A. Datasets

ISIC2018 dataset. ISIC2018 dataset is a skin lesion segmentation dataset [4]. It consists of 2596 images with corresponding annotations. In our experiments, we resize the images to 384 $\times$ 384 resolution unless otherwise mentioned. We randomly split the images into 80% for training, 10% for validation, and 10% for testing.

Polyp datasets. Kvasir contains 1,000 polyp images collected from the polyp class in the Kvasir-SEG dataset [8]. CVC-ClinicDB [1] consists of 612 images extracted from 31 colonoscopy videos. Following CASCADE [14], we adopt the same 900 and 550 images from Kvasir and CVC-ClinicDB, respectively as the training set. We use the remaining 100 and 62 images as the respective testsets. To assess the generalizability of our proposed decoder, we use two unseen test datasets, namely EndoScene [21], and ColonDB [18]. EndoScene and ColonDB consists of 60 and 380 images, respectively.

Retinal vessels segmentation datasets. The DRIVE [17] dataset has 40 retinal images with segmentation annotations. All the retinal images in this dataset are 8-bit color images of resolution 565×584 pixels. The official splits contain a training set of 20 images and a test set of 20 images. The CHASE_DB1 [2] dataset contains 28 color retina images of 999×960 pixels resolution. There are two manual annotations of each image for segmentation. We use the first annotation as the ground truth. Following [11], we use the first 20 images for training, and the remaining 8 images for testing.

B. Experiments

B.1. Implementation details and evaluation metrics

In this subsection, we discuss the implementation details of our proposed decoder for Retinal vessel segmentation. We have conducted experiments on two retinal datasets such as DRIVE [17] and CHASE_DB1 [2]. In both cases, we first extend the training set using horizontal flips, vertical flips, horizontal-vertical flips, random rotations, random colors, and random Gaussian blurs. Through this process, we get 260 images including our 20 original training images. We use 26 of these images for validation that belong to 4 randomly selected original images. In the case of the DRIVE dataset, we resize the images into 768×768 resolution for PVT and (768×768, 672×672) resolutions for MERIT. In the case of CHASE_DB1, we use 960×960 resolution inputs for PVT and (768×768, 672×672) resolution inputs for MERIT. However, we resize the output segmentation maps to the original resolution to get evaluation metrics during inference. We use random flips and rotations with a probability of 0.5 as augmentation methods during training. To train our models, we use the AdamW optimizer with both learning rate and weight decay of 1e-4. We optimize the combined weighted BCE and weighted mIoU loss function. The MUTATION is used to aggregate multi-stage loss. We train our networks for 200 epochs with a batch size of 4 and 2 for DRIVE and CHASE_DB, respectively.

We use accuracy (Acc), sensitivity (Sen), specificity (Sp), DICE, and IoU scores as evaluation metrics. We report the percentage (%) score averaging over five runs for both datasets.

Methods	Avg DICE	Avg mIoU
UNet [16]	85.5	78.5
UNet++ [28]	80.9	72.9
PraNet [6]	87.5	78.7
CaraNet [12]	87.0	78.2
TransUNet [3]	88.0	80.9
TransFuse [27]	90.1	84.0
UCTransNet [22]	90.5	83.0
PolypPVT [5]	91.3	85.2
PVT-CASCADE [14]	91.1	84.9
PVT-GCASCADE (**Ours**)	**91.51**$\pm$**0.61**	**86.53**$\pm$**0.54**

Table 1. Results on ISIC2018 dataset. The results of UNet, UNet++, PraNet, CaraNet, TransUNet, TransFuse, UCTransNet, and PolypPVT are taken from [19]. We produce the results of PVT-CASCADE using our experimental settings for this dataset. All PVT-GCASCADE results are averaged over five runs. The best results are in bold.

B.2. Experimental results on ISIC2018 dataset

Table 1 presents the average DICE scores of our PVT-GCASCADE and MERIT-GCASCADE along with other SOTA methods on the ISIC2018 dataset. This dataset is different than the CT and MRI images used in the above experiments. In this case also, it is evident from the table that our PVT-GCASCADE achieves the best average DICE (91.51%) and mIoU (86.53%) scores. PVT-GCASCADE outperforms its counterpart PVT-CASCADE by 0.4% DICE and 0.6% mIoU scores.

B.3. Experimental results on Polyp datasets

We evaluate the performance and generalizability of our G-CASCADE decoder on four different polyp segmentation testsets among which two are completely unseen datasets collected from different labs. Table 2 displays the DICE and mIoU scores of SOTA methods along with our G-CASCADE decoder. From Table 2, we can see that G-CASCADE significantly outperforms all other methods on both DICE and mIoU scores. It is noteworthy that G-CASCADE outperforms the best CNN-based model UACANet by a large margin on unseen datasets (i.e., 9.8% DICE score improvement in ColonDB). Therefore, we can conclude that due to using transformers as a backbone

Methods	CVC-ClinicDB		Kvasir		ColonDB		EndoScene	
	DICE	mIoU	DICE	mIoU	DICE	mIoU	DICE	mIoU
UNet [16]	82.3	75.5	81.8	74.6	51.2	44.4	71.0	62.7
UNet++ [28]	79.4	72.9	82.1	74.3	48.3	41.0	70.7	62.4
PraNet [6]	89.9	84.9	89.8	84.0	71.2	64.0	87.1	79.7
CaraNet [12]	93.6	88.7	91.8	86.5	77.3	68.9	90.3	83.8
UACANet-L [9]	91.07	86.7	90.83	85.95	72.57	65.41	88.21	80.84
SSFormerPVT [23]	92.88	88.27	91.11	86.01	79.34	70.63	89.46	82.68
PolypPVT [5]	93.08	88.28	91.23	86.3	80.75	71.85	88.71	81.89
PVT-CASCADE [14]	94.34	89.98	92.58	87.76	82.54	74.53	90.47	83.79
PVT-GCASCADE (**Ours**)	**94.68**	**90.18**	**92.74**	**87.90**	**82.61**	**74.60**	**90.56**	**83.87**

Table 2. Results on polyp segmentation datasets. Training on combined Kvasir [8] and CVC-ClinicDB [1] trainset. The results of UNet, UNet++ and PraNet are taken from [6]. We get the results of PolypPVT, SSFormerPVT, and UACANet from [14]. PVT-GCASCADE results are averaged over five runs. The best results are shown in bold.

Methods	Acc	Sen	Sp	DICE	IoU
UNet [16]	96.78	80.57	98.33	81.41	68.64
UNet++ [28]	96.79	78.91	**98.50**	81.14	68.27
Attention UNet [13]	96.62	79.06	98.31	80.39	67.21
FR-UNet [11]	97.05	**83.56**	98.37	**83.16**	**71.20**
PVTV2-b2 (only) [24]	96.24	82.02	97.61	79.14	65.48
PVT-CASCADE [14]	96.79	83.07	98.10	81.73	69.10
MERIT-CASCADE [15]	96.89	82.94	98.22	82.21	69.08
PVT-GCASCADE (**Ours**)	96.89	83.00	98.22	82.10	69.70
MERIT-GCASCADE (**Ours**)	**97.07**	82.81	98.44	82.90	70.81

Table 3. Results (%) of Retinal Vessel Segmentation on DRIVE dataset. The results of UNet, UNet++, Attention UNet, and FR-UNet are taken from [11]. All other results are averaged over five runs in our experimental setups. The best results are in bold.

Methods	Acc	Sen	Sp	DICE	IoU
UNet [16]	97.43	76.50	98.84	78.98	65.26
UNet++ [28]	97.39	83.57	98.32	80.15	66.88
Attention UNet [13]	97.30	83.84	98.20	79.64	66.17
FR-UNet [11]	97.48	**87.98**	98.14	81.51	68.82
PVTV2-b2 (only) [24]	97.25	85.07	98.07	79.58	66.12
PVT-CASCADE [14]	97.55	85.83	98.34	81.50	68.80
MERIT-CASCADE [15]	97.60	84.97	98.45	81.68	69.06
PVT-GCASCADE (**Ours**)	97.71	85.84	98.51	82.51	70.24
MERIT-GCASCADE (**Ours**)	**97.76**	84.93	**98.62**	**82.67**	**70.50**

Table 4. Results (%) of Retinal Vessel Segmentation on CHASE_DB1 dataset. The results of UNet, UNet++, Attention UNet, and FR-UNet are taken from [11]. All other results are averaged over five runs in our experimental setups. The best results are in bold.

network and our graph-based convolutional attention decoder, PVT-GCASCADE inherits the merits of transformers, GCNs, CNNs, and local attention which makes them highly generalizable for unseen datasets.

B.4. Experimental results on Retinal vessels segmentation datasets

We have conducted experiments on two retinal vessel segmentation datasets, namely DRIVE and CHASE_DB1. The experimental results are reported in Tables 3 and 4. Our G-CASCADE decoder outperforms the baseline CASCADE decoder with significantly lower computational costs. Specifically, our PVT-GCASCADE shows 0.37% and 1.01% improvements in DICE score over PVT-CASCADE in DRIVE and CHASE_DB1 datasets, respectively. Similarly, our MERIT-GCASCADE exhibits 0.69% and 0.99% improvements in DICE score in DRIVE and CHASE_DB1 datasets, respectively. From Tables 3 and 4, we can conclude that our methods show competitive performance compared to the SOTA approaches. Although FR-UNet achieves a 0.26% better DICE score in the DRIVE dataset, it has a 1.16% lower DICE score in CHASE_DB1 than our MERIT-GCASCADE. Besides, FR-UNet splits the retinal images into 48×48 pixels patches in a stride of 6 pixels during training but we use the whole retinal images during both training and inference. Consequently, we have a significantly lower number of samples for training compared to FR-UNet. We can conclude from the results that our G-CASCADE decoder equally performs well in retinal vessel segmentation.

C. Ablation Study

C.1. Comparison among different graph convolutions in GCAM

We report the experimental results of our decoder with different graph convolutions in Table 5. As shown in Table 5, Max-Relative (MR) [10] graph convolution provides the best DICE score (83.28%) with only 0.342G FLOPs and 1.78M parameters. Although GIN [26] has slightly lower FLOPs and parameters, it provides the lowest DICE score

Graph Convolutions	#FLOPs(G)	#Params(M)	DICE (%)
GIN [26]	0.313	1.59	82.22
EdgeConv [25]	0.957	1.78	82.81
GraphSAGE [7]	0.520	1.88	83.10
Max-Relative [10] (**Ours**)	0.342	1.78	**83.28**

Table 5. Experimental results of different graph convolutions in GCAM block on Synapse Multi-organ dataset. We use the PVTV2-b2 encoder and only report the #FLOPs and #parameters of the decoder. All the results are averaged over five runs. The best results are shown in bold.

Architectures	#FLOPs(G)	#Params(M)	DICE (%)
PVT-CASCADE	5.84	34.13	83.28
PVT-GCASCADE	**4.252**	**26.64**	**83.40**
MERIT-CASCADE	33.31	147.86	84.54
MERIT-GCASCADE	**26.143**	**132.93**	**84.63**

Table 6. Comparison of overall computational complexity. We use the PVTV2-b2 backbone with an input resolution of 224×224 in both PVT-CASCADE and PVT-GCASCADE. We use two Small MaxViT backbones with input resolutions of 256×256 and 224×224 in MERIT architectures.

Input resolutions	DICE (%)	mIoU (%)	HD95 (%)
224×224	83.28	73.91	15.83
256×256	84.21	75.32	14.58
384×384	86.01	78.10	13.67

Table 7. Experimental results of PVT-GCASCADE with different input resolutions on Synapse Multi-organ dataset. All the results are averaged over five runs.

(82.22%). EdgeConv [25] and GraphSAGE [7] graph convolutions have lower DICE scores than the MR graph convolution with higher computational costs.

C.2. Overall computational complexity

We report the total #parameters and #FLOPs of encoder backbones and our decoder in Table 6. We can see from Table 6 that overall computational complexity depends on the #parameters and #FLOPs of the encoder backbones. We implement our decoder on top of PVTV2-b2 [24] and Small MaxViT [20] backbones. Our PVT-GCASCADE has 4.252G FLOPs and 26.64M parameters, which is 1.588G and 7.49M lower than the corresponding PVT-CASCADE architecture. Due to the larger size of two Small MaxViT backbones in MERIT-CASCADE architecture (i.e., 33.31G FLOPs and 147.86M parameters), our MERIT-GCASCADE (i.e., 26.143G FLOPs and 132.93M parameters) is also larger in size. In both cases, the savings in #FLOPs and #parameters come only from our decoder. Our proposed decoder can easily be plugged into other hierarchical encoders; if a lightweight encoder is used, the total computational cost will be reduced.

C.3. Influence of input resolution

Table 7 presents the quantitative segmentation performance of PVT-GCASCADE network with different input resolutions. We conduct experiments with three input resolutions such as 224×224, 256×256, and 384×384. It is evident from the table that performance improved in all three evaluation metrics for higher input resolutions. We get the best DICE and mIoU 86.01% and 78.10%, respectively with the input resolution of 384×384.

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] Adrian Carballal, Francisco J Novoa, Carlos Fernandez-Lozano, Marcos García-Guimaraes, Guillermo Aldama-López, Ramón Calviño-Santos, José Manuel Vazquez-Rodriguez, and Alejandro Pazos. Automatic multiscale vascular image segmentation algorithm for coronary angiography. *Biomedical Signal Processing and Control*, 46:1–9, 2018.

[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] 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.

[5] 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.

[6] 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.

[7] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, 30, 2017.

[8] 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 *Inter-*

national Conference on Multimedia Modeling, pages 451–462. Springer, 2020.

[9] Taehun Kim, Hyemin Lee, and Daijin Kim. Uacanet: Uncertainty augmented context attention for polyp segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2167–2175, 2021.

[10] 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.

[11] Wentao Liu, Huihua Yang, Tong Tian, Zhiwei Cao, Xipeng Pan, Weijin Xu, Yang Jin, and Feng Gao. Full-resolution network and dual-threshold iteration for retinal vessel and coronary angiograph segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26(9):4623–4634, 2022.

[12] Ange Lou, Shuyue Guan, Hanseok Ko, and Murray H Loew. Caranet: context axial reverse attention network for segmentation of small medical objects. In *Medical Imaging 2022: Image Processing*, volume 12032, pages 81–92. SPIE, 2022.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4):501–509, 2004.

[18] 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.

[19] Feilong Tang, Qiming Huang, Jinfeng Wang, Xianxu Hou, Jionglong Su, and Jingxin Liu. Duat: Dual-aggregation transformer network for medical image segmentation. *arXiv preprint arXiv:2212.11677*, 2022.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics*, 38(5):1–12, 2019.

[26] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[27] 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.

[28] 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.