G-CASCADE: Efficient Cascaded Graph Convolutional Decoding for 2D Medical Image Segmentation

Md Mostafijur Rahman   Radu Marculescu

The University of Texas at Austin

{mostafijur.rahman, radum}@utexas.edu

Abstract

In this paper, we are the first to propose a new graph convolution-based decoder namely, Cascaded Graph Convolutional Attention Decoder (G-CASCADE), for 2D medical image segmentation. G-CASCADE progressively refines multi-stage feature maps generated by hierarchical transformer encoders with an efficient graph convolution block. The encoder utilizes the self-attention mechanism to capture long-range dependencies, while the decoder refines the feature maps preserving long-range information due to the global receptive fields of the graph convolution block. Rigorous evaluations of our decoder with multiple transformer encoders on five medical image segmentation tasks (i.e., Abdomen organs, Cardiac organs, Polyp lesions, Skin lesions, and Retinal vessels) show that our model outperforms other state-of-the-art (SOTA) methods. We also demonstrate that our decoder achieves better DICE scores than the SOTA CASCADE decoder with 80.8% fewer parameters and 82.3% fewer FLOPs. Our decoder can easily be used with other hierarchical encoders for general-purpose semantic and medical image segmentation tasks. The implementation can be found at: https://github.com/SLDGroup/G-CASCADE.

1. Introduction

Automatic medical image segmentation plays a crucial role in the diagnosis, treatment planning, and post-treatment evaluation of various diseases; this involves classifying pixels and generating segmentation maps to identify lesions, tumours, or organs. Convolutional neural networks (CNNs) have been extensively utilized for medical image segmentation tasks [25, 22, 40, 13, 10, 21]. Among them, the U-shaped networks such as UNet [25], UNet++ [40], UNet 3+ [13], and DC-UNet [21] exhibit reasonable performance and produce high-resolution segmentation maps. Additionally, researchers have incorporated attention modules into their architectures [22, 5, 10] to enhance feature maps

and improve pixel-level classification of medical images by capturing salient features. Although these attention-based methods have shown improved performance, they still struggle to capture long-range dependencies [23].

Recently, vision transformers [9] have shown great promise for capturing long-range dependencies among pixels and demonstrated improved performance, particularly for medical image segmentation [3, 2, 8, 31, 23, 24, 39, 29]. The self-attention (SA) mechanism used in transformers learns correlations among input patches; this enables capturing the long-range dependencies among pixels. Recently, hierarchical vision transformers such as the Swin transformer [19], the pyramid vision transformer (PVT) [32], MaxViT [27], MERIT [24], have been introduced to enhance performance. These hierarchical vision transformers are effective in medical image segmentation tasks [3, 2, 8, 31, 23, 24]. As self-attention modules employed in transformers have limited capacity to learn (local) spatial relationships among pixels [6, 15], some methods [36, 34, 33, 8, 31, 23, 24] incorporate local convolutional attention modules in the decoder. However, due to the locality of convolution operations, these methods have difficulties at capturing long-range correlations among pixels.

To overcome the aforementioned limitations, we introduce a new Graph based CAScaded Convolutional Attention DEcoder (G-CASCADE) using graph convolutions. More precisely, G-CASCADE enhances the feature maps by preserving long-range attention due to the global receptive field of the graph convolution operation, while incorporating local attention through the spatial attention mechanism. Our contributions are as follows:

- **New Graph Convolutional Decoder:** We introduce a new graph-based cascaded convolutional attention decoder (G-CASCADE) for 2D medical image segmentation; this takes the multi-stage features of vision transformers and learns multiscale and multiresolution spatial representations. To the best of our knowledge, we are the first to propose this graph convolutional network-based decoder for semantic segmentation.

- **Efficient Graph Convolutional Attention Block:** We introduce a new graph convolutional attention module to build our decoder; this preserves the long-range attention of the vision transformer and highlights salient features by suppressing irrelevant regions. The use of graph convolution makes our decoder efficient.

- **Efficient Design of Up-Convolution Block:** We design an efficient up-convolution block that enables computational gains without degrading performance.

- **Improved Performance:** We empirically show that G-CASCADE can be used with any hierarchical vision encoder (e.g., PVT [33], MERIT [3]), while significantly improving the performance of 2D medical image segmentation. When compared against multiple baselines, G-CASCADE produces better results than SOTA methods on ACDC, Synapse Multi-organ, ISIC2018 skin lesion, Polyp, and Retinal vessels segmentation benchmarks with a significantly lower computational cost.

The remaining of this paper is organized as follows: Section 2 summarizes the related work in vision transformers, graph convolutional networks, and medical image segmentation. Section 3 describes the proposed method. Section 4 explains experimental setup and results on multiple medical image segmentation benchmarks. Section 5 covers different ablation experiments. Lastly, Section 6 concludes the paper.

2. Related Work

We divide the related work into three parts, i.e., vision transformers, vision graph convolutional networks, and medical image segmentation; these are described next.

2.1. Vision transformers

Dosovitskiy et al. [9] pioneered the development of the vision transformer (ViT), which enables the learning of long-range relationships between pixels through self-attention. Subsequent works have focused on enhancing ViT in various ways, such as integration of convolutional neural networks (CNNs) [33, 27], introducing new SA blocks [19, 27], and novel architectural designs [32, 36]. Liu et al. [19] introduce a sliding window attention mechanism within the hierarchical Swin transformer. Xie et al. [36] present SegFormer, a hierarchical transformer utilizing Mix-FFN blocks. Wang et al. [32] develop the pyramid vision transformer (PVT) with a spatial reduction attention mechanism, and subsequently extend it to PVTv2 [33] by incorporating overlapping patch embedding, a linear complexity attention layer, and a convolutional feedforward network. Most recently, Tu et al. [27] introduce MaxViT, which employs a multi-axis self-attention mechanism to construct a hierarchical CNN-transformer encoder.

Although vision transformers exhibit remarkable performance, they have certain limitations in their (local) spatial information processing capabilities. In this paper, we aim to overcome these limitations by introducing a new graph-based cascaded attention decoder that preserves the long-range attention through graph convolution and incorporates local attention by a spatial attention mechanism.

2.2. Vision graph convolutional networks

Graph convolutional networks (GCNs) are developed primarily focusing on point clouds classification [17, 18], scene graph generation [37], and action recognition [38] in computer vision. Vision GNN (ViG) [11] introduces the first graph convolutional backbone network to directly process the image data. ViG devides the image into patches and then uses K-nearest neighbors (KNN) algorithm to connect various patches; this enables the processing of long-range dependencies similar to vision transformers. Besides, due to using 1×1 convolutions before and after the graph convolution operation, the graph convolution block used in ViG is significantly faster than the vision transformer and 3×3 convolution-based CNN blocks. Therefore, we propose to use the graph convolution block to decode feature maps for dense prediction. This will make our decoder computationally efficient, while preserving long-range information.

2.3. Medical image segmentation

Medical image segmentation is the task of classifying pixels into lesions, tumours, or organs in a medical image (e.g., endoscopy, MRI, and CT) [3]. To address this task, U-shaped architectures [25, 22, 40, 13, 21] have been commonly utilized due to their sophisticated encoder-decoder structure. Ronneberger et al. [13] introduce UNet, an encoder-decoder architecture that utilizes skip connections to aggregate features from multiple stages. In UNet++ [40], nested encoder-decoder sub-networks are connected through dense skip connections. UNet 3+ [13] further extends this concept by exploring full-scale skip connections with intra-connections among the decoder blocks. DC-UNet [21] incorporates the multi-resolution convolution block and residual path within skip connections. These architectures have proven to be effective in medical image segmentation tasks.

Recently, transformers have gained popularity in the field of medical image segmentation [2, 3, 8, 23, 24, 29, 39]. In TransUNet [3], a hybrid architecture combining CNNs and transformers is proposed to capture both local and global pixel relationships. Swin-Unet [2] adopts a pure U-shaped transformer structure by utilizing Swin transformer blocks [19] in both the encoder and decoder. More recently, Rahman et al. [24] propose a multi-scale hierarchical transformer network with cascaded attention decoding (MERIT) that calculates self attention in varying window sizes to cap-
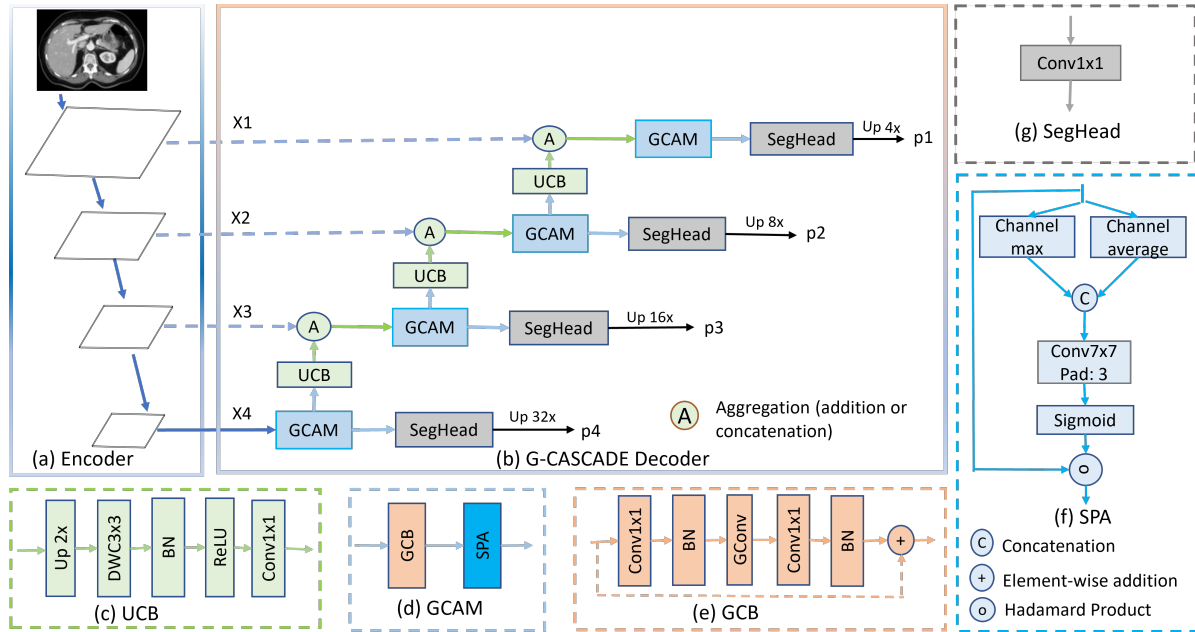
Figure 1. Hierarchical encoder with G-CASCADE network architecture. (a) PVTv2-b2 Encoder backbone with four stages, (b) G-CASCADE decoder, (c) Up-convolution block (UCB), (d) Graph convolutional attention module (GCAM), (e) Graph convolution block (GCB), (f) Spatial attention (SPA), and (g) Segmentation head (SegHead). X1, X2, X3, and X4 are the output features of the four stages of hierarchical encoder. p1, p2, p3, and p4 are output segmentation maps from four stages of our decoder.

ture effective multi-scale features.

Attention mechanisms have also been explored in combination with both CNNs [22, 10] and transformer-based architectures [8] in medical image segmentation. PraNet [10] utilizes the reverse attention mechanism [5]. In PolypPVT [8], authors employ PVTv2 [33] as the encoder and integrates CBAM [35] attention blocks in the decoder, along with other modules. CASCADE [23] proposes a cascaded decoder that utilizes both channel attention [12] and spatial attention [4] modules for feature refinement. CASCADE extracts features from four stages of the transformer encoder and uses cascaded refinement to generate high-resolution segmentation maps. Due to incorporating local information with global information of transformers, CASCADE exhibits remarkable performance in medical image segmentation. However, CASCADE decoder has two main limitations: i) long-range attention deficit due to using only convolution operations during decoding and ii) high computational inefficiency due to using three 3×3 convolutions in each stage of the decoder. We propose to use graph convolution to overcome these limitations.

3. Method

In this section, we first introduce a new G-CASCADE decoder, then explain two different transformer-based architectures (i.e., PVT-GCASCADE and MERIT-GCASCADE) incorporating our proposed decoder.

3.1. Cascaded Graph Convolutional Decoder (G-CASCADE)

Existing transformer-based models have limited (local) contextual information processing ability among pixels. As a result, the transformer-based model faces difficulties in locating the more discriminating local features. To address this issue, some works [8, 23, 24] utilize computationally expensive 2D convolution blocks in the decoder. Although the convolution block helps to incorporate the local information, it results in long-range attention deficits. To overcome this problem, we propose a new cascaded graph convolutional decoder, G-CASCADE, for pyramid encoders.

As shown in Figure 1(b), G-CASCADE consists of efficient up-convolution blocks (UCBs) to upsample the features, graph convolutional attention modules (GCAMs) to robustly enhance the feature maps, and segmentation heads (SegHeads) to get the segmentation output. We have four GCAMs for the four stages of pyramid features from the encoder. To aggregate the multi-scale features, we first aggregate (via addition or concatenation) the upsampled features from the previous decoder block with the features from the skip connections. Afterward, we process the concatenated features using our GCAM for enhancing semantic information. We then send the output from each GCAM to a prediction head. Finally, we aggregate four different prediction maps to produce the final segmentation output.

3.1.1 Graph convolutional attention module (GCAM)

We use the graph convolutional attention modules to refine the feature maps. GCAM consists of a graph convolution block ($GCB(.)$) to refine the features preserving long-range attention and a spatial attention [4] ($SPA(\cdot)$) block to capture the local contextual information as in Equation 1:

$$GCAM(x) = SPA(GCB(x)) \qquad (1)$$

where x is the input tensor and $GCAM(\cdot)$ represents the convolutional attention module. Due to using graph convolution, our GCAM is significantly more efficient than the convolutional attention module (CAM) proposed in [23].

Graph Convolution Block (GCB): The GCB is used to enhance the features generated using our cascaded expanding path. In our GCB, we follow the Grapher design of Vision GNN [11]. GCB consists of a graph convolution layer $GConv(.)$ and two 1×1 convolution layers $C(\cdot)$ each followed by a batch normalization layer $BN(\cdot)$ and a ReLU activation layer $R(.)$. $GCB(\cdot)$ is formulated as Equation 2:

$$GCB(x) = R(BN(C(GConv(R(BN(C(x))))))) \quad (2)$$

where $GConv$ can be formulated using Equation 3:

$$GConv(x) = GELU(BN(DynConv(x))) \qquad (3)$$

where $DynConv(.)$ is a graph convolution (e.g., max-relative, edge, GraphSAGE, and GIN) in dense dilated K-nearest neighbour (KNN) graph. $BN(.)$ and $GELU(.)$ are batch normalization and GELU activation, respectively.

SPatial Attention (SPA): The SPA determines *where* to focus in a feature map; then it enhances those features. The spatial attention is formulated as Equation 4:

$$SPA(x) = Sigmoid(Conv([C_{max}(x), C_{avg}(x)])) \circledast x \quad (4)$$

where $Sigmoid(\cdot)$ is a Sigmoid activation function. $C_{max}(\cdot)$ and $C_{avg}(\cdot)$ represent the maximum and average values obtained along the channel dimension, respectively. $Conv(\cdot)$ is a 7×7 convolution layer with padding 3 to enhance local contextual information (as in [8]). $\circledast$ is the Hadamard product.

3.1.2 Up-convolution block (UCB)

UCB progressively upsamples the features of the current layer to match the dimension to the next skip connection. Each UCB layer consists of an UpSampling $Up(\cdot)$ with scale-factor 2, a 3×3 depth-wise convolution $DWC(\cdot)$ with groups equal input channels, a batch normalization $BN(\cdot)$, a $ReLU(.)$ activation, and a 1×1 convolution $Conv(.)$. The $UCB(\cdot)$ can be formulated as Equation 5:

$$UCB(x) = Conv(ReLU(BN(DWC(Up(x))))) \quad (5)$$

Our UCB is light-weight as we replace the 3×3 convolution with a depth-wise convolution after upsampling.

3.1.3 Segmentation head (SegHead)

SegHead takes refined feature maps from the four stages of the decoder as input and predicts four output segmentation maps. Each SegHead layer consists of a 1×1 convolution $Conv_{1 \times 1}(\cdot)$ which takes feature maps having N_i channels (N_i is the number of channels in the feature map of stage i) as input and gives output with channels equal to number of target classes for multi-class but 1 channel for binary prediction. The $SegHead(\cdot)$ is formulated as in Equation 6:

$$SegHead(x) = Conv_{1 \times 1}(x) \qquad (6)$$

3.2. Overall architecture

To ensure effective generalization and the ability to process multi-scale features in medical image segmentation, we integrate our proposed G-CASCADE decoder with two different hierarchical backbone encoder networks such as PVTv2 [33] and MERIT [24]. PVTv2 utilizes convolution operations instead of traditional transformer patch embedding modules to consistently capture the spatial information. MERIT utilizes two MaxViT [27] encoders with varying window sizes for self-attention, thus enabling it to capture multi-scale features.

By utilizing the PVTv2-b2 (Standard) encoder [33], we create the PVT-GCASCADE architecture. To adopt PVTv2-b2, we first extract the features (X1, X2, X3, and X4) from four layers and feed them (i.e., X4 in the upsample path and X3, X2, X1 in the skip connections) into our G-CASCADE decoder as shown in Figure 1(a-b). Then, the G-CASCADE processes them and produces four prediction maps that correspond to the four stages of the encoder network.

Moreover, we introduce the new MERIT-GCASCADE architecture by adopting the architectural design of the MERIT network [24]. In the case of MERIT, we only replace their decoder with our proposed decoder and keep their hybrid CNN-transformer MaxViT [27] encoder networks. In our MERIT-GCASCADE architecture, we extract hierarchical feature maps from four stages of first encoder and then feed them to the corresponding decoder. Afterward, we aggregate the feedback from final stage of the decoder to the input image and feed them to the second encoder which has different window sizes for self-attention. We extract feature maps from four stages of the second decoder and feed them to the second decoder. We send cascaded skip connections like MERIT [24] to the second decoder. We get four output segmentation maps from the four stages of our second decoder. Finally, we aggregate the segmentation maps from the two decoders for four stages separately to produce four output segmentation maps. Our proposed decoder is designed to be adaptable and seamlessly integrates with other hierarchical backbone networks.

3.3. Multi-stage outputs and loss aggregation

We get four output segmentation maps p_1, p_2, p_3, and p_4 from the four prediction heads for the four stages of our G-CASCADE decoder.

Output segmentation maps aggregation: We compute the final segmentation output using additive aggregation:

$$seg\_output = \alpha p_1 + \beta p_2 + \gamma p_3 + \zeta p_4 \qquad (7)$$

where α, β, γ, and ζ are the weights of each prediction head. We set α, β, γ, and ζ to 1.0 in all our experiments. We get the final prediction output by applying the Sigmoid activation for binary segmentation and Softmax activation for multi-class segmentation.

Loss aggregation: Following MERIT [24], we use the combinatorial loss aggregation strategy, MUTATION in all our experiments. Therefore, we compute the loss for $2^n - 1$ combinatorial predictions synthesized from n heads separately and then do a summation of them. We optimize this additive combinatorial loss during training.

4. Experimental Evaluation

In this section, we first describe the datasets and evaluation metrics, followed by implementation details. Then, we conduct a comparative analysis between our proposed G-CASCADE decoder-based architectures and SOTA methods to highlight the superior performance of our approach.

4.1. Datasets

We present the description of Synapse Multi-organ and ACDC datasets below. The description of ISIC2018, polyp, and retinal vessel segmentation datasets are available in supplementary materials (Section A).

Synapse Multi-organ dataset. The Synapse Multi-organ dataset[1] contains 30 abdominal CT scans which have 3779 axial contrast-enhanced slices. Each CT scan has 85-198 slices of 512×512 pixels. Similar to TransUNet [3], we divide the dataset randomly into 18 scans for training (2212 axial slices) and 12 scans for validation. We segment only 8 abdominal organs, i.e., aorta, gallbladder (GB), left kidney (KL), right kidney (KR), liver, pancreas (PC), spleen (SP), and stomach (SM).

ACDC dataset. The ACDC dataset[2] contains 100 cardiac MRI scans each of which consists of three sub-organs, right ventricle (RV), myocardium (Myo), and left ventricle (LV). Following TransUNet [3], we use 70 cases (1930 axial slices) for training, 10 for validation, and 20 for testing.

4.2. Evaluation metrics

We use DICE, mIoU, and 95% Hausdorff Distance (HD95) to evaluate performance on the Synapse Multi-

---

organ dataset. However, for the ACDC dataset, we use only DICE score as an evaluation metrics. We use DICE and mIoU as the evaluation metrics in polyp segmentation and ISIC2018 datasets. The DICE score $DSC(Y, \hat{Y})$, $IoU(Y, \hat{Y})$, and HD95 distance $D_H(Y, \hat{Y})$ are calculated using Equations 8, 9, and 10, respectively:

$$DSC(Y, \hat{Y}) = \frac{2 \times |Y \cap \hat{Y}|}{|Y| + |\hat{Y}|} \times 100 \qquad (8)$$

$$IoU(Y, \hat{Y}) = \frac{|Y \cap \hat{Y}|}{|Y \cup \hat{Y}|} \times 100 \qquad (9)$$

$$D_H(Y, \hat{Y}) = \max\{\max_{y \in Y} \min_{\hat{y} \in \hat{Y}} d(y, \hat{y}), \{\max_{\hat{y} \in \hat{Y}} \min_{y \in Y} d(y, \hat{y})\} \quad (10)$$

where Y and $\hat{Y}$ are the ground truth and predicted segmentation map, respectively.

4.3. Implementation details

We use Pytorch 1.11.0 to implement our network and conduct experiments. We train all models on a single NVIDIA RTX A6000 GPU with 48GB of memory. We use the PVTv2-b2 and Small CascadedMERIT as representative networks. We use the pre-trained weights on ImageNet for both PVT and MERIT backbone networks. We train our model using AdamW optimizer [20] with both learning rate and weight decay of 0.0001.

GCB: We construct dense dilated graph using $K = 11$ neighbors for KNN and use the *Max-Relative (MR)* graph convolution in all our experiments. The *batch normalization* is used after MR graph convolution. Following ViG [11], we also use the relative position vector for graph construction and reduction ratios of [1, 1, 4, 2] for graph convolution block in different stages.

Synapse Multi-organ dataset. We use a batch size of 6 and train each model for maximum of 300 epochs. We use the input resolution of 224×224 for PVT-GCASCADE and $(256 \times 256, 224 \times 224)$ for MERIT-GCASCADE. We apply random rotation and flipping for data augmentation. The combined weighted Cross-entropy (0.3) and DICE (0.7) loss are utilized as the loss function.

ACDC dataset. For the ACDC dataset, we train each model for a maximum of 150 epochs with a batch size of 12. We set the input resolution as 224×224 for PVT-GCASCADE and $(256 \times 256, 224 \times 224)$ for MERIT-GCASCADE. We apply random flipping and rotation for data augmentation. We optimize the combined weighted Cross-entropy (0.3) and DICE (0.7) loss function.

Polyp datasets. We resize the image to 352×352 and use a multi-scale $\{0.75, 1.0, 1.25\}$ training strategy with a gradient clip limit of 0.5 like CASCADE [23]. We use a batch size of 4 and train each model a maximum of

Architectures	DICE↑	Average HD95↓	mIoU↑	Aorta	GB	KL	KR	Liver	PC	SP	SM
UNet [25]	70.11	44.69	59.39	84.00	56.70	72.41	62.64	86.98	48.73	81.48	67.96
AttnUNet [22]	71.70	34.47	61.38	82.61	61.94	76.07	70.42	87.54	46.70	80.67	67.66
R50+UNet [3]	74.68	36.87	−	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92
R50+AttnUNet [3]	75.57	36.97	−	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
SSFormerPVT [31]	78.01	25.72	67.23	82.78	63.74	80.72	78.11	93.53	61.53	87.07	76.61
PolypPVT [8]	78.08	25.61	67.43	82.34	66.14	81.21	73.78	94.37	59.34	88.05	79.4
TransUNet [3]	77.61	26.9	67.32	86.56	60.43	80.54	78.53	94.33	58.47	87.06	75.00
SwinUNet [2]	77.58	27.32	66.88	81.76	65.95	82.32	79.22	93.73	53.81	88.04	75.79
MT-UNet [30]	78.59	26.59	−	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
MISSFormer [14]	81.96	18.20	−	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
PVT-CASCADE [23]	81.06	20.23	70.88	83.01	70.59	82.23	80.37	94.08	64.43	90.1	83.69
TransCASCADE [23]	82.68	17.34	73.48	86.63	68.48	87.66	84.56	94.43	65.33	90.79	83.52
Cascaded MERIT [24]	84.32	14.27	75.44	86.67	72.63	87.71	84.62	95.02	**70.74**	**91.98**	**85.17**
PVT-GCASCADE (**Ours**)	83.28	15.83	73.91	86.50	71.71	87.07	83.77	95.31	66.72	90.84	83.58
MERIT-GCASCADE (**Ours**)	**84.54**	**10.38**	**75.83**	**88.05**	**74.81**	**88.01**	**84.83**	**95.38**	69.73	91.92	83.63

Table 1. Results of Synapse Multi-organ segmentation. We report only DICE scores for individual organs. We get the results of UNet, AttnUNet, PolypPVT, SSFormerPVT, TransUNet, and SwinUNet from [23]. We reproduce the results of Cascaded MERIT with a batch size of 6. ↑ (↓) denotes the higher (lower) the better. G-CASCADE results are averaged over five runs. The best results are shown in bold.

Methods	Avg Dice	RV	Myo	LV
R50+UNet [3]	87.55	87.10	80.63	94.92
R50+AttnUNet [3]	86.75	87.58	79.20	93.47
ViT+CUP [3]	81.45	81.46	70.71	92.18
R50+ViT+CUP [3]	87.57	86.07	81.88	94.75
TransUNet [3]	89.71	86.67	87.27	95.18
SwinUNet [2]	88.07	85.77	84.42	94.03
MT-UNet [30]	90.43	86.64	89.04	95.62
MISSFormer [14]	90.86	89.55	88.04	94.99
PVT-CASCADE [23]	91.46	89.97	88.9	95.50
TransCASCADE [23]	91.63	90.25	89.14	95.50
Cascaded MERIT [24]	91.85	90.23	89.53	95.80
PVT-GCASCADE (**Ours**)	91.95	90.31	89.63	95.91
MERIT-GCASCADE (**Ours**)	**92.23**	**90.64**	**89.96**	**96.08**

Table 2. Results on ACDC dataset. DICE scores are reported for individual organs. We get the results of SwinUNet from [23]. G-CASCADE results are averaged over five runs. The best results are shown in bold.

200 epochs. We optimize the combined weighted BinaryCrossEntropy (BCE) and weighted IoU loss function.

ISIC2018 dataset: We resize the images into 384×384 resolution. Then, we train our model for 200 epochs with a batch size of 4 and a gradient clip of 0.5. We optimize the combined weighted BCE and weighted IoU loss function.

4.4. Results

We compare our architectures (i.e., PVT-GCASCADE and MERIT-GCASCADE) with SOTA CNN and transformer-based segmentation methods on Synapse Multi-organ, ACDC, ISIC2018 [7], and Polyp (i.e., Endoscene [28], CVC-ClinicDB [1], Kvasir [16], ColonDB [26]) datasets. The results of ISIC2018, polyp, and retinal vessel segmentation datasets are reported in the supplementary materials (Section B).

4.4.1 Quantitative results on Synapse Multi-organ dataset

Table 1 presents the performance of different CNN- and transformer-based methods on Synapse Multi-organ segmentation dataset. We can see from Table 1 that our MERIT-GCASCADE significantly outperforms all the SOTA CNN- and transformer-based 2D medical image segmentation methods thus achieving the best average DICE score of 84.54%. Our PVT-GCASCADE and MERIT-GCASCADE outperform their counterparts PVT-CASCADE and Cascaded MERIT by 2.22% and 0.22% DICE scores, respectively, with significantly lower computational costs. Similarly, our PVT-GCASCADE and MERIT-GCASCADE outperform their counterparts by 4.4 and 3.89 in HD95 distance. Our MERIT-GCASCADE has the lowest HD95 distance (10.38) which is 3.89 lower than the best SOTA method Cascaded MERIT (HD95 of 14.27). The lower HD95 scores indicate that our G-CASCADE decoder can better locate the boundary of organs.

Our proposed decoder also shows a boost in the DICE scores of individual organ segmentation. We can see from Table 1 that our proposed MERIT-GCASCADE significantly outperforms SOTA methods on five out of eight organs. We believe that G-CASCADE decoder demonstrates
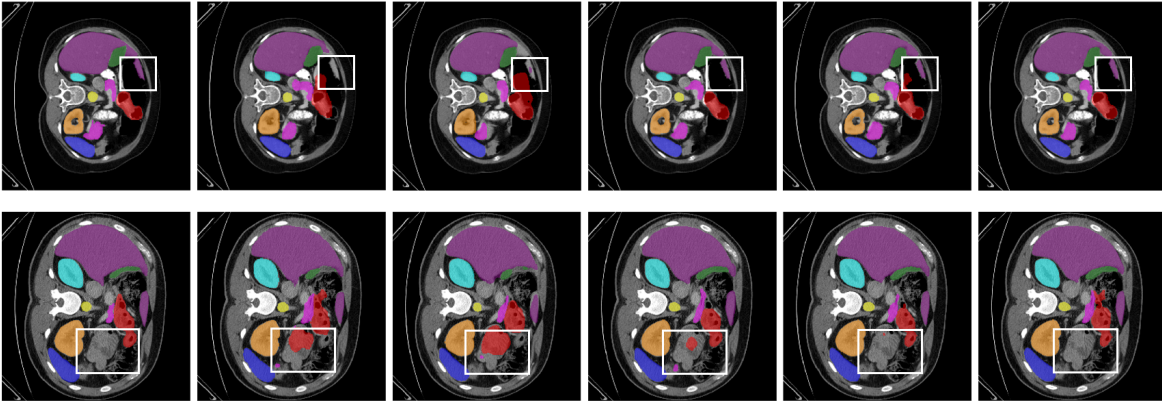
Figure 2. Qualitative results on Synapse multi-organ dataset. (a) Ground Truth (GT), (b) PVT-CASCADE, (c) TransCASCADE, (d) Cascaded MERIT, (e) PVT-GCASCADE, and (f) MERIT-GCASCADE. We overlay the segmentation maps on top of original image/slice. We use the white bounding box to highlight regions where most of the methods have incorrect predictions.

better performance due to using graph convolution together with the transformer encoder.

4.4.2 Quantitative results on ACDC dataset

We have conducted another set of experiments on the MRI images of the ACDC dataset using our architectures. Table 2 presents the average DICE scores of our PVT-GCASCADE and MERIT-GCASCADE along with other SOTA methods. Our MERIT-GCASCADE achieves the highest average DICE score of 92.23% thus improving about 0.38% over Cascaded MERIT though our decoder has significantly lower computational cost (see Table 5). Our PVT-GCASCADE gains 91.95% DICE score which is also better than all other methods. Besides, both our PVT-GCASCADE and MERIT-GCASCADE have better DICE scores in all three organs segmentation.

4.4.3 Qualitative results on Synapse Multi-organ dataset

We present the segmentation outputs of our proposed method and three other SOTA methods on two sample images in Figure 2. If we look into the highlighted regions in both samples, we can see that MERIT-GCASCADE consistently segments the organs with minimal false negative and false positive results. PVT-GCASCADE and Cascaded MERIT show comparable results. PVT-GCASCADE has false positives in first sample (i.e., first row) and has better segmentation in second sample (i.e., second row), whereas Cascaded MERIT provides better segmentation in first sample but it has larger false positives in second sample. Tran-

Components			#FLOPs	#Params	Avg
Cascaded	GCB	SPA	(G)	(M)	DICE
No	No	No	0	0	80.1±0.2
Yes	No	No	0.102	0.225	81.1±0.2
Yes	No	Yes	0.102	0.225	82.1±0.3
Yes	Yes	No	0.341	1.78	83.0±0.2
Yes	Yes	Yes	0.342	1.78	**83.3±0.2**

Table 3. Quantitative results of different components of G-CASCADE with PVTv2-b2 encoder on Synapse multi-organ dataset. We use *additive aggregation* for adding skip connections and an input resolution of 224×224. All results are averaged over five runs. The best results are shown in bold.

Arrangements	DICE (%)
SPA → GCB	82.93±0.2
GCB → SPA (**Ours**)	**83.28±0.2**

Table 4. Comparison of different arrangements of GCB and SPA in GCAM on Synapse Multi-organ dataset. We use PVTv2-b2 as the encoder to produce these results. All the results are averaged over five runs. The best results are in bold.

sCASCADE and PVT-CASCADE provide larger incorrect segmentation outputs in both samples.

5. Ablation Study

In this section, we perform a set of ablation experiments that address various questions concerning our proposed architectures and experimental setup. More ablation studies are available in supplementary materials (Section C).

Decoders	UCB	#FLOPs(G)	#Params(M)	DICE (%)
CASCADE	Original	1.93	9.27	82.78
CASCADE	Modified	1.22	7.58	82.79
G-CASCADE (**Ours**)	Original	1.06	3.47	83.15
G-CASCADE (**Ours**)	Modified	**0.342**	**1.78**	**83.28**

Table 5. Comparison with the baseline decoder on Synapse Multi-organ dataset. We only report the #FLOPs and the #parameters of the respective decoder. We produce these results using PVTv2-b2 encoder. All the results are averaged over five runs. The best results are shown in bold.

Architectures	Aggregation	#FLOPs(G)	#Params(M)	DICE (%)
PVT-GCASCADE	Addition	**0.342**	**1.78**	83.28
PVT-GCASCADE	Concat	0.975	3.32	**83.40**
MERIT-GCASCADE	Addition	**1.523**	**3.55**	84.54
MERIT-GCASCADE	Concat	4.27	5.99	**84.63**

Table 6. Comparison of different skip-aggregations in G-CASCADE decoder on Synapse Multi-organ dataset. We only report the #FLOPs and #parameters of the respective decoder. PVTV2-b2 encoder has 3.91G FLOPS and 24.86M parameters. Small MERIT encoder has 24.62G FLOPs and 129.38M parameters. All results are averaged over five runs. The best results are shown in bold.

5.1. Effect of different components of G-CASCADE

We carry out ablation studies on the Synapse Multi-organ dataset to evaluate the effectiveness of different components of our proposed G-CASCADE decoder. We use the same PVTv2-b2 backbone pre-trained on ImageNet and the same experimental settings for Synapse Multi-organ dataset in all experiments. We remove different modules such as Cascaded structure, GCB, and SPA from the G-CASCADE decoder and compare the results. It is evident from Table 3 that the cascaded structure of the decoder improves performance over the non-cascaded decoder. GCB and SPA modules also help improve performance. However, the use of both SPA and GCB modules together produces the best DICE score of 83.3%. We can also see from the table that DICE score is improved by about 3.2% with 0.342G and 1.78M additional FLOPs and parameters, respectively.

5.2. Effect of arrangements of GCB and SPA in GCAM

We have conducted an ablation study to see the effect of the order of GCB and SPA in GCAM. Table 4 presents the experimental results of two different arrangements. We can conclude from Table 4 that GCB followed by SPA block performs better than SPA followed by GCB. Therefore, in our G-CASCADE decoder, we use a GCB followed by a SPA block in each GCAM.

5.3. Comparison with the baseline decoder

Table 5 reports the experimental results with the computational complexity of our baseline CASCADE decoder and our proposed G-CASCADE decoder. We also report the results of original UpConv used in the CASCADE decoder and our modified efficient UCB. From Table 5, we can see that our modified UCB performs equal or better with significantly lower #FLOPs and #parameters. Our G-CASCADE decoder provides 0.5% better DICE score than the CASCADE decoder with 80.8% fewer parameters and 82.3% fewer FLOPs.

5.4. Effect of different skip-aggregations in G-CASCADE decoder

We conduct some experiments to see the effect of Additive or Concatenation in aggregating upsampled features with the skip-connections. Table 6 presents the results of PVT-GCASCADE and MERIT-GCASCADE with Additive and concatenation aggregations. We can see from Table 6 that Concatenation-based aggregation achieves marginally better DICE scores than Additive aggregation while having significantly higher #FLOPs and #parameters. The reason behind this increase in computational complexity is the use of GCAM with the concatenated channels (i.e., $2\times$ number of original channels). Considering the lower computational complexity of Additive aggregation, we have used Additive aggregation in all of our experiments.

6. Conclusion

In this paper, we have introduced a new graph-based cascaded convolutional attention decoder namely G-CASCADE for multi-stage feature aggregation. G-CASCADE enhances feature maps while preserving long-range information captured by transformers which is crucial for accurate medical image segmentation. Due to using graph convolution blocks instead of 3×3 convolution blocks, G-CASCADE is computationally very efficient. Our experimental results show that G-CASCADE outperforms a recent decoder, CASCADE, in DICE scores with 80.8% fewer parameters and 82.3% fewer FLOPs. Our experimental results also demonstrate the superiority of our G-CASCADE decoder over SOTA methods on five public medical image segmentation benchmarks. Finally, we believe that our proposed decoder will improve other downstream medical image segmentation and semantic segmentation tasks.

Acknowledgment

References

[1] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015.

[2] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

[3] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[4] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5659–5667, 2017.

[5] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–250, 2018.

[6] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 2021.

[7] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

[8] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*, 2021.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 263–273. Springer, 2020.

[11] Kai Han, Yunhe Wang, Jianyuan Guo, Yehui Tang, and Enhua Wu. Vision gnn: An image is worth graph of nodes. *arXiv preprint arXiv:2206.00272*, 2022.

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[13] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059. IEEE, 2020.

[14] Xiaohong Huang, Zhifang Deng, Dandan Li, and Xueguang Yuan. Missformer: An effective medical image segmentation transformer. *arXiv preprint arXiv:2109.07162*, 2021.

[15] Md Amirul Islam, Sen Jia, and Neil DB Bruce. How much position information do convolutional neural networks encode? *arXiv preprint arXiv:2001.08248*, 2020.

[16] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, pages 451–462. Springer, 2020.

[17] Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4558–4567, 2018.

[18] Guohao Li, Matthias Muller, Ali Thabet, and Bernard Ghanem. Deepgcns: Can gcns go as deep as cnns? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9267–9276, 2019.

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[21] Ange Lou, Shuyue Guan, and Murray Loew. Dc-unet: rethinking the u-net architecture with dual channel efficient cnn for medical image segmentation. In *Medical Imaging 2021: Image Processing*, volume 11596, pages 758–768. SPIE, 2021.

[22] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[23] Md Mostafijur Rahman and Radu Marculescu. Medical image segmentation via cascaded attention decoding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6222–6231, January 2023.

[24] Md Mostafijur Rahman and Radu Marculescu. Multi-scale hierarchical vision transformer with cascaded attention decoding for medical image segmentation. In *Medical Imaging with Deep Learning*, 2023.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation.

In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.

[26] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Medical Imaging*, 35(2):630–644, 2015.

[27] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 459–479. Springer, 2022.

[28] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of Healthcare Engineering*, 2017, 2017.

[29] Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2441–2449, 2022.

[30] Hongyi Wang, Shiao Xie, Lanfen Lin, Yutaro Iwamoto, Xian-Hua Han, Yen-Wei Chen, and Ruofeng Tong. Mixed transformer u-net for medical image segmentation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2390–2394. IEEE, 2022.

[31] Jinfeng Wang, Qiming Huang, Feilong Tang, Jia Meng, Jionglong Su, and Sifan Song. Stepwise feature fusion: Local guides global. *arXiv preprint arXiv:2203.03635*, 2022.

[32] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 568–578, 2021.

[33] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.

[34] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17683–17693, 2022.

[35] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[36] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[37] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5419, 2017.

[38] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[39] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 14–24. Springer, 2021.

[40] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018.