GC-UNet: Efficient Network for Medical Image Segmentation

Khaled Alrfou[1*] and Tian Zhao[2]

[1*]Engineering, Computing and Mathematical Sciences, Lewis University, 1 University Parkway, Romeoville, 60446, IL, USA.

[2]Electrical Engineering and Computer Science, University of Wisconsin-Milwaukee, 3203 North Cramer Street, Milwaukee, 53211, WI, USA.

*Corresponding author(s). E-mail(s): kalrfou@lewisu.edu;
Contributing authors: tzhao@uwm.edu;

Abstract

Medical image segmentation is crucial for disease diagnosis and monitoring, but existing methods face challenges in capturing both local and global features efficiently. Convolutional Neural Network (CNN)-based approaches such as UNet, excel at modeling local features but struggle with capturing long-range features. Transformer-based methods, such as Swin-UNet, can model global context but lack the spatial inductive bias needed for local feature extraction. Hybrid methods such as TransUNet and CS-UNet, which combine CNNs and Transformers, have shown promise but often come with increased model complexity and computational cost, limiting their practical applicability. To address these limitations, we propose a neural network GC-UNet a lightweight and efficient segmentation network that leverages the Global Context Vision Transformer (GC-ViT) in its encoder and decoder. GC-UNet combines global context self-attention with local self-attention to model both long and short-range spatial dependencies effectively. For further enhancement, we also introduce two variations of GC-UNet: (1) Hi-GC-UNet, which adds depthwise convolution to improve local feature extraction, and (2) ECA-GC-UNet, which replaces the Squeeze-and-Excitation (SE) block with Efficient Channel Attention (ECA) block to reduce model complexity in the encoders and decoders.

The proposed methods and its variants are evaluated on multiple medical image datasets, including the Synapse multi-organ abdominal CT dataset, the ACDC cardiac MRI dataset, and several Polyp segmentation datasets. In terms of Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD) metrics, GC-UNet outperforms CNN-based and Transformer-based approaches, with notable gains in the segmentation of complex anatomical structures. Hi-GC-UNet performs better than GC-UNet for ACDC dataset with slightly larger model size. ECA-GC-UNet performs better than GC-UNet for most datasets with slightly smaller model size.

Furthermore, pre-training GC-UNet on the MedNet dataset, which contains over 200,000 medical images, yields better performance than pre-training on natural images (ImageNet). The proposed GC-UNet and its variants offer a practical and efficient solution for medical image segmentation, making them suitable for real-world clinical applications.

1 Introduction

Medical image segmentation plays a crucial role in modern healthcare, enabling accurate disease diagnosis, treatment planning, and progression monitoring. Automated segmentation of anatomical structures and pathological regions from medical images, such as CT scans and MRIs, provides valuable quantitative information for clinicians. However, medical image segmentation is a challenging task due to the complexity of anatomical structures, variability in image quality, and the need to capture both local details and global context.

Existing segmentation methods can be broadly categorized into CNN-based and Transformer-based approaches. CNN-based methods, such as UNet [1] and its variants (e.g., UNet++ [2], Att-UNet [3]), excel at capturing local features due to their convolutional operations. However, they struggle to model long-range dependencies, which are essential for segmenting large or complex structures. On the other hand, Transformer-based methods, such as Swin-UNet [4] and TransDeepLab [5], leverage self-attention mechanisms to capture global context. While these methods perform well in modeling long-range dependencies, they often lack the spatial inductive bias needed for effective local feature extraction, leading to suboptimal performance in segmenting small or intricate structures.

Past research explored CNN-Transformer hybrid architectures such as TransUnet [6] to capture global and local information but these models often significantly increase the number of parameters. This, in turn, translates to higher computational complexity, potentially limiting their practical applications.

Recently, Hatamizadeh et al. [7] proposed a Global Context Vision Transformer (GC-ViT) that leverages global context self-attention modules and is joined with local self-attention to effectively and efficiently model both long and short-range spatial interactions. GC-ViT achieved state-of-the-art results across image classification, object detection, and semantic segmentation tasks.

The strength of GC-ViT is its small model size though there is no detailed comparison with the state-of-the-art methods for medical image segmentation. In this paper, we investigate the performance of segmentation algorithms using GC-ViT. To this end, we introduce GC-UNet, a UNet-like segmentation network that captures long and short-range semantic features using GC-ViT [7] as encoders and decoders. This architecture enhances performance while requiring fewer model parameters, with higher inference speed, and lower computational complexity. As further enhancements, we also explored two variations of GC-UNet.

1. Since GC-ViT only uses convolution for downsampling, it is unclear whether parallel application of convolution can improve its performance in smaller objects. To answer this question, we extended GC-UNet with parallel depthwise convolution in the encoders and decoders. The resulting architecture, Hi-GC-UNet, has better performance for ACDC dataset with slightly higher model complexity.
2. GC-ViT uses Squeeze-and-Excitation (SE) block in its feature extraction and downsampling modules to adaptively recalibrate channel-wise feature responses. To further reduce model complexity, we replaced the SE block in GC-ViT with the Efficient Channel Attention (ECA) block, which models channel relationships directly. The resulting architecture, ECA-GC-UNet, has better performance than GC-UNet in most datasets.

We evaluated the segmentation and runtime performance of GC-UNet and its variations on several medical image datasets including Synapse, ACDC, and several polyp image datasets. Our results show that GC-UNet has better or comparable performance than the state-of-the-art segmentation algorithms including CNN-based, Transformer-based, and hybrid segmentation networks. Also, GC-UNet has smaller model sizes and uses the least amount of training and inference time. In addition, we pre-trained GC-UNet on both ImageNet and MedNet, which is a set of 200,000 medical images collected from public sources. GC-UNet pre-trained on in-domain images (i.e. MedNet) yielded better accuracy than GC-UNet pre-trained on natural images (i.e. ImageNet). Hi-GC-UNet has better performance than GC-UNet in ACDC dataset with slight increase in model size. We speculate that the improvement is related to the inclusion of depthwise convolution, which is better at detecting local features of the smaller objects. ECA-GC-UNet has better performance than GC-UNet in most datasets and its model size is smaller due to the use of ECA blocks.

To summarize, we make the following contributions.

1. We propose GC-UNet, a lightweight segmentation network based on GC-ViT, which achieves state-of-the-art performance on multiple medical image datasets.
2. We introduce two variants of GC-UNet (Hi-GC-UNet and ECA-GC-UNet) to further enhance its performance and efficiency.
3. We demonstrate that pre-training GC-UNet on the MedNet dataset, which contains over 200,000 medical images, yields better performance than pre-training on natural images (ImageNet).
4. We provide extensive experimental results showing that GC-UNet outperforms existing CNN-based, Transformer-based, and hybrid methods in terms of segmentation accuracy, model size, and computational efficiency.

The source code for the model is available at github.com/Kalrfou/GC-UNet.

2 Related work

Medical image segmentation has seen significant advancements with the adoption of deep learning techniques. Existing methods can be broadly categorized into CNN-based, Transformer-based, and hybrid approaches, each with its own strengths and limitations.

2.1 CNN-Based Methods

The CNN-based methods are widely used and regarded as one of the most prominent approaches for medical image segmentation. Encoder-decoder based architectures, such as UNet [1] and its derivatives, have shown exceptional efficacy in medical image segmentation. For instance, Att UNet [3] enhanced segmentation through attention gates while UNet++ [8] introduced an alternative skip connection mechanism, nested and dense, alleviating the semantic gap between levels of UNet to a certain degree. This modification yields notable performance improvements compared to UNet. However, UNet++ cannot capture semantic features at full scale. Huang et al. [9] proposed UNet3+ to maximize the use of full-scale feature maps by combining low-level details from various scales with high-level semantics. CNN-based methods have found application in diverse medical image segmentation tasks, such as retinal image segmentation [10] and skin segmentation [11], showcasing promising performance and practicality in implementation and training. Segmentation algorithms based on ResNet architecture have established its presence in medical image segmentation [12]. For example, Res-UNet [13] enhanced retinal vessel segmentation with a weighted attention mechanism.

Despite their success, CNN-based methods struggle to model long-range dependencies due to their localized receptive fields, which limits their performance in segmenting large or complex structures.

2.2 Transformer-Based Methods

Transformers, originally designed for natural language processing, have gained traction in medical image segmentation due to their ability to model global context through self-attention mechanisms. The self-attention mechanism (MSA) inherent in Transformers empowers them to perform global correlation modeling, enabling them to handle long-range dependencies effectively. Leveraging this capability, Transformers have made significant strides in both natural language processing and computer vision tasks due to their superior global modeling abilities. Several pioneering studies have introduced Transformer-based architectures for medical image segmentation. Cao et al. [4] presented Swin-UNet, integrating a Swin Transformer [14] into a U-shaped segmentation network for multi-organ segmentation. Azad et al. [5] proposed TransDeepLab for skin lesion segmentation, enhancing DeepLab with diverse window strategies. Additionally, Huang et al., [15] introduced MISSFormer to leverage global information across different scales for cardiac segmentation, while Azad et al. [16] introduced TransCeption, refining the patch merging module to capture multi-scale representations within a single stage.

Swin Transformer [14] introduced local-window-self-attention to reduce the cost so that it grows linearly with the image size, used shifted-window-attention to capture cross-window information, and exploited multi-resolution information with hierarchical architecture. However, the shifted-window-attention struggles to capture long-range information due to small coverage area of shifted-window-attention and lacks inductive bias like ViT [17].

2.3 Hybrid Methods

While Transformers excel at capturing global context, they often lack the spatial inductive bias inherent in CNNs, making them less effective at modeling local features, especially in small or intricate structures. To address the limitations of both CNNs and Transformers, hybrid approaches have been proposed, combining the strengths of both architectures.

TransUNet [6] integrated Transformers into the encoder of a U-Net, enabling the model to capture both local and global features. However, this approach significantly increases the number of parameters and computational complexity. HiFormer [18] introduced a hierarchical multi-scale Transformer for medical image segmentation, achieving state-of-the-art results on several benchmarks. CS-UNet [19] proposed a generalizable and flexible segmentation algorithm by combining CNNs and Transformers, demonstrating improved performance on diverse medical imaging tasks.

Combining convolution operations with a Transformer on the encoder side, Transclaw UNet [20] enables detailed segmentation and long-distance relationship learning. UNETR [21] adopts the sequence-to-sequence prediction for 3D medical image segmentation. These developments underscore the transformative impact of Transformer-based approaches on medical image segmentation, charting a path toward broader adoption and deep learning advancement.

2.4 Benchmarks in Medical Image Segmentation

Several benchmarks have been widely used to evaluate medical image segmentation methods:

1. The Synapse multi-organ segmentation dataset provides abdominal CT images for evaluating the segmentation of multiple organs, such as the liver, pancreas, and kidneys [4–6, 9, 15, 18, 19, 22].
2. The ACDC cardiac MRI dataset focuses on the segmentation of cardiac structures, including the left ventricle, right ventricle, and myocardium [4, 6, 15, 19, 22].
3. Polyp segmentation datasets, such as CVC-ClinicDB and Kvasir-SEG, are used to evaluate the detection and segmentation of polyps in colonoscopy images [2, 19, 23].

These benchmarks have been instrumental in driving advancements in medical image segmentation, providing standardized datasets for comparing different methods.

3 GC-UNet Architecture

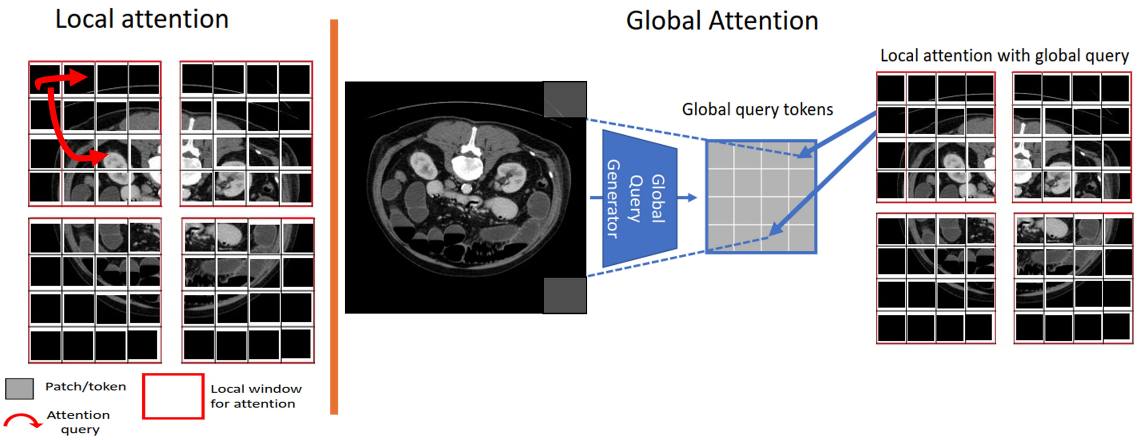

Fig. 1 An illustration of the local and global attention mechanisms in GC-ViT [7]. Local attention is computed on feature patches within local window only (left). The global attention mechanism extracts query patches from the entire input feature map, aggregating information from all windows. The global query is interacted with local key and value tokens, hence allowing to capture long-range information.

The core component of GC-UNet is GC-ViT block, the local and global attention mechanisms of which are illustrated in Figure 1. GC-ViT [7] is a hierarchical architecture like Swin Transformer but utilizes global-window attention instead of shifted-window attention for effectively capturing long-range information. GC-ViT also uses convolution layers for downsampling to provide the network with desirable properties such as locality bias and cross-channel interactions which are missing in both ViT and Swin Transformer. GC-ViT has 4 stages, each of which consists of alternating blocks of local and global Multi-head Self-Attention (MSA) layers. As shown in Figure 1, at each stage, global query tokens are computed by using novel fused inverted residual blocks that encompass global contextual information from different image regions. While the local self-attention modules are responsible for modeling short-range information, the global query tokens are shared across all global self-attention modules to interact with local key and value representations.

As shown in Figure 2, Each GC-ViT block includes a local and global Multi-head Self-Attention (MSA), Multilayer Perceptron (MLP), a Global Token Generator (GTG) and a downsampling layer. The GTG component adds global context to the computations. Local MSA can only query patches within a local window, while global MSA can query different image regions while still operating within the window. At each stage, the global query component is pre-computed. The block also introduces a CNN-based module in the downsampling layer to include inductive bias, a useful feature for images that have been missing in both ViT and Swin Transformer.

GC-UNet is a GC-ViT-based U-shaped Encoder-Decoder architecture with skip-connections for long and short-range semantic feature learning. As shown in Figure 3, the GC-UNet consists of encoder, bottleneck, decoder, and skip connections.

1. Both encoder and decoder used GC-ViT [7] to model long and short-range spatial interactions, without the need for expensive operations such as computing attention masks or shifting local windows.

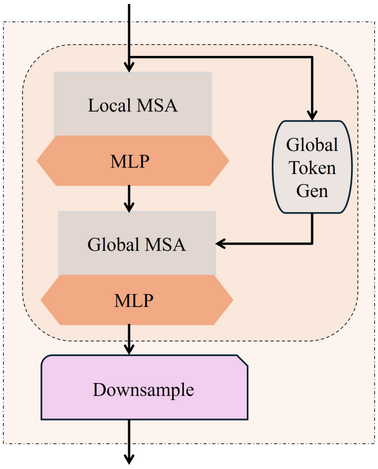

Fig. 2 A GC-ViT block has a local and global attention, a global token generator, and a downsampling layer.

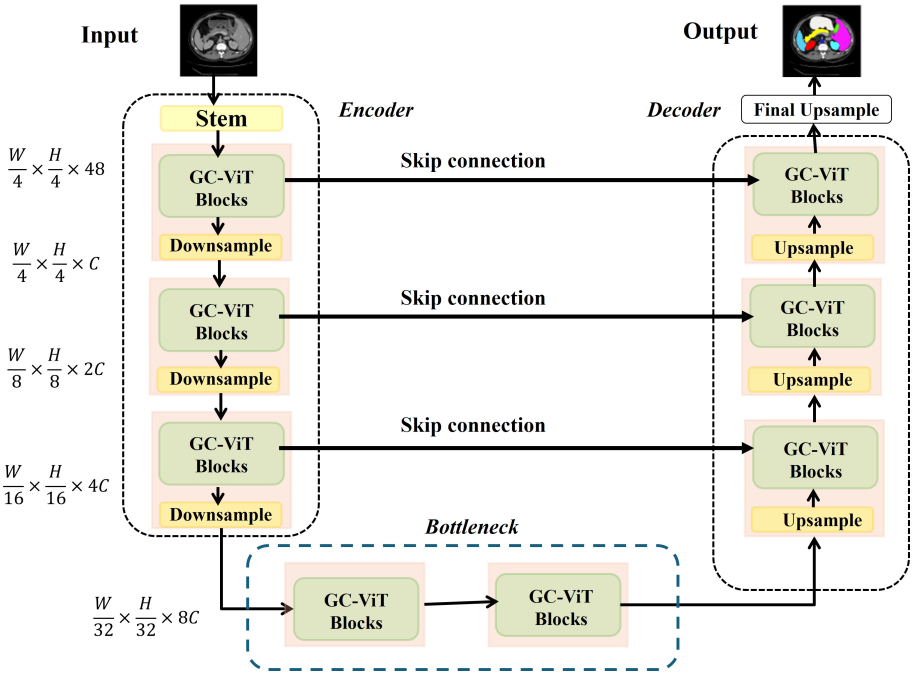
Input
Output
Encoder
Decoder
Final Upsample
Stem
$\frac{W}{4} \times \frac{H}{4} \times 48$
GC-ViT Blocks
Skip connection
Downsample
$\frac{W}{4} \times \frac{H}{4} \times C$
Upsample
$\frac{W}{8} \times \frac{H}{8} \times 2C$
$\frac{W}{16} \times \frac{H}{16} \times 4C$
Bottleneck
$\frac{W}{32} \times \frac{H}{32} \times 8C$

Fig. 3 GC-UNet architecture includes encoders, bottlenecks, skip connections, and decoder. Encoder, bottleneck and decoder are all constructed based on GC-ViT block

2. At each stage, the GC-ViT encoder and decoder consist of alternating local and global self-attention modules to extract spatial features. Both operate in local windows like Swin Transformer.

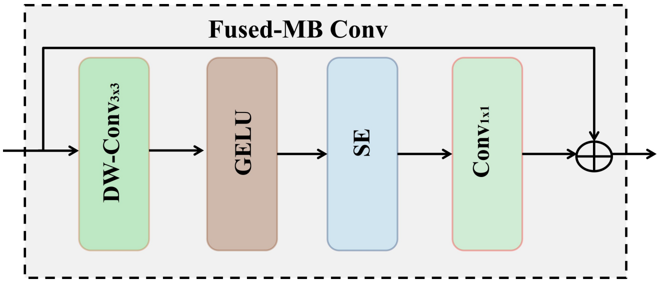

Fig. 4 Fused-MBConv module

3. Skip connections concatenate the feature maps from the GC-ViT encoder with the corresponding decoder stages. The bottleneck is employed to acquire the deep feature representation, maintaining both feature dimension and resolution unchanged within this component.
4. The downsampler between stages in the encoder part and upsampler between stages in the decoder part provide desirable properties such as inductive bias and modeling of inter-channel dependencies.

Within the encoder, the initial image is partitioned into four patch blocks, which act as input for the four-stage GC-VIT module. Following the encoding process, the image dimensions are decreased to (H/32) × (W/32). In the decoder, the upsample operations are utilized to increase the image dimensions by 2 and reduce the number of channels by 2. The features from each stage of the encoder are concatenated with their corresponding stage in the decoder using skip connections. The decoder accomplishes its task through upsampling.

3.1 Encoder

The encoder utilizes a hierarchical GC-ViT approach to acquire feature representations at various resolutions. This is achieved by reducing spatial dimensions while simultaneously increasing embedding dimensions by a factor of 2 across 4 stages. Initially, the input image $x \in \mathbb{R}^{H \times W \times 3}$ undergoes processing through the patchify layer. This layer comprises a 3×3 convolution operation with a stride of 2, along with padding, to generate overlapping patches. Subsequently, these patches are projected into an embedding space of dimension C via another 3×3 convolution layer. After each stage in the GC-ViT backbone, the spatial resolution is decreased while the number of channels is increased through a downsampling layer. This downsampling operation helps in extracting hierarchical features at different resolutions.

3.2 Downsampler

The downsampler incorporates the fused-MBConv module to generates hierarchical representations by injecting inductive bias into the network and modeling inter-channel correlations, where a convolution layer with a kernel size of 3 and a stride of 2 is used to downsample the spatial feature resolution by 2 while doubling the number of channels. The fused-MBConv module, as shown in Figure 4, includes DW-Conv$_{3\times3}$, GELU, SE, and Conv$_{1\times1}$. The fused-MBConv operation can be defined by the following equations:

$$\begin{aligned}\hat{x} &= \text{DW-Conv}_{3\times3}(x),\\ \hat{x} &= \text{GELU}(\hat{x}),\\ \hat{x} &= \text{SE}(\hat{x}),\end{aligned}$$

$$x = \mathrm{Conv}_{1\times 1}(\hat{x}) + x.$$

where DW-Conv refers to depthwise convolution, SE refers to the Squeeze and Excitation block, and GELU represents the Gaussian Error Linear Unit function.

3.3 Bottleneck

Similar to Swin-UNet [4], two GC-ViT blocks are used for bottleneck construction. The bottleneck is strategically designed to facilitate the learning of deep feature representations. Within this structure, the feature dimension and resolution remain unchanged.

3.4 Decoder

The symmetric decoder, corresponding to the encoder, is constructed using the GC-ViT Transformer block. The decoder mirrors the encoder design, replacing the patchy block with an unpatched block, the embedding layer with a de-embedding layer, and the downsample block with an upsample block. The decoder's upsample block replaces the encoder's downsample block. The upsample block restructures the feature map of adjacent dimensions into a higher-resolution feature map and reduces the feature dimension by half. This upsample block effectively increases the spatial resolution while refining and normalizing feature representations, making it suitable for decoding and reconstructing higher-resolution features in segmentation models. The skip connection fuses the features of the encoder with the deep features recovered from the up-sample, therefore mitigating the loss of spatial data produced by the downsampling.

3.5 Hi-GC-ViT Architecture

GC-ViT is lightweight since it does not include expensive convolution operation though this may impact the segmentation accuracy of small objects. To remedy this deficiency, we evaluated the effect of adding depthwise convolution [24] to the algorithm. Depthwise convolution is much more efficient than traditional convolution since it applies separate convolution to each channel. To this end, we developed an architecture, Hi-GC-UNet, whose encoders and decoders include the depthwise convolution blocks running in parallel to the GC-ViT blocks. These components are designed to extract both local and global features from input images, enhancing the model's ability to capture fine-grained details and long-range dependencies.

The convolution block, as shown in the left side of Figure 5 is designed to extract local features from input images. The block utilizes a series of convolution operations followed by batch normalization to refine the feature maps. Specifically, it consists of:

- **Depthwise convolution**: A depthwise convolution is performed on the input feature maps using a 3×3 kernel with a stride of 1. This operation is followed by a ReLU activation function to introduce non-linearity.
- **Pointwise convolution**: A 1×1 pointwise convolution is then applied to combine the output of the depthwise convolution across the channel dimension.
- **Normalization**: The resulting feature map is normalized using batch normalization to stabilize the learning process.
- **Residual connection**: A residual connection is added between the input and the normalized output, enabling better gradient flow during training.

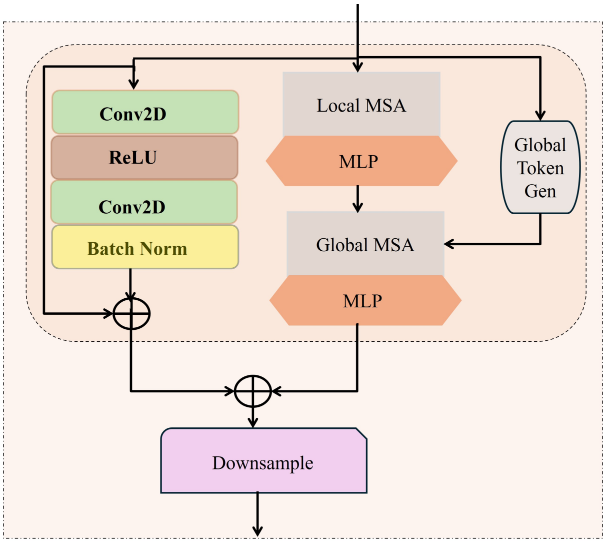

Fig. 5 A Hi-GC-ViT block combines depthwise convolution operations with a GC-ViT block.

The output of the convolution block is a refined feature map that captures local patterns in the input image. The combination of the convolution block and GC-ViT block allows for a comprehensive feature extraction process, effectively capturing both local details and global relationships in the input images.

3.6 ECA-GC-ViT Architecture

The Efficient Channel Attention (ECA) [25] block, a refined version of the Squeeze-and-Excitation (SE) block, enhances channel attention mechanisms in convolutional neural networks. Unlike the SE block's indirect approach, the ECA block directly models interactions between each channel and its K-nearest neighbors. This, coupled with an adaptive kernel size determined based on the number of channels, leads to a more efficient and effective block. In our ECA-GC-UNet architecture, we replace the SE block in the GC-ViT with the ECA block to reduce computational complexity and achieve improved performance. As shown in Figure 6, we also replace the SE block with the ECA block in the Fused-MBConv module to further enhance the model's efficiency and effectiveness.

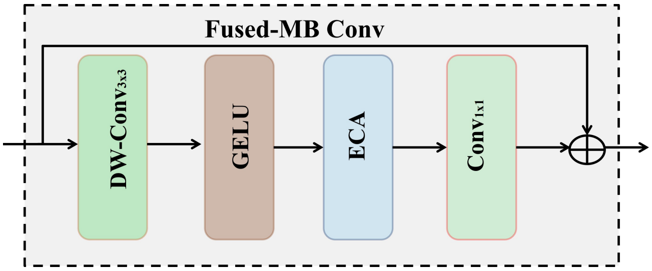

Fig. 6 Modified Fused-MBConv module

4 Pre-training on MedNet dataset

The majority of CNN and Transformer-based segmentation models are pre-trained on natural images such as ImageNet. However, this is suboptimal for medical image segmentation due to the semantic gap between natural and medical image modalities [19, 26]. In this work, we pre-trained the GC-ViT [7] model, specifically GCVit xxTiny, on a large medical image dataset called *MedNet* that contains more than 200,000 medical images collected from several public datasets [27] and Kaggle [28–30].

MedNet consists of different types of microscopy images such as X-ray, computed tomography (CT), optical coherence tomography (OCT), and MRI. Images in MedNet are divided into 65 classes. Similar to the approach of Stuckner *et al.* [31] and Alrfou *et al.* [19], the MedNet dataset is divided into training and validation sets, with each class having 100 images in the validation set, resulting in 96.75%/3.25% training/validation split. Using 100 images per class for validation is sufficient to obtain reliable accuracy metrics and to prevent overfitting during training. Although the validation sets are balanced, the training sets exhibited some class imbalance. There are a few classes, each of which contains less than 0.12% of the total images. Three classes contain 6.2% of the images. Most classes have over 2000 images representing one to two percent of the training set. MedNet includes images from various modalities such as X-ray, CT, OCT, and MRI, and encompassed a wide range of medical diseases such as Kidney Cancer, Cervical Cancer, Alzheimer's, Covid-19, Pneumonia, Tuberculosis, Monkeypox, Breast Cancer, and Malaria.

We trained and tested GC-ViT xxTiny with the AdamW optimizer [32] for 100 epochs with an initial learning rate of 0.0001, weight decay of 0.05, and cosine decay scheduler. The training data had been augmented using the albumentations library, which included random changes to the contrast and brightness, vertical and horizontal flips, photometric distortions, and added noise.

The training process continued until the validation score showed no improvement, employing an early stopping criterion with a patience of 10 epochs. Performance was evaluated using top-1 and top-5 accuracy metrics. Top-1 accuracy measures the percentage of test samples for which the correct label is the top prediction, while top-5 accuracy measures the percentage of test samples for which the correct label appears within the top five predictions. The top-1 accuracy of the GC-ViT xx-Tiny model is 82.3%, and the top-5 accuracy is 98.2%.

5 Experimental Evaluation

5.1 Dataset

To evaluate the effectiveness of our proposed method, we utilized multiple medical image datasets, including the Synapse multi-organ abdominal CT dataset, the ACDC cardiac MRI dataset, and several Polyp segmentation datasets. These datasets were chosen to assess the performance of GC-UNet across a variety of medical imaging tasks, including organ segmentation, cardiac structure segmentation, and polyp detection.

Synapse Synapse multi-organ segmentation dataset (Synapse) includes 30 patient cases with 3779 axial abdominal clinical CT images, where 18 cases are used for training and 12 cases are used for testing. The dataset contains 8 abdominal organs (aorta, gallbladder, left kidney, right kidney, liver, pancreas, spleen, and stomach). Each CT volume includes 85 $\sim$ 198 slices of 512 $\times$ 512 pixel images, with a voxel spatial resolution of $[0.54 \sim 0.54] \times [0.98 \sim 0.98] \times [2.5 \sim 5.0]\ mm^3$.

ACDC Automated Cardiac Diagnosis Challenge dataset (ACDC) [33] compiles MRI scan results of various patients from the MICCAI 2017 dataset. The ACDC dataset contains 100 cardiac MRI scans, each containing three organs: the right ventricle (RV), the myocardium (Myo), and the left ventricle (LV). Following

TransUNet [6], we partitioned the dataset into 70 training cases, 10 validation cases, and 20 test cases.

Polyp We used 5 polyp datasets with early colorectal cancer diagnosis images. The CVC-ClinicDB [34] and Kvasir-SEG [35] datasets are used for binary segmentation. The CVC-ClinicDB dataset contains 612 RGB colonoscopy images with labeled olyps from MICCAI 2015 with a pixel resolution of 288 × 384. The Kvasir-SEG dataset contains 1000 polyp images with a pixel resolution ranging from 332 × 487 to 1920 × 1072 and their corresponding ground truth. Following the setting in PraNet [23], we used 900 images from the CVC-ClinicDB dataset and 548 images from the Kvasir dataset for training. The remaining 64 images from CVC-ClinicDB and 100 images from Kvasir were used as test sets. To evaluate the generalization performance, we tested the model on three unseen datasets: CVC-300 [36], CVC-ColonDB, and ETIS-LaribDB.

Evaluation Method To evaluate the performance of our segmentation model, we used the Dice Similarity Coefficient (DSC) and the 95th percentile Hausdorff Distance (HD95).

1. The Dice Similarity Coefficient (DSC) is a metric used to evaluate the similarity between the predicted segmentation and the ground truth segmentation in medical image analysis. It is defined as:

$$\mathrm{DSC} = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

   where:

   TP (True Positive) is the number of pixels that are correctly classified as the region of interest in the predicted segmentation and are also present in the ground truth segmentation.
   FP (False Positive) is the number of pixels that are incorrectly classified as the region of interest in the predicted segmentation but are not present in the ground truth segmentation.
   FN (False Negative) is the number of pixels that are part of the region of interest in the ground truth segmentation but are not classified as such in the predicted segmentation.

2. The HD95 is defined as:

$$\mathrm{HD95}(A, B) = \mathrm{Percentile}\left(D_{A \to B} \cup D_{B \to A}, 95\right)$$

   where:

   A is the set of points in the predicted segmentation,
   B is the set of points in the ground truth segmentation,
   $D_{A \to B} = \{\min_{b \in B} d(a, b) \mid a \in A\}$ is the set of distances from each point in A to the nearest point in B,
   $D_{B \to A} = \{\min_{a \in A} d(b, a) \mid b \in B\}$ is the set of distances from each point in B to the nearest point in A,
   $d(a, b)$ is the Euclidean distance between points a and b,
   $\mathit{Percentile}(D, 95)$ is the 95th percentile of the combined set of distances $D = D_{A \to B} \cup D_{B \to A}$.

   The HD95 metric is useful for evaluating boundary accuracy in medical image segmentation, as it reduces sensitivity to outliers by focusing on the 95th percentile of distances.

5.2 Implementation

The GC-UNet is implemented with PyTorch library and the model is trained on an Nvidia GeForce GTX TITAN X with 12 GB of memory. The input image size was reduced to 224 × 224 pixels for consistency, and data augmentation techniques (e.g., random cropping, flipping, and rotation) were applied during training to improve model robustness. Our model is trained with batch size 24, learning rate 0.0001, and AdamW optimizer with momentum 0.9 and weight decay 0.0001. The loss function was a weighted combination of Dice loss (0.3) and cross-entropy loss (0.7). The model was trained for 150 epochs on the Synapse and ACDC datasets, while the Polyp dataset was trained for 100 epochs.

5.3 Results

We compared GC-UNet with several state-of-the-art methods, including CNN-based, Transformer-based, and hybrid methods. Table 1 summarizes the list of methods that used in our experiments.

Table 1 List of the related networks included in our experiments.

Architecture	**Method**
CNN	U-Net [1] Att-UNet [3] R50-UNet [6] R50-AttUNet [6] UNet++ [2] PraNet [23]
Transformer	Swin-UNet [4] TransDeepLab [5] MISSForme [15]
Hybrid CNN and Transformer	TransUNet [6] HiFormer [18] R50-ViT [17] CS-UNet [19] GPA-TUNet [22]

5.3.1 Synapse

Performance of GC-UNet and Its Variants

We evaluated the performance of GC-UNet and its variants on the Synapse multi-organ abdominal CT dataset, which includes 8 abdominal organs (aorta, gallbladder, left kidney, right kidney, liver, pancreas, spleen, and stomach). The results are summarized in Table 2.

1. GC-UNet: GC-UNet achieved an DSC of 82.39% and an average HD of 15.94 mm, outperforming most state-of-the-art methods. Specifically, GC-UNet performed exceptionally well on larger organs with clear boundaries, such as the liver with DSC 94.64% and spleen with DSC 91.81%. However, it struggled slightly with smaller organs like the gallbladder with DSC 69.32%, likely due to the difficulty in capturing fine-grained details.
2. Hi-GC-UNet: The Hi-GC-UNet variant, which integrates depthwise convolution to enhance local feature extraction, achieved an average DSC of 82.28% and an HD of 17.44 mm. While its overall performance was slightly lower than GC-UNet, Hi-GC-UNet showed improvements in specific organs, such as the gallbladder with DSC 70.79%, indicating that depthwise convolution is effective for detecting smaller structures.
3. ECA-GC-UNet: The ECA-GC-UNet variant, which replaces the SE block with the ECA block, achieved the best performance among the variants, with an average DSC of 82.54% and an HD of 16.50 mm. ECA-GC-UNet performed particularly well on the right kidney with DSC 84.33% and spleen with DSC 91.79%, demonstrating that the ECA block improves feature representation while reducing model complexity.

Comparison with State-of-the-Art Methods

We compare GC-UNet with current state-of-the-art methods on the Synapse dataset, including U-Net, Att-UNet, TransUnet, SwinUnet, MISSFormer, TransDeepLab, HiFormer, GPA-TUNet, and CS-UNet. The results are summarized in Table 2.

GC-UNet with DSC 82.39% and HD 15.94 mm performs competitively with the best hybrid method, CS-UNet with DSC 83.27 and HD 15.26 mm, while maintaining a smaller size and lower computational cost. This demonstrates that GC-UNet archives a better trade-off between performance and efficiency, making it more suitable for real-world clinical applications. Also, note that GC-UNet significantly outperforms CNN-based methods UNet and Att-UNet. For instance, UNet has an average DSC of 76.85% and an average HD of 39.70 mm and Att-UNet has an average DSC of 77.77% and an average HD of 36.02 mm. GC-UNet has better performance than Transformer-based methods like Transdeeplab, MISSFormer, and Swin-UNet, which has an average DSC ranging from 79.13 to 81.96% and an average HD ranging from 18.20 to 21.25 mm.

Table 2 Comparison of GC-UNet/Hi-GC-UNet and state-of-the-art algorithms on Synapse (the columns are average DSC in %, average HD in mm, and DSC in % for each organ). Blue indicates the best result and red displays the second-best. The superscript 1 and 2 indicate pre-training on ImageNet and MedNet respectively. Other models are pre-trained on ImageNet.

Algorithm	DSC↑	HD↓	Aorta	Gallbladder	Kid(L)	Kid(R)	Liver	Pancreas	Spleen	Stomach
U-Net [1]	76.85	39.70	**89.07**	69.72	77.77	68.60	93.43	53.98	86.67	75.58
Att-UNet [3]	77.77	36.02	**89.55**	68.88	77.98	71.11	93.57	58.04	87.30	75.75
Swin-UNet [4]	79.13	21.55	85.47	66.53	83.2	79.61	94.29	56.58	90.66	76.60
TransDeepLab [5]	80.16	21.25	86.04	69.16	84.08	79.88	93.53	61.19	89.00	78.40
MISSFormer [15]	81.96	18.20	86.99	68.65	85.21	82.00	94.41	**65.67**	**91.92**	80.81
TransUNet [6]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
GPA-TUNet [22]	80.37	20.55	88.74	65.63	83.51	80.37	94.84	63.89	87.58	78.40
HiFormer [18]	80.39	**14.70**	86.21	65.69	85.23	79.77	94.61	59.52	90.99	81.08
CS-UNet [19]	**83.27**	**15.26**	88.07	**71.32**	**88.00**	**84.38**	94.80	65.64	89.95	**83.81**
GC-UNet^1	81.95	16.80	86.96	66.26	**87.75**	83.86	94.53	61.06	91.42	83.74
GC-UNet^2	82.39	15.94	86.30	69.32	86.11	81.89	94.64	64.88	91.81	**84.15**
Hi-GC-UNet^1	81.76	18.77	85.86	69.76	82.82	79.39	**95.08**	**65.77**	**91.93**	83.48
Hi-GC-UNet^2	82.28	17.44	85.67	**70.79**	85.43	82.68	**95.10**	64.31	91.18	83.08
ECA-GC-UNet^1	81.91	18.17	86.60	70.16	86.29	83.33	94.18	62.76	90.24	81.69
ECA-GC-UNet^2	**82.54**	16.50	86.56	69.81	87.28	**84.33**	94.09	64.65	91.79	81.88

Organ-wise performance

Figure 7 shows qualitative comparison on two Synapse images, where GC-UNet has superior performance than Swin-UNet and CS-UNet. GC-UNet segmented most of the organs correctly, with a few misclassifications in the gallbladder area. In comparison, Swin-UNet over-segmented the spleen (some areas belonging to the spleen were misclassified as the left kidney) and CS-UNet over-segmented the pancreas. We speculate that this improvement is due to the use of GC-ViT, which introduces a parameter-efficient downsampling module with modified Fused MB-Conv blocks. These modifications address the lack of inductive bias in ViTs, enabling GC-UNet to accurately capture relatively large regions and perform well with organs close to each other.

The segmentation of larger organs with clear boundaries such as kidney, pancreas, and spleen require the network to capture global features. We speculate that this is why Transformer-based models are more accurate compared to CNN-based models. The segmentation of smaller organ like aorta benefits more from the detection of local features. This is probably why CNN-based models have more accurate results than Transformed-based models. The segmentation of larger organs with complex boundaries, such as liver and stomach, requires to capture local and global features. That is probably why hybrid models have more accurate result.

Impact of Pre-training

The average performance of GC-UNet and its variants are consistently better when they are pretrained on MedNet. Also, ECA-GC-UNet^2 has slightly better performance than GC-UNet^2 in average DSC. This highlights the importance of domain-specific pre-training for medical image segmentation.

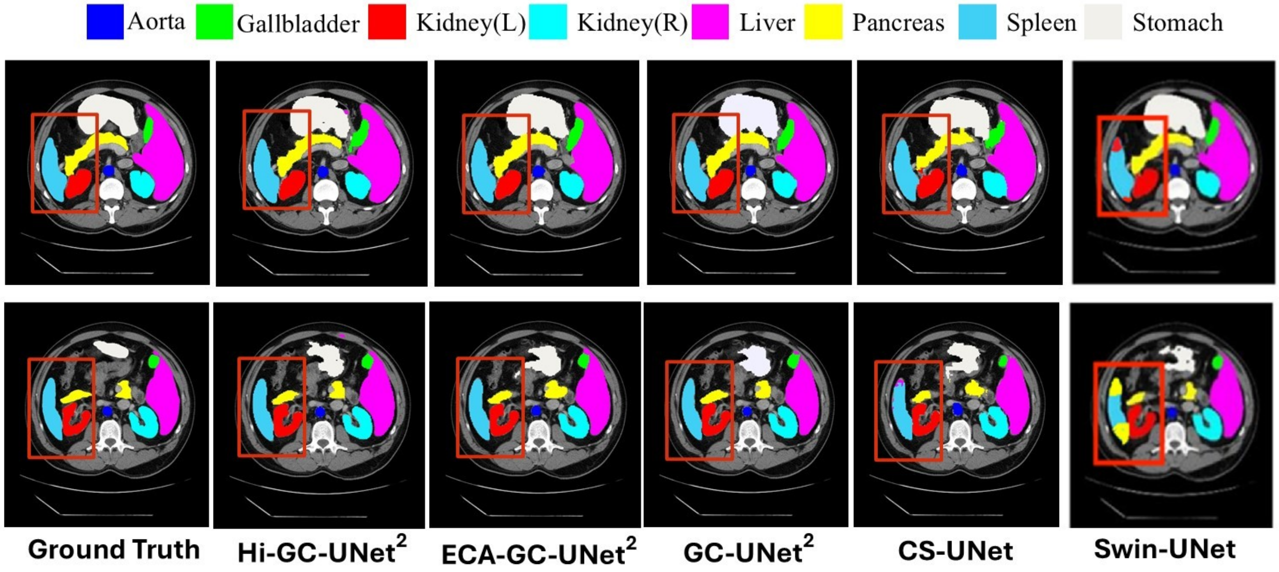

Fig. 7 Comparison of GC-UNet and its variants with the ground truth, CS-UNet, and Swin-UNet on two sample images in Synapse dataset. The superscript 2 indicates pretraining on MedNet. The red rectangles identify the regions where Swin-UNet/CS-UNet tend to have over/under-segmentation problems compared to the rest.

5.3.2 ACDC

Performance of GC-UNet and Its Variants

We evaluated the performance of GC-UNet and its variants on Automated Cardiac Diagnosis Challenge dataset (ACDC), which includes 3 organs: the right ventricle (RV), the myocardium (Myo), and the left ventricle (LV). The results are summarized in Table 3.

1. GC-UNet achieves an average DSC of 91.23% on the ACDC dataset, with performance in segmenting LV with DSC of 96.57% and RV with DSC of 89.88%.

2. Hi-GC-UNet, which integrates depthwise convolution, further improves performance, achieving an average DSC of 91.76%, with the best results for the LV with a DSC of 96.58% and RV DSC of 90.80%.
3. ECA-GC-UNet, which replaces the SE block with the ECA block, Achieved comparable performance to Hi-GC-UNet, with an average DSC of 91.73%, with the best results for the LV with a DSC 96.86%.

These results demonstrate that both Hi-GC-UNet and ECA-GC-UNet provide incremental improvements over the base GC-UNet model, particularly in segmenting the LV and RV.

Table 3 Comparison of GC-UNet/Hi-GC-UNet with the state-of-the-art methods on ACDC dataset in DSC. Blue denotes the best results and red denotes the second best. The superscript 1 and 2 indicate pre-training on ImageNet and MedNet respectively. Other models are pre-trained on ImageNet.

Algorithm	DSC(%)↑	Right Ventricle	Myocardium	Left Ventricle
R50-UNet [6]	87.55	87.10	80.63	94.92
R50-Atten-UNet [6]	86.75	87.58	79.20	93.47
R50-ViT [6]	87.57	86.07	81.88	94.75
Swin-UNet [4]	90.00	88.55	85.62	95.83
MISSFormer [15]	90.86	89.55	**88.04**	94.99
TransUNet [6]	89.71	88.86	84.53	95.73
GPA-TUNet [22]	90.37	89.44	**87.98**	93.68
CS-UNet [19]	90.38	88.28	86.50	96.35
GC-UNet^1	90.98	89.63	86.77	96.55
GC-UNet^2	91.23	89.88	87.25	96.57
Hi-GC-UNet^1	**91.48**	**90.56**	87.37	96.51
Hi-GC-UNet^2	**91.76**	**90.80**	87.90	**96.58**
ECA-GC-UNet^1	91.43	90.16	87.48	96.65
ECA-GC-UNet^2	91.73	90.45	87.90	**96.86**

Comparison with State-of-the-Art Methods

We compare the performance of GC-UNet and its variants on the ACDC dataset with some state-of-the-art methods including CNN-based methods (R50-UNet and R50-Atten-UNet), Transformer-based methods (R50-ViT, Swin-UNet, MISSFormer), and hybrid methods (TransUNet, GPA-TUNet, and CS-UNet). The results are shown in Table 3. GC-UNet and its variants are better than all other methods in terms of average DSC, where Hi-GC-UNet pre-trained on MedNet has the best performance (91.76%). This represents a significant improvement over previous approaches, with DSC scores below 90.86% for all other networks. Compared to CS-UNet, our model achieves a 0.85% higher DSC, demonstrating the effectiveness of the global context self-attention and local self-attention mechanisms in capturing both long and short-range spatial dependencies.

Figure 8 includes 3 example images in the ACDC dataset for a qualitative comparison between GC-UNet, its variants, CS-UNet and Swin-UNet. GC-UNet and its variants (pretrained on MedNet) are able to segment right ventricle and left ventricle more accurately than CS-UNet (pretrained on ImageNet).

Impact of Pre-training

We evaluated the impact of pre-training GC-UNet on two datasets: ImageNet and MedNet. When pre-trained on ImageNet, GC-UNet achieved an average DSC of 90.98%, Hi-GC-UNet reached 91.48%, and ECA-GC-UNet obtained 91.43%. Pre-training on MedNet led to further improvements, with GC-UNet achieving an average DSC of 91.23%, Hi-GC-UNet improving to 91.76%, and ECA-GC-UNet reaching 91.73%. These

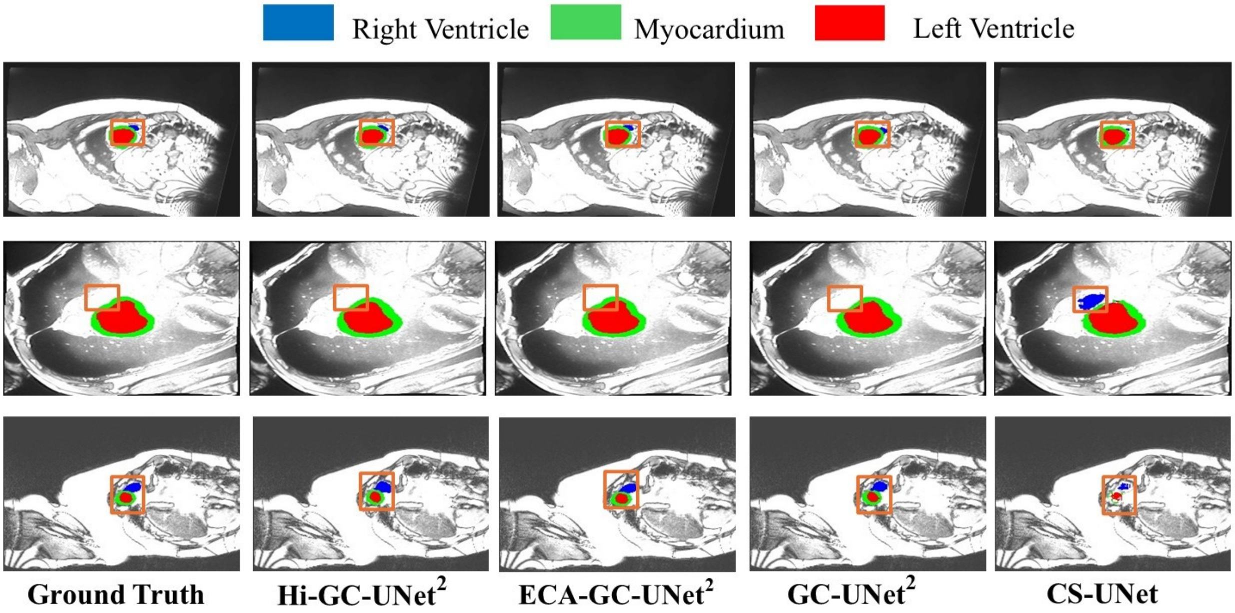

Fig. 8 Comparison of GC-UNet and its variants with the ground truth and CS-UNet on 3 example images in ACDC dataset. The superscript 2 indicates pretraining on MedNet. The orange rectangle box identifies the regions where CS-UNet have over or under segmentation problems compared to the rest.

results demonstrate that pre-training on MedNet, a domain-specific dataset, consistently enhances the performance of GC-UNet and its variants, highlighting the importance of domain-specific pre-training for medical image segmentation tasks.

5.3.3 Polyp datasets

We evaluated the performance of GC-UNet and its variants on several Polyp datasets by first training it on 2 seen datasets (CVC-ClinicDB and Kvasir) and then use the trained models on 3 unseen datasets (CVC-ColonDB, ETIS-LaribDB, and CVC-300) to evaluate the generalizability of the models.

Table 4 compares the performance of GC-UNet and its variants with state-of-the-art CNN-based algorithms (UNet, UNet++, PraNet) and hybrid method (CS-UNet). While GC-UNet pre-trained on MedNet has the best DSC metric for the Kvasir dataset, it is behind CS-UNet on the CVC-ClinicDB dataset though the difference is relatively small, with DSC scores of 90.60% compared to 90.67%.

Generalizability

As shown in Table 4 (column 3–5), GC-UNet is more generalizable to unseen datasets (CVC-ColonDB, ETIS-LaribDB, and CVC-300). Compared to other approaches, GC-UNet models have the best and second best DSC value in CVC-ColonDB and ETIS-LaribDB datasets and the second best DSC value in CVC-300 dataset. For example, ECA-GC-UNet pre-trained on MedNet achieved a DSC of 77.84% on CVC-ColonDB and 74.26% on ETIS-LaribDB, outperforming CS-UNet (DSC: 72.00% and 64.50%) and PraNet (DSC: 70.9% and 62.8%). The overall performance of GC-UNet is rather remarkable.

Table 4 Comparison of GC-UNet/Hi-GC-UNet and State-Of-The-Art algorithms on Pylop datasets (the columns are average DSC in %). Blue indicates the best result and red displays the second-best. The superscript 1 and 2 indicate pre-training on ImageNet and MedNet respectively. Other models are pre-trained on ImageNet.

UNet[1]	82.3	81.8	71.0	51.2	39.8
UNet++ [2]	79.4	82.1	70.7	48.3	40.1
PraNet [23]	89.9	89.8	**87.1**	70.9	62.8
CS-UNet [19]	**90.67**	90.00	85.59	72.00	64.50
GC-UNet[1]	89.48	89.26	86.52	74.95	65.97
GC-UNet[2]	89.44	90.02	86.20	**78.08**	72.03
Hi-GC-UNet[1]	89.14	89.70	85.18	75.90	72.76
Hi-GC-UNet[2]	88.87	**90.58**	86.42	76.64	**73.39**
ECA-GC-UNet[1]	90.20	**90.43**	85.33	78.67	71.68
ECA-GC-UNet[2]	**90.60**	90.27	**86.54**	**77.84**	**74.26**

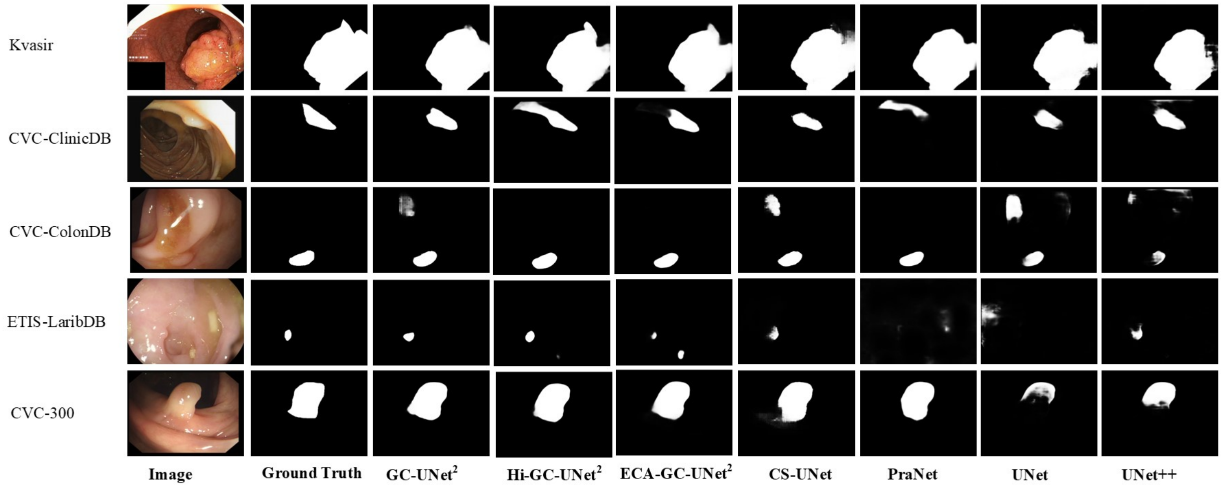

Fig. 9 Comparison of GC-UNet/Hi-GC-UNet with CS-UNet, PraNet, UNet, and UNet++ on 5 example images in the Polyp datasets. The superscript 2 indicates pretraining on MedNet.

Figure 9 presents the qualitative segmentation results of various methods, including GC-UNet that trained on MedNet and ImageNet. Five samples, one from each dataset, are selected to highlight ambiguous boundaries and small polyps, facilitating a differentiated comparison of segmentation performance. GC-UNet, pre-trained on MedNet, shows a significant reduction in false positives and false negatives. This improvement is attributed to its enhanced ability to distinguish between the obscure boundaries of polyp regions and normal regions.

Impact of Pre-training

The impact of pre-training on MedNet versus ImageNet was also evaluated. GC-UNet and its variants consistently performed better when pre-trained on MedNet compared to ImageNet. For instance, GC-UNet pre-trained on MedNet achieved a DSC of 90.02% on Kvasir, compared to 89.26% when pre-trained on ImageNet. Similarly, ECA-GC-UNet pre-trained in MedNet achieved a DSC of 90.60% in CVC-ClinicDB, compared to 90.20% when pre-trained in ImageNet. This improvement further confirms the benefit of domain-specific pre-training for medical image segmentation.

6 Computation Complexity

GC-UNet not only has good segmentation performance but also lower computation complexity than the state-of-the-art segmentation methods. We compared the computation complexity of GC-UNet and its variants with that of Transformer-based and hybrid methods in terms of model parameter numbers, floating point operations (FLOPs) per epoch of training, inference time, and model sizes. This assessment is evaluated on the Synapse dataset. As shown in Table 5, ECA-GC-UNet has the least number of parameters at 12.12 million, while the smallest Transformer-based model, TransDeepLab, has 21.14 million parameters. Correspondingly, ECA-GC-UNet has the smallest model size at 48.92 MB while TransDeepLab has 86.343 MB. ECA-GC-UNet also has the least number of FLOPs per epoch of training at 30.28G while the closest Transformer-based model, Swim-UNet, has 61.64G. Similarly, ECA-GC-UNet has the least training time per epoch and highest FPS for inference.

The low computation cost and small memory footprint of GC-UNet make it the best trade-off between performance and model complexity. It has better or comparable performance than most of the state-of-the-art models, which have substantially larger model size and higher computation cost. The efficient design of GC-UNet highlights its potential for achieving better segmentation results in clinical applications.

Table 5 Comparison of GC-UNet, its variants, Transformer-based networks, and hybrid networks based on model parameters, the floating point operations (FLOPs) per training epoch, number of epochs to train, training time per epoch (TTPE), and inference time in Frame Per Second (FPS) for 1568 axial abdominal clinical CT images, and model size. For FLOPs and TTPE, the batch size is 10. Training epochs is what is needed for the final result.

Algorithm	# of params (M)	FLOPs (G)	# of epochs	TTPE (m:s)	FPS	Model size (MB)
TransDeepLab [5]	21.14	160.00	200	1:36	20	86.343
Swin-UNet [4]	27.17	61.64	150	1:45	27	108.058
MISSFormer [15]	42.46	98.86	400	4:01	19	166.124
HiFormer [18]	25.51	80.45	400	1:27	19	101.161
CS-UNet [19]	44.96	110.00	150	2:45	26	177.613
TransUNet [6]	105.28	290.00	150	2:54	17	414.412
GC-UNet	12.34	30.41	150	**1:16**	30	49.75
Hi-GC-UNet	13.06	31.54	150	1:18	28	52.58
ECA-GC-UNet	**12.12**	**30.28**	150	**1:16**	**30.5**	**48.92**

7 Ablation Study

To investigate the impact of various factors on model performance, we conducted ablation studies using the Synapse dataset. Below, we discuss the effects of upsampling and the optimal hyperparameters for training our model.

7.1 Hyper-parameter Tuning

GC-UNet is trained with a combination of two loss functions, dice loss and cross-entropy loss, which aligns with many current segmentation methods. During the training process, we improve the performance by choosing an optimal combination of dice and cross-entropy losses and an optimal learning rate. We conducted experiments to identify the optimal settings for the combined losses and the learning rate. Table 6 compares the performance of GC-UNet in terms of the Dice-Similarity Coefficient (DSC) and Hausdorff Distance (HD) values for various hyper-parameter values. The optimal DSC and HD are achieved when (dice loss, cross-entropy loss) = (0.3, 0.7) and learning rate is 0.0001. This configuration was used in all of our subsequent experiments.

Table 6 Ablation study on the impact of the training hyper-parameters to the performance of GC-UNet. The hyper parameters include the loss function (cross entropy loss and dice loss) and learning rate.

AdamW	0.6	0.4	0.00001	81.67	22.40
AdamW	0.6	0.4	0.0001	81.58	22.35
AdamW	0.7	0.3	0.00001	81.53	23.16
AdamW	0.7	0.3	0.0001	82.39	15.94

7.2 Upsampling

Similar to Swin-UNet [4], to complement the downsampling layer in the encoder, we specifically designed an upsampling layer in the decoder to perform upsampling and feature dimension increase. To assess the effectiveness of this upsampling layer, we evaluated GC-UNet on the Synapse dataset to compare GC-UNet with bilinear interpolation (with or without SE (Squeeze and Excitation) block) to GC-UNet with transposed convolution (Fused-MBConv module with or without SE block) in the upsampling layer. The results in Table 7 indicate that GC-UNet with transposed convolution (Fused-MBConv module with SE block) in the upsampling layer has the best performance.

Table 7 Ablation study on the impact of the upsampling types to the performance of GC-UNet.

Upsampling Type	DSC (%)	HD (mm)
bilinear interpolation	80.31	22.64
bilinear interpolation + SE	80.89	25.76
transposed convolution (Fused-MBConv)	81.80	21.12
transposed convolution (Fused-MBConv + SE)	82.39	15.94

8 Conclusion

We introduced GC-UNet, a U-shaped network that incorporates a lightweight vision transformer to enhance medical image segmentation by effectively capturing both global and local features. The downsampling and upsampling blocks between encoder and decoder components provide inductive bias and model inter-channel dependencies effectively. GC-UNet and its variants have better or comparable performance than traditional CNN-based, Transformer-based, and hybrid methods on various medical image datasets. At the same time, GC-UNet has lower model complexity with less number of model parameters, lower model size, lower training and inference time, and lower FLOPs for training. The ability of GC-UNet to model long-range spatial dependencies and its competitive performance in segmenting complex and small anatomical structures make it a promising tool for clinical applications. The architecture's design, which includes GC-ViT encoders and decoders with skip connections, contributes to its high performance while maintaining a lower computational complexity compared to state-of-the-art methods. The pretraining on a medical image dataset, MedNet, and the subsequent evaluations on multiple medical imaging tasks show the model's robustness and generalization capabilities, positioning GC-UNet as a practical and powerful approach for medical image segmentation.

In future work, we plan to introduce the GC-UNet 3D model for voxel segmentation of medical images. Another potential direction is to leverage the capsule network (CapsNet) [37], which has better handling of spatial hierarchies, is more robust to transformation, and can generalize better to unseen data. CapsNet has a high computational cost and is not suitable for large images. This was overcome with local constraints on the routing and sharing of the transformation matrix and was shown to be effective for the segmentation of medical images [38]. In future work, we would like to compare the performance of more efficient capsule networks [39, 40] with CNN- and Transformer-based segmentation algorithms.

Acknowledgments

This research was partially funded by the Northwestern Mutual Data Science Institute.

Author Contributions

Khaled Alrfou: Developed the GC-UNet method, conceived and designed the study, developed the software, evaluated results, provided datasets, and contributed to the formal analysis and writing of the original draft. **Tian Zhao**: evaluated the results, contributed to the formal analysis and writing of the original draft, and proofread and reviewed the final manuscript.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4_28 . Springer

[2] Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, pp. 3–11 (2018). Springer

[3] Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., *et al.*: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018) https://doi.org/10.48550/arXiv.1804.03999

[4] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision, pp. 205–218 (2022). https://doi.org/10.1007/978-3-031-25066-8_9 . Springer

[5] Azad, R., Heidari, M., Shariatnia, M., Aghdam, E.K., Karimijafarbigloo, S., Adeli, E., Merhof, D.: Transdeeplab: Convolution-free transformer-based deeplab v3+ for medical image segmentation. In: Predictive Intelligence in Medicine: 5th International Workshop, PRIME 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings, pp. 91–102 (2022). https://doi.org/10.1007/978-3-031-16919-9_9 . Springer

[6] 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 preprint arXiv:2102.04306 (2021) https://doi.org/10.48550/arXiv.2102.04306

[7] Hatamizadeh, A., Yin, H., Heinrich, G., Kautz, J., Molchanov, P.: Global context vision transformers. In: International Conference on Machine Learning, pp. 12633–12646 (2023). PMLR

[8] Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging **39**(6), 1856–1867 (2019)

[9] Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1055–1059 (2020). IEEE

[10] Fu, Z., Li, J., Hua, Z.: Deau-net: Attention networks based on dual encoder for medical image segmentation. Computers in Biology and Medicine **150**, 106197 (2022)

[11] Zhang, W., Lu, F., Zhao, W., Hu, Y., Su, H., Yuan, M.: Accpg-net: A skin lesion segmentation network with adaptive channel-context-aware pyramid attention and global feature fusion. Computers in Biology and Medicine **154**, 106580 (2023)

[12] Li, C., Wang, L., Cheng, S.: Enhanced transformer encoder and hybrid cascaded upsampler for medical image segmentation. Expert Systems with Applications **238**, 121965 (2024)

[13] Xiao, X., Lian, S., Luo, Z., Li, S.: Weighted res-unet for high-quality retina vessel segmentation. In: 2018 9th International Conference on Information Technology in Medicine and Education (ITME), pp. 327–331 (2018). IEEE

[14] 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 IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021). https://doi.org/10.1109/ICCV48922.2021.00986

[15] Huang, X., Deng, Z., Li, D., Yuan, X., Fu, Y.: Missformer: an effective transformer for 2d medical image segmentation. IEEE transactions on medical imaging (2022)

[16] Azad, R., Jia, Y., Aghdam, E.K., Cohen-Adad, J., Merhof, D.: Enhancing medical image segmentation

with transception: a multi-scale feature fusion approach. arXiv preprint arXiv:2301.10847 (2023)

[17] Dosovitskiy, A., Beyer, L., Kolesnikov, 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. arXiv preprint arXiv:2010.11929 (2020) https://doi.org/10.48550/arXiv.2010.11929

[18] Heidari, M., Kazerouni, A., Soltany, M., Azad, R., Aghdam, E.K., Cohen-Adad, J., Merhof, D.: Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 6202–6212 (2023). https://doi.org/WACV56688.2023.00614

[19] Alrfou, K., Zhao, T., Kordijazi, A.: CS-UNet: A generalizable and flexible segmentation algorithm. Multimedia Tools and Applications, 1–28 (2024)

[20] Chang, Y., Menghan, H., Guangtao, Z., Xiao-Ping, Z.: Transclaw u-net: Claw u-net with transformers for medical image segmentation. arXiv preprint arXiv:2107.05188 (2021)

[21] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574–584 (2022)

[22] Li, C., Wang, L., Li, Y.: Gpa-tunet: Transformer and gpa attention co-encoder for medical image segmentation (2022)

[23] 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: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 263–273 (2020). Springer

[24] Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)

[25] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11534–11542 (2020)

[26] Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H.: Transformers in medical imaging: A survey. Medical Image Analysis, 102802 (2023)

[27] Subramoniam, M., Aparna, T., Anurenjan, P., Sreeni, K.: Deep learning-based prediction of alzheimer's disease from magnetic resonance images. In: Intelligent Vision in Healthcare, pp. 145–151. Springer, ??? (2022)

[28] ”Kaggle:CT scan”. https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images (2020)

[29] ”Kaggle:CT KIDNEY DATASET”. https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone/data (2021)

[30] ”Kaggle:Medical Scan Classification Dataset”. https://www.kaggle.com/datasets/arjunbasandrai/medical-scan-classification-dataset (2024)

[31] Stuckner, J., Harder, B., Smith, T.M.: Microstructure segmentation with deep learning encoders pre-trained on a large microscopy dataset. npj Computational Materials **8**(1), 200 (2022)

[32] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[33] Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., *et al.*: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging **37**(11), 2514–2525 (2018)

[34] 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. Computerized medical imaging and graphics **43**, 99–111 (2015)

[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: MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26, pp. 451–462 (2020). Springer

[36] Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, A.C.: A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of Healthcare Engineering (2017)

[37] Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS’17, pp. 3859–3869. Curran Associates Inc., Red Hook, NY, USA (2017)

[38] LaLonde, R., et al.: Capsules for biomedical image segmentation. Medical Image Analysis **68** (2021)

[39] Liu, Y., Zhang, D., Zhang, Q., Han, J.: Part-object relational visual saliency. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(7), 3688–3704 (2022) https://doi.org/10.1109/TPAMI.2021.3053577

[40] Liu, Y., Zhou, L., Wu, G., Xu, S., Han, J.: Tcgnet: Type-correlation guidance for salient object detection. IEEE Transactions on Intelligent Transportation Systems **25**(7), 6633–6644 (2024) https://doi.org/10.1109/TITS.2023.3342811