EMCAD: Efficient Multi-scale Convolutional Attention Decoding for Medical Image Segmentation

Supplementary Material

7. Experimental Details

This section extends our Section 4 in the original paper by describing the datasets and evaluation metrics, followed by additional experimental results.

7.1. Datasets

To evaluate the performance of our EMCAD decoder, we carry out experiments across 12 datasets that belong to six medical image segmentation tasks, as described next.

Polyp segmentation: We use five polyp segmentation datasets: Kvasir [29] (1,000 images), ClinicDB [3] (612 images), ColonDB [51] (379 images), ETIS [51] (196 images), and BKAI [40] (1,000 images). These datasets contain images from different imaging centers/clinics, having greater diversity in image nature as well as size and shape of polyps.

Abdomen organ segmentation: We use the Synapse multi-organ dataset[1] for abdomen organ segmentation. This dataset contains 30 abdominal CT scans which have 3,779 axial contrast-enhanced slices. Each CT scan has 85-198 slices of 512 $\times$ 512 pixels. Following TransUNet [8], we use the same 18 scans for training (2,212 axial slices) and 12 scans for validation. We segment only eight abdominal organs, namely aorta, gallbladder (GB), left kidney (KL), right kidney (KR), liver, pancreas (PC), spleen (SP), and stomach (SM).

Cardiac organ segmentation: We use ACDC dataset[2] for cardiac organ segmentation. It contains 100 cardiac MRI scans having three sub-organs, namely right ventricle (RV), myocardium (Myo), and left ventricle (LV). Following TransUNet [8], we use 70 cases (1,930 axial slices) for training, 10 for validation, and 20 for testing.

Skin lesion segmentation: We use ISIC17 [15] (2,000 training, 150 validation, and 600 testing images) and ISIC18 [14] (2,594 images) for skin lesion segmentation.

Breast cancer segmentation: We use BUSI [1] dataset for breast cancer segmentation. Following [50], we use 647 (437 benign and 210 malignant) images from this dataset.

Cell nuclei/structure segmentation: We use the DSB18 [4] (670 images) and EM [6] (30 images) datasets of biological imaging for cell nuclei/structure segmentation.

We use a train-val-test split of 80:10:10 in ClinicDB, Kvasir, ColonDB, ETIS, BKAI, ISIC18, DSB18, EM, and BUSI datasets. For ISIC17, we use the official train-val-test sets provided by the competition organizer.

7.2. Evaluation metrics

We use the DICE score to evaluate performance on all the datasets. However, we also use 95% Hausdorff Distance (HD95) and mIoU as additional evaluation metrics for Synapse multi-organ segmentation. The DICE score $DSC(Y, P)$, $IoU(Y, P)$, and HD95 distance $D_H(Y, P)$ are calculated using Equations 12, 13, and 14, respectively:

$$DSC(Y, P) = \frac{2 \times |Y \cap P|}{|Y| + |P|} \times 100 \tag{12}$$

$$IoU(Y, P) = \frac{|Y \cap P|}{|Y \cup P|} \times 100 \tag{13}$$

$$D_H(Y, P) = \max\{\max_{y \in Y} \min_{p \in P} d(y, p), \{\max_{p \in P} \min_{y \in Y} d(y, p)\} \tag{14}$$

where Y and P are the ground truth and predicted segmentation map, respectively.

7.3. Qualitative results

This subsection describes the qualitative results of different methods including our EMCAD. From, the qualitative results on Synapse Multi-organ dataset in Figure 4, we can see that most of the methods face challenges segmenting the left kidney (orange) and part of the pancreas (pink). However, our PVT-EMCAD-B0 (Figure 4g) and PVT-EMCAD-B2 (Figure 4h) can segment those organs more accurately (see red rectangular box) with significantly lower computational costs. Similarly, qualitative results of polyp segmentation on a representative image from ClinicDB dataset in Figure 5 show that predicted segmentation outputs of our PVT-EMCAD-B0 (Figure 5p) and PVT-EMCAD-B2 (Figure 5q) have strong overlaps with the GroundTruth mask (Figure 5r), while existing SOTA methods exhibit false segmentation of polyp (see red rectangular box).

8. Additional Ablation Study

This section further elaborates on Section 5 by detailing five additional ablation studies related to our architectural design and experimental setup.

8.1. Parallel vs. sequential depth-wise convolution

We have conducted another set of experiments to decide whether we use multi-scale depth-wise convolutions in parallel or sequential. Table 7 presents the results of these experiments which show that there is no significant impact of

[1] https://www.synapse.org/#!Synapse:syn3193805/wiki/217789

[2] https://www.creatis.insa-lyon.fr/Challenge/acdc/

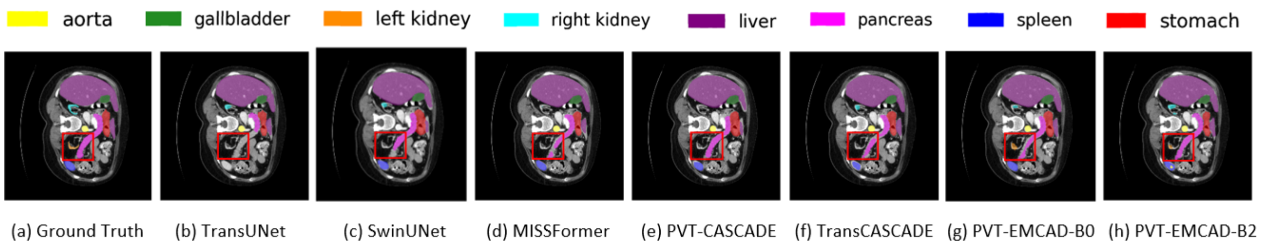

Figure 4. Qualitative results of multi-organ segmentation on Synapse Multi-organ dataset. The red rectangular box highlights incorrectly segmented organs by SOTA methods.

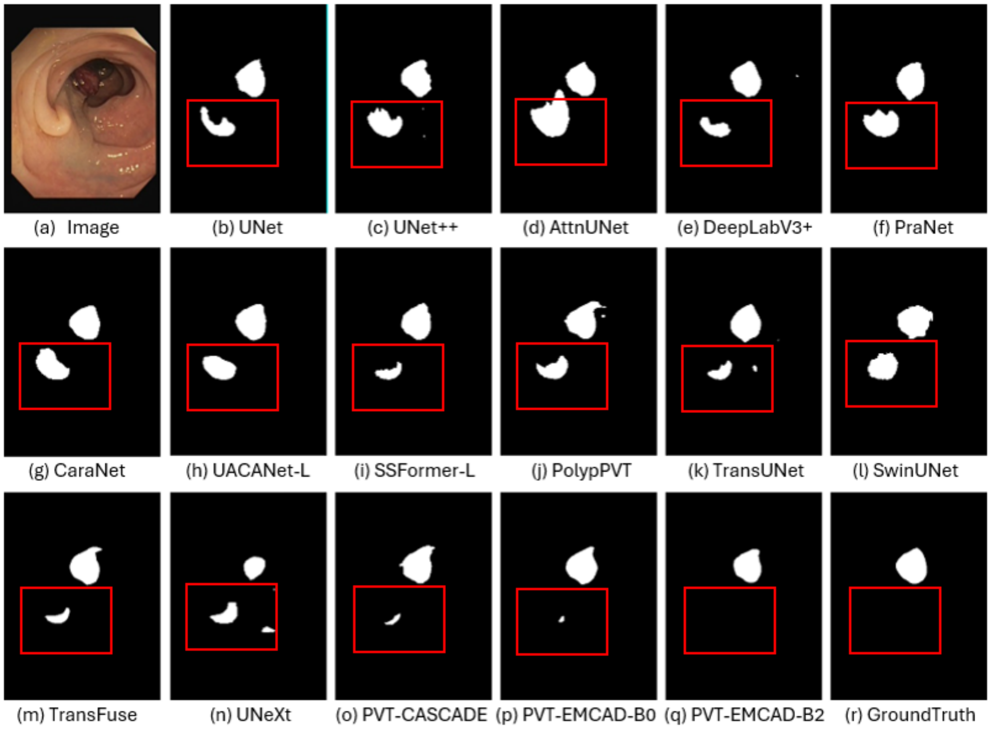

Figure 5. Qualitative results of polyp segmentation. The red rectangular box highlights incorrectly segmented polyps by SOTA methods.

Architectures	Depth-wise convolutions	Synapse	ClinicDB
PVT-EMCAD-B0	Sequential	81.82±0.3	94.57±0.2
PVT-EMCAD-B0	Parallel	**81.97±0.2**	**94.60±0.2**
PVT-EMCAD-B2	Sequential	83.54±0.3	95.15±0.3
PVT-EMCAD-B2	Parallel	**83.63±0.2**	**95.21±0.2**

Table 7. Results of parallel and sequential depth-wise convolution in MSDC on Synapse multi-organ and ClinicDB datasets. All results are averaged over five runs. Best results are in bold.

Architectures	Module	Params(K)	FLOPs(M)	Synapse
PVT-EMCAD-B0	AG	31.62	15.91	81.74
PVT-EMCAD-B0	**LGAG**	**5.51**	**5.24**	**81.97**
PVT-EMCAD-B2	AG	124.68	61.68	83.51
PVT-EMCAD-B2	**LGAG**	**11.01**	**10.47**	**83.63**

Table 8. LGAG vs. AG (Attention gate) [41] on Synapse multi-organ dataset. The total #Params and #FLOPs of three AG/LGAGs in our decoder are reported for an input resolution of 256 × 256. All results are averaged over five runs. Best results are in bold.

the arrangements though the parallel convolutions provide a slightly improved performance (0.03% to 0.15%). We also observe higher standard deviations among runs in the case of sequential convolutions. Hence, in all our experiments, we use multi-scale depth-wise convolutions *in parallel*.

8.2. Effectiveness of our large-kernel grouped attention gate (LGAG) over attention gate (AG)

Table 8 presents experimental results of EMCAD with original AG [41] and our LGAG. We can conclude that our LGAG achieves better DICE scores with significant re-

Architectures	Pretrain		Average		Aorta	GB	KL	KR	Liver	PC	SP	SM
		DICE↑	HD95↓	mIoU↑								
PVT-EMCAD-B0	No	77.47	19.93	66.72	81.96	**69.41**	83.88	74.82	93.45	54.41	88.97	72.85
PVT-EMCAD-B0	Yes	**81.97**	**17.39**	**72.64**	**87.21**	66.62	**87.48**	**83.96**	**94.57**	**62.00**	**92.66**	**81.22**
PVT-EMCAD-B2	No	80.18	18.83	70.21	85.98	68.10	84.62	79.93	93.96	61.61	90.99	76.23
PVT-EMCAD-B2	Yes	**83.63**	**15.68**	**74.65**	**88.14**	**68.87**	**88.08**	**84.10**	**95.26**	**68.51**	**92.17**	**83.92**

Table 9. Effect of transfer learning from ImageNet pre-trained weights on Synapse multi-organ dataset. ↑ (↓) denotes the higher (lower) the better. All results are averaged over five runs. Best results are in bold.

DS	EM	BUSI	Clinic	Kvasir	ISIC18	Synapse	ACDC
No	95.74	79.64	94.96	92.51	90.74	82.03	92.08
Yes	95.53	80.25	95.21	92.75	90.96	83.63	92.12

Table 10. Effect of deep supervision (DS). PVT-EMCAD-B2 with DS achieves slightly better DICE scores in 6 out of 7 datasets.

Architectures	Resolutions	FLOPs(G)	DICE
PVT-EMCAD-B0	224 × 224	0.64	81.97
PVT-EMCAD-B0	256 × 256	0.84	82.63
PVT-EMCAD-B0	384 × 384	1.89	84.81
PVT-EMCAD-B0	512 × 512	3.36	85.52
PVT-EMCAD-B2	224 × 224	4.29	83.63
PVT-EMCAD-B2	256 × 256	5.60	84.47
PVT-EMCAD-B2	384 × 384	12.59	85.78
PVT-EMCAD-B2	512 × 512	22.39	86.53

Table 11. Effect of input resolutions on Synapse multi-organ dataset. All results are averaged over five runs.

ductions in #Params (82.57% for PVT-EMCAD-B0 and 91.17% for PVT-EMCAD-B2) and #FLOPs (67.06% for PVT-EMCAD-B0 and 83.03% for PVT-EMCAD-B2) than AG. The reduction in #Params and #FLOPs is bigger for the larger models. Therefore, our LGAG demonstrates improved scalability with models that have a greater number of channels, yielding enhanced DICE scores.

8.3. Effect of transfer learning from ImageNet pre-trained weights

We conduct experiments on the Synapse multi-organ dataset to show the effect of transfer learning from the ImageNet pre-trained encoder. Table 9 reports the results of these experiments which show that transfer learning from ImageNet pre-trained PVT-v2 encoders significantly boosts the performance. Specifically, for PVT-EMCAD-B0, the DICE, mIoU, and HD95 scores are improved by 4.5%, 5.92%, and 2.54, respectively. Likewise, for PVT-EMCAD-B2, the DICE, mIoU, and HD95 scores are improved by 3.45%, 4.44%, and 3.15, respectively. We can also conclude that transfer learning has a comparatively greater impact on the smaller PVT-EMCAD-B0 model than the larger PVT-EMCAD-B2 model. For individual organs, transfer learning significantly boosts the performance of all organ segmentation, except the Gallbladder (GB).

8.4. Effect of deep supervision

We have conducted an ablation study that drops the Deep Supervision (DS). Results of our PVT-EMCAD-B2 on seven datasets are given in Table 10. Our PVT-EMCAD-B2 with DS achieves slightly better DICE scores in six out of seven datasets. Among all the datasets, the DS has the largest impact on the Synapse Multi-organ dataset.

8.5. Effect of input resolutions

Table 11 presents the results of our PVT-EMCAD-B0 and PVT-EMCAD-B2 architectures with different input resolutions. From this table, it is evident that the DICE scores improve with the increase in input resolution. However, these improvements in DICE score come with the increment in #FLOPs. Our PVT-EMCAD-B0 achieves an 85.52% DICE score with only 3.36G FLOPs when using 512 × 512 inputs. On the other hand, our PVT-EMCAD-B2 achieves the best DICE score (86.53%) with 22.39G FLOPs when using 512 × 512 inputs. We also observe that our PVT-EMCAD-B2 with 5.60G FLOPs when using 256 × 256 inputs shows a 1.05% lower DICE score than PVT-EMCAD-B0 with 3.36G FLOPs. Therefore, we can conclude that PVT-EMCAD-B0 is more suitable for larger input resolutions than PVT-EMCAD-B2.