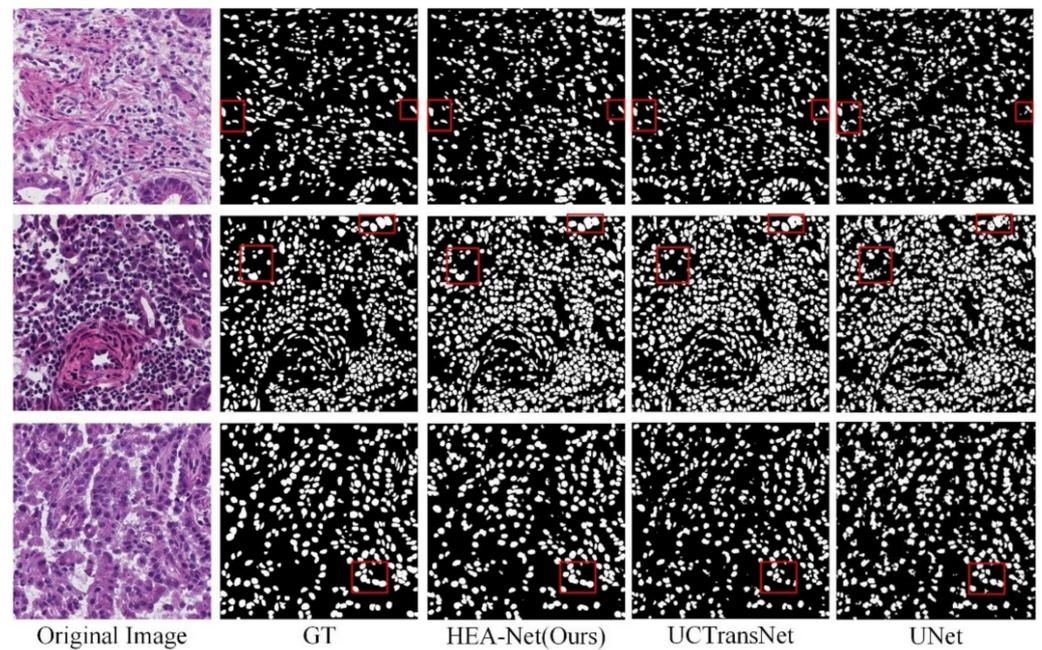


Figure 7. Qualitative results of MoNuSeg dataset based on UNet. In the red box are the parts with significant differences in the segmentation results of each network.

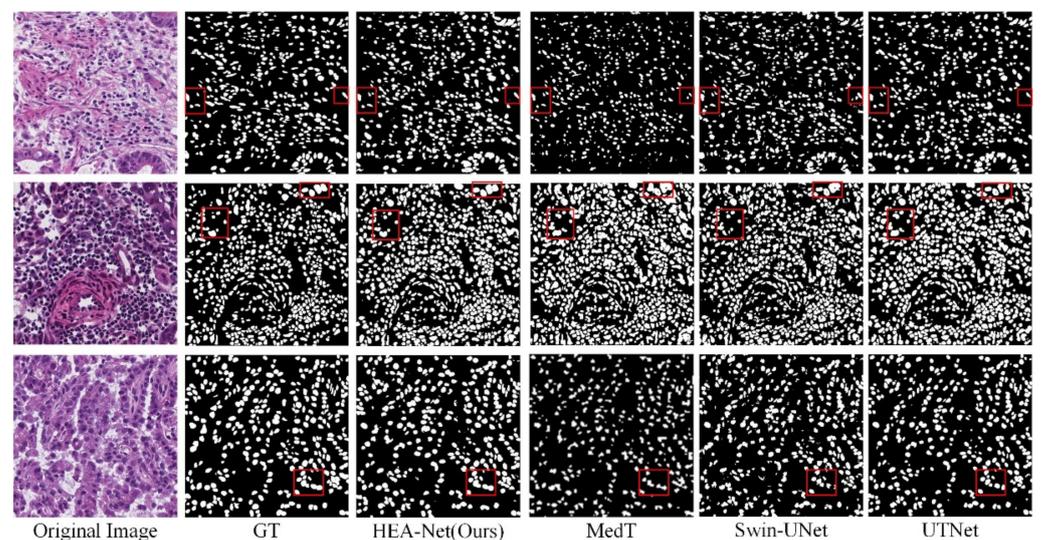


Figure 8. Qualitative results of MoNuSeg dataset based on Transformer. In the red box are the parts with significant differences in the segmentation results of each network.

5.2. Ablation Studies

To intuitively demonstrate the relative contribution of each module, we conducted ablation experiments on the GlaS and MoNuSeg datasets. We compared our model to the baseline (UTransNet). As you can see from Tables 3 and 4, both EAM and DS-MLP are indispensable, as removing either of them can result in network performance degradation. In addition, to demonstrate the critical role of the BN layer in the EAM module in our model, we included Baseline+EAM (without BN) in the ablation experiment. As shown in Tables 3 and 4, the EAM with a BN layer performs better than the EAM without it. This is

because the channel weights generated by the BN layer can highlight the importance of individual channels.

Table 3. Ablation experiments on GlaS datasets. The superscript “1” indicates the best experimental result.

Method	Dice (%)	IoU (%)
Baseline (UCTransNet)	89.54	82.26
Baseline + EAM (without BN)	87.72	79.51
Baseline + EAM	89.86	82.48
Baseline + DS-MLP	90.20	83.04
Baseline + EAM + DS-MLP (Ours)	90.56 ¹	83.62 ¹

Table 4. Ablation experiments on MoNuSeg datasets. The superscript “1” indicates the best experimental result.

Method	Dice (%)	IoU (%)
Baseline (UCTransNet)	79.51	66.15
Baseline + EAM (without BN)	77.99	64.82
Baseline + EAM	80.20	67.57
Baseline + DS-MLP	80.41	67.51
Baseline + EAM + DS-MLP (Ours)	80.80 ¹	68.26 ¹

On the GlaS dataset, Dice and IoU are 1.02% and 1.36% higher than the baseline (UC-TransNet), respectively. As shown in Figure 9, our model has remarkable performance in segmenting the edges of diseased tissue. The EAM module assigns unique weights to the pixel values from the channel dimension and spatial dimension, respectively, so as to strengthen the foreground information and weaken the background information. Taking the fourth row in Figure 9 as an example, the EAM module has a more accurate and smooth processing of the boundary of the foreground part, and the segmentation of the large areas is closer to the label graph (GT). DS-MLP simply further enhances the acquired features to avoid information loss. As you can see in the fourth row of Figure 9, the results with the addition of the DS-MLP module are significantly more complete than the baseline (UCTransNet), but slightly worse than the results with the addition of the EAM module. The combination of EAM and DS-MLP can better capture the foreground information of the samples. Therefore, HEA-Net has a solid ability to segment highly similar foreground and background information.

On the MoNuSeg dataset, Dice and IoU are 1.29% and 2.11% higher than the baseline (UCTransNet), respectively. As shown in Figure 10, most of the nucleus boundaries are closely connected, which increases the segmentation difficulty of our model. As can be seen from Figure 10, although the segmentation result of our model is improved compared to the baseline (UCTransNet), it is not very complete compared to the label map. From the fourth row in Figure 10, we derive a set of local enlargements (Figure 11). As shown in Figure 11, the segmentation graph with the DS-MLP module is more complete than that with the EAM module. This is because the DS-MLP module obtains salient features through hierarchical operation, which is better able to avoid the loss of important information. In order to correctly identify the boundary information of the nucleus, the network needs to pay more attention to the extraction of local details. The combination of EAM and DS-MLP can more accurately grasp the nuclear boundary information for local modeling. Thus, the HEA-Net segmentation of the nucleus is complete.

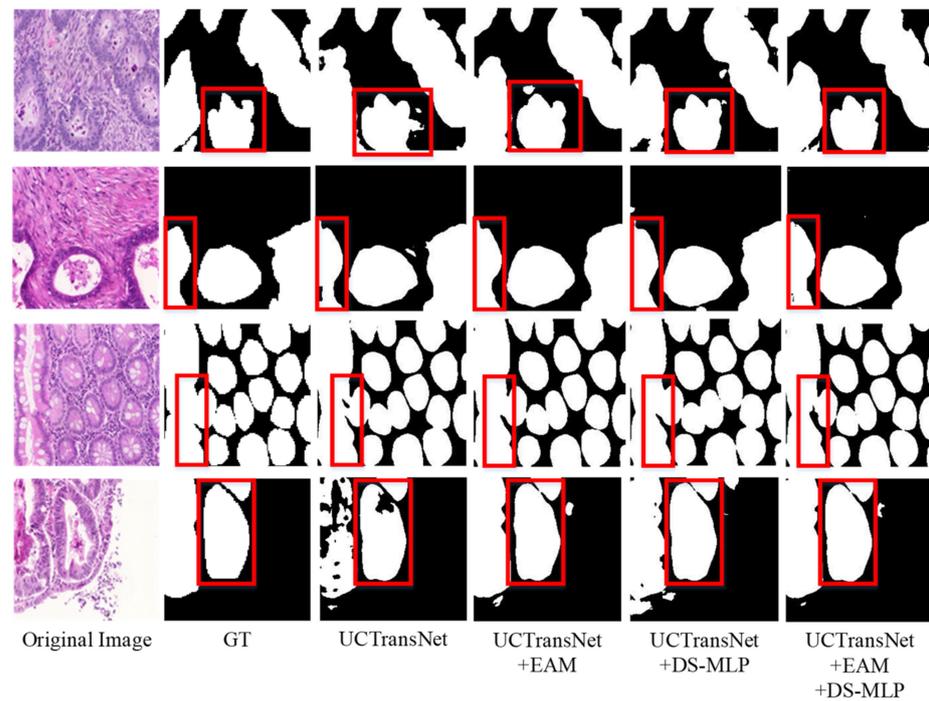


Figure 9. Visual comparison of each module in GlaS dataset. In the red box are the parts with significant differences in the segmentation results of each network.

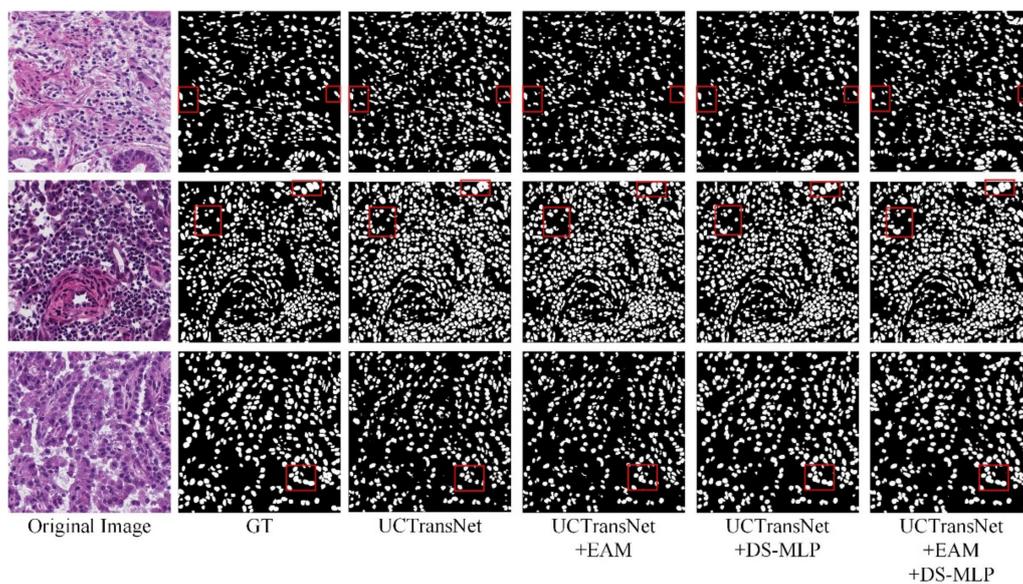


Figure 10. Visual comparison of each module in MoNuSeg dataset. In the red box are the parts with significant differences in the segmentation results of each network.

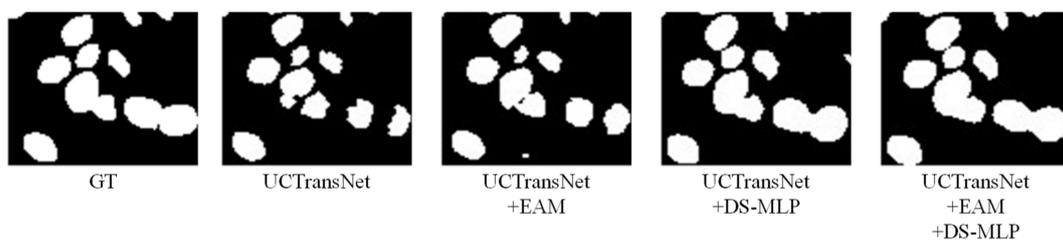


Figure 11. A local enlargement of the visualization result for the fourth line in Figure 10.

6. Conclusions

The accurate segmentation of medical images helps doctors to observe and judge the diseased parts of the human body more effectively. In this work, this paper proposes a hybrid attention encoder architecture (HEA-Net). In HEA-Net, we propose the EAM module can refine the convolutional output at different scales of the encoder network for better local modeling. The DS-MLP module is also proposed to achieve further feature enhancement by shifting operations in the channel and spatial dimensions. We evaluate HEA-Net on two public datasets. The experiments show that our network outperforms the most advanced methods. By analysis of ablation experiments on two different datasets, the overall performance of HEA-Net is significantly improved compared to baseline. Dice and IoU are (1.02%, 1.29%) and (1.36%, 2.11%) higher than baseline, respectively, indicating that the proposed HEA-Net, including EAM and DS-MLP modules, is helpful and valid for the segmentation of lesions.

Author Contributions: Conceptualization, L.A.; methodology, L.A.; software, L.A. and Y.L.; validation, Y.L. and L.W.; formal analysis, L.W. and Y.L.; data curation, L.A.; writing—original draft preparation, L.A.; writing—review and editing, L.A. and L.W.; visualization, L.A. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Tianshan Innovation Team of Xinjiang Uygur Autonomous Region under Grant 2020D14044.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Glas and MoNuSeg datasets are openly available at: https://warwick.ac.uk/fac/cross_fac/tia/data/glascontest (accessed on 8 July 2022) and <https://monuseg.grand-challenge.org/Data/> (accessed on 8 July 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chen, X.; Zhang, X.; Xie, H.; Tao, X.; Wang, F.L.; Xie, N.; Hao, T. A bibliometric and visual analysis of artificial intelligence technologies-enhanced brain MRI research. *Multim. Tools Appl.* **2021**, *80*, 17335–17363. [CrossRef]
2. Chen, X.; Tao, X.; Wang, F.H.; Xie, H. Global research on artificial intelligence-enhanced human electroencephalogram analysis. *Neural Comput. Appl.* **2022**, *34*, 11295–11333. [CrossRef]
3. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
4. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015—18th International Conference, Part III, Munich, Germany, 5–9 October 2015; Volume 9351, pp. 234–241.
5. Reiß, S.; Seibold, C.; Freytag, A.; Rodner, E.; Stiefelhagen, R. Every Annotation Counts: Multi-Label Deep Supervision for Medical Image Segmentation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 9527–9537.
6. He, Y.; Yang, D.; Roth, H.; Zhao, C.; Xu, D. DiNTS: Differentiable Neural Network Topology Search for 3D Medical Image Segmentation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 5837–5846.
7. Gao, Y.; Zhou, M.; Liu, D.; Metaxas, D.N. A Multi-scale Transformer for Medical Image Segmentation: Architectures, Model Efficiency, and Benchmarks. *arXiv* **2022**, arXiv:2203.00131.
8. Pham, Q.D.; Nguyen, T.H.; Phuong, N.N.; Nguyen, K.N.A.; Nguyen, C.D.T.; Bui, T.; Truong, S.Q.H. SegTransVAE: Hybrid CNN—Transformer with Regularization for Medical Image Segmentation. In Proceedings of the 19th IEEE International Symposium on Biomedical Imaging, Kolkata, India, 28–31 March 2022; pp. 1–5.
9. Ji, Y.; Zhang, R.; Wang, H.; Li, Z.; Wu, L.; Zhang, S.; Luo, P. Multi-Compound Transformer for Accurate Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI2021—24th International Conference, Part I, Strasbourg, France, 27 September–1 October 2021; Volume 12901, pp. 326–336.
10. Yan, X.; Jiang, W.; Shi, Y.; Zhuo, C. MS-NAS: Multi-Scale Neural Architecture Search for Medical Image Segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI2020—23rd International Conference, Part I, Lima, Peru, 4–8 October 2020; Volume 12261, pp. 388–397.

11. Li, C.; Tan, Y.; Chen, W.; Luo, X.; He, Y.; Gao, Y.; Li, F. ANU-Net: Attention-based Nested U-Net to exploit full resolution features for medical image segmentation. *Comput. Graph.* **2020**, *90*, 11–20. [[CrossRef](#)]
12. Zhang, Z.; Wu, C.; Coleman, S.; Kerr, D. DENSE-INception U-net for medical image segmentation. *Comput. Methods Programs Biomed.* **2020**, *192*, 105395. [[CrossRef](#)] [[PubMed](#)]
13. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.C.H.; Heinrich, M.P.; Misawa, K.; Mori, K.; McDonagh, S.G.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv* **2018**, arXiv:1804.03999.
14. Ibteha, N.; Rahman, M.S. MultiResUNet: Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation. *Neural Netw.* **2020**, *121*, 74–87. [[CrossRef](#)] [[PubMed](#)]
15. Hu, X.; Wang, L.; Cheng, S.; Li, Y. HDC-Net: A hierarchical dilation convolutional network for retinal vessel segmentation. *PLoS ONE* **2021**, *16*, e0257013. [[CrossRef](#)] [[PubMed](#)]
16. Dong, Y.; Wang, L.; Cheng, S.; Li, Y. FAC-Net: Feedback Attention Network Based on Context Encoder Network for Skin Lesion Segmentation. *Sensors* **2021**, *21*, 5172. [[CrossRef](#)] [[PubMed](#)]
17. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306.
18. Zhang, Y.; Liu, H.; Hu, Q. Transfuse: Fusing transformers and cnns for medical image segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI2021—24th International Conference, Part I, Strasbourg, France, 27 September–1 October 2021; pp. 14–24.
19. Dosovitskiy, A.; Beyer, L.; Kosovitskiy, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the 9th International Conference on Learning Representations, Virtual Event, Austria, 3–7 May 2021.
20. Valanarasu, J.M.J.; Oza, P.; Hacihaliloglu, I.; Patel, V.M. Medical transformer: Gated axial-attention for medical image segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI2021—24th International Conference, Part I, Strasbourg, France, 27 September–1 October 2021; Volume 12901, pp. 36–46.
21. Wang, H.; Zhu, Y.; Green, B.; Adam, H.; Yuille, A.L.; Chen, L.-C. Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation. In Proceedings of the Computer Vision—ECCV2020—16th European Conference, Part IV, Glasgow, UK, 23–28 August 2020; Volume 12349, pp. 108–126.
22. Alom, M.Z.; Hasan, M.; Yakopcic, C.; Taha, T.M.; Asari, V.K. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *arXiv* **2018**, arXiv:1802.06955.
23. Li, X.; Chen, H.; Qi, X.; Dou, Q.; Fu, C.-W.; Heng, P.-A. H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation From CT Volumes. *IEEE Trans. Med. Imaging* **2018**, *37*, 2663–2674. [[CrossRef](#)] [[PubMed](#)]
24. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In Proceedings of the Deep Learning in Medical Image Analysis—and—Multimodal Learning for Clinical Decision Support—4th International Workshop, DLMIA2018, and 8th International Workshop, ML-CDS2018, Held in Conjunction with MICCAI2018, Granada, Spain, 20 September 2018; Volume 11045, pp. 3–11.
25. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Johansen, D.; de Lange, T.; Halvorsen, P.; Johansen, H.D. Resunet++: An advanced architecture for medical image segmentation. In Proceedings of the IEEE International Symposium on Multimedia, San Diego, CA, USA, 9–11 December 2019; pp. 225–230.
26. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [[CrossRef](#)] [[PubMed](#)]
27. Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. BAM: Bottleneck Attention Module. In Proceedings of the British Machine Vision Conference 2018, Newcastle, UK, 3–6 September 2018; p. 147.
28. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the Computer Vision—ECCV2018—15th European Conference, Part VII, Munich, Germany, 8–14 September 2018; Volume 11211, pp. 3–19.
29. Misra, D.; Nalamedda, T.; Arasanipalai, A.U.; Hou, Q. Rotate to attend: Convolutional triplet attention module. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 3138–3147.
30. Gao, Y.; Zhou, M.; Metaxas, D.N. UTransNet: A Hybrid Transformer Architecture for Medical Image Segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI2021—24th International Conference, Part III, Strasbourg, France, 27 September–1 October 2021; Volume 12903, pp. 61–71.
31. Wang, H.; Cao, P.; Wang, J.; Zaiane, O.R. UTransNet: Rethinking the Skip Connections in U-Net from a Channel-wise Perspective with Transformer. In Proceedings of the Twelfth Symposium on Educational Advances in Artificial Intelligence, Virtual Event, 22 February–1 March 2022; pp. 2441–2449.
32. 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.
33. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002.
34. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; Volume 37, pp. 448–456.

35. Liu, Y.; Shao, Z.; Teng, Y.; Hoffmann, N. NAM: Normalization-based Attention Module. *arXiv* **2021**, arXiv:2111.12419.
36. Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. A Simple and Light-Weight Attention Module for Convolutional Neural Networks. *Int. J. Comput. Vis.* **2020**, *128*, 783–798. [[CrossRef](#)]
37. Hariharan, B.; Malik, J.M.; Ramanan, D.K. Discriminative Decorrelation for Clustering and Classification. *Eur. Conf. Comput. Vis.* **2012**, *7575*, 459–472.
38. Yu, T.; Li, X.; Cai, Y.; Sun, M.; Li, P. S²-MLP: Spatial-Shift MLP Architecture for Vision. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 3615–3624.
39. Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Lin, H.; Zhang, Z.; Sun, Y.; He, T.; Mueller, J.; Manmatha, R.; et al. ResNeSt: Split-Attention Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, 19–20 June 2022; pp. 2735–2745.
40. Sirinukunwattana, K.; Pluim, J.P.W.; Chen, H.; Qi, X.; Heng, P.; Guo, Y.B.; Wang, L.Y.; Matuszewski, B.J.; Bruni, E.; Sanchez, U.; et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical Image Anal.* **2017**, *35*, 489–502. [[CrossRef](#)] [[PubMed](#)]
41. Kumar, N.; Verma, R.; Anand, D.; Zhou, Y.; Onder, O.F.; Tsougenis, E.; Chen, H.; Heng, P.; Li, J.; Hu, Z.; et al. A Multi-Organ Nucleus Segmentation Challenge. *IEEE Trans. Med. Imaging* **2020**, *39*, 1380–1391. [[CrossRef](#)] [[PubMed](#)]