CVPR
SEATTLE, WA JUNE 17-21, 2024

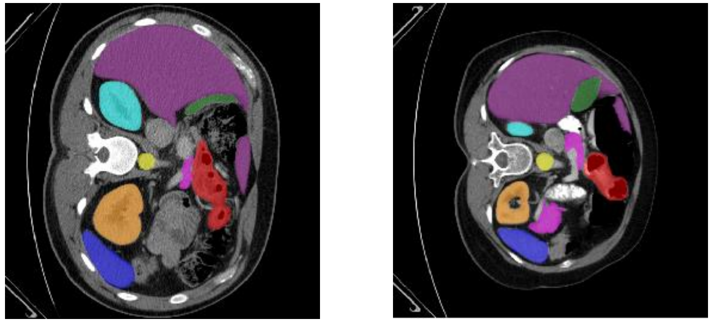
TEXAS
The University of Texas at Austin

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

Md Mostafijur Rahman, Mustafa Munir, and Radu Marculescu
The University of Texas at Austin

Motivation

- Medical image segmentation is a critical step in pre-treatment diagnosis, treatment planning, and post-treatment assessments of various diseases.
- An efficient and effective decoding mechanism is crucial in medical image segmentation, especially in scenarios with limited computational resources.
- However, these decoding mechanisms usually come with high computational costs.



Segmentation mask overlayed on images

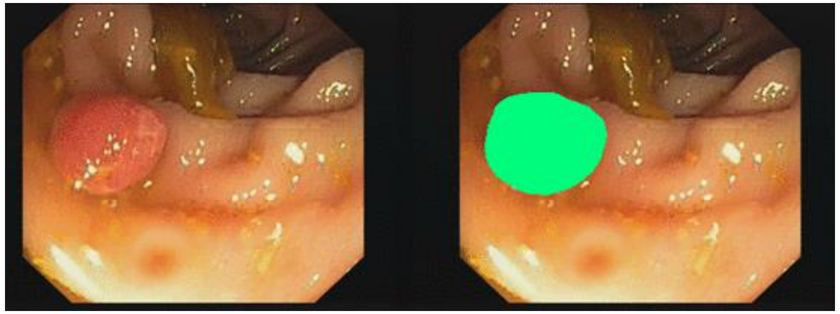
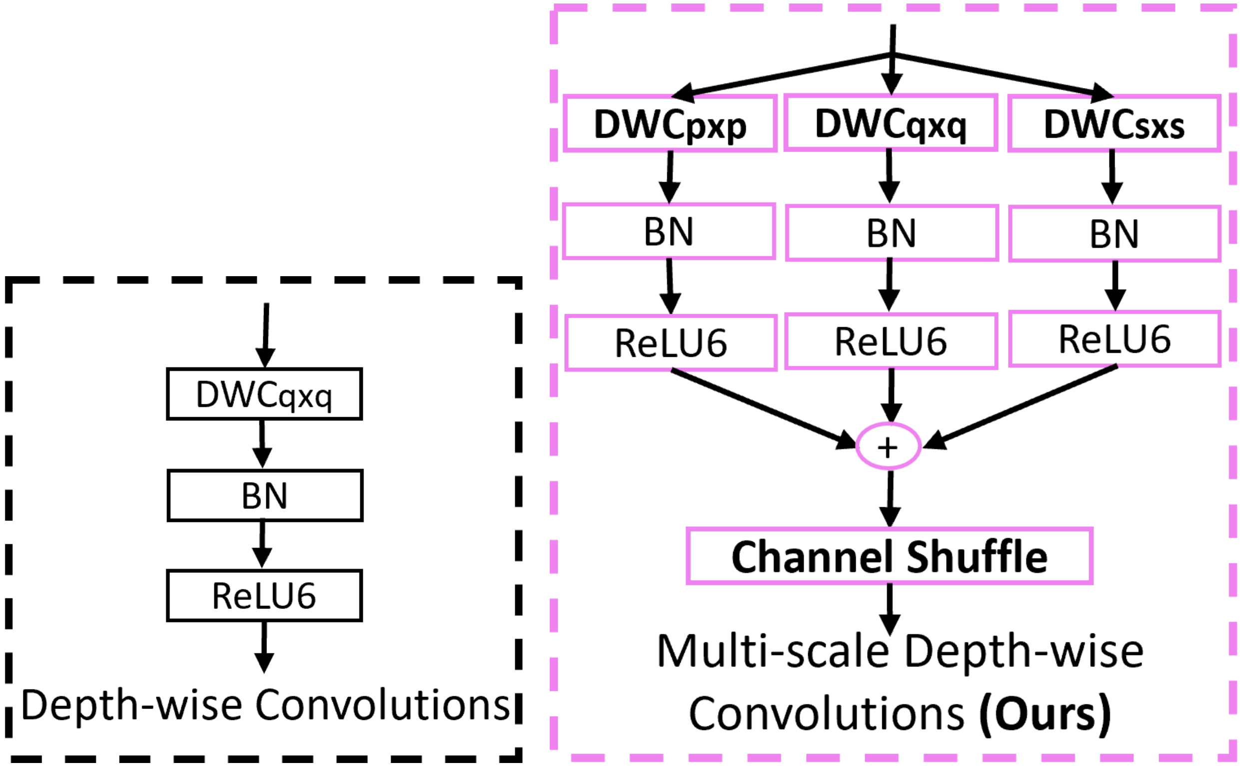


Image Prediction

We introduce EMCAD, a new efficient multi-scale convolutional attention decoder, designed to optimize both performance and computational efficiency.

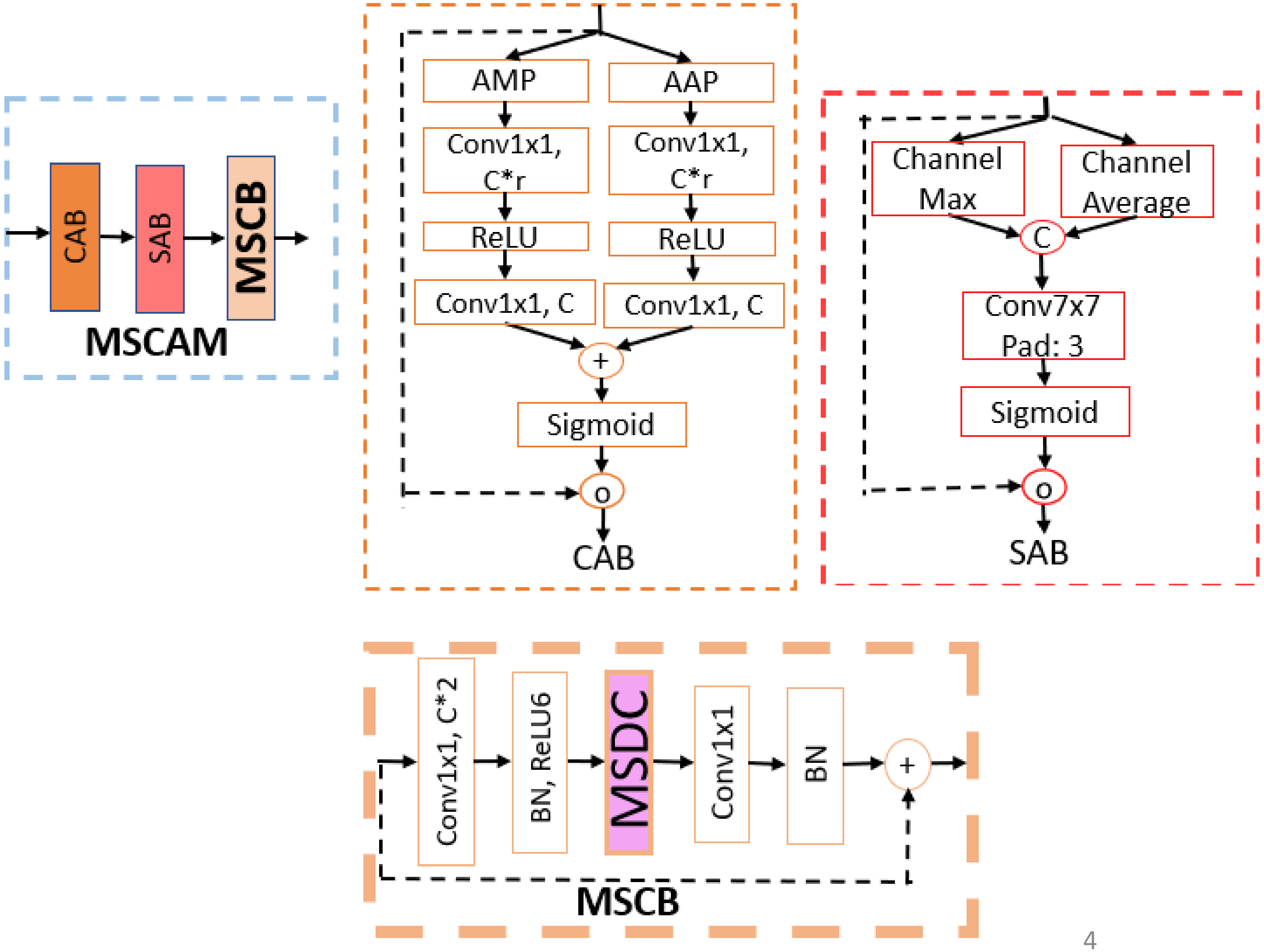
Depth-wise Convolutions vs Our Multi-scale Depth-wise Convolutions

- Basic Depth-wise convolutions apply convolutions in a single scale (qxq).
- Our Multi-scale Depth-wise Convolutions have multiple branches to apply convolutions on multiple scales (e.g., pxp, qxq, sxs) and add the outputs together. We empirically choose (1x1, 3x3, 5x5) kernels for multi-scale depth-wise convolutions in our EMCAD.



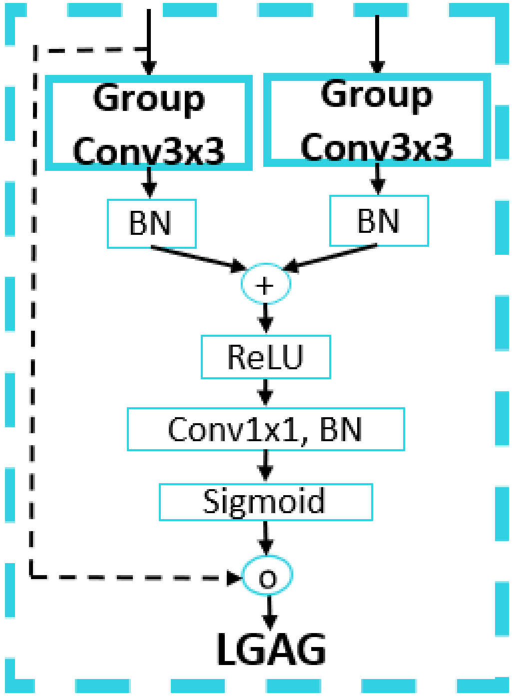
Efficient Multi-scale Convolutional Attention Module (MSCAM)

- Consists of a Channel Attention Block (CAB), a Spatial Attention Block (SAB), and a Multi-scale Convolution Block (MSCB).
- Captures multi-scale salient features by suppressing irrelevant regions.
- Depth-wise convolutions make MSCAM very efficient.



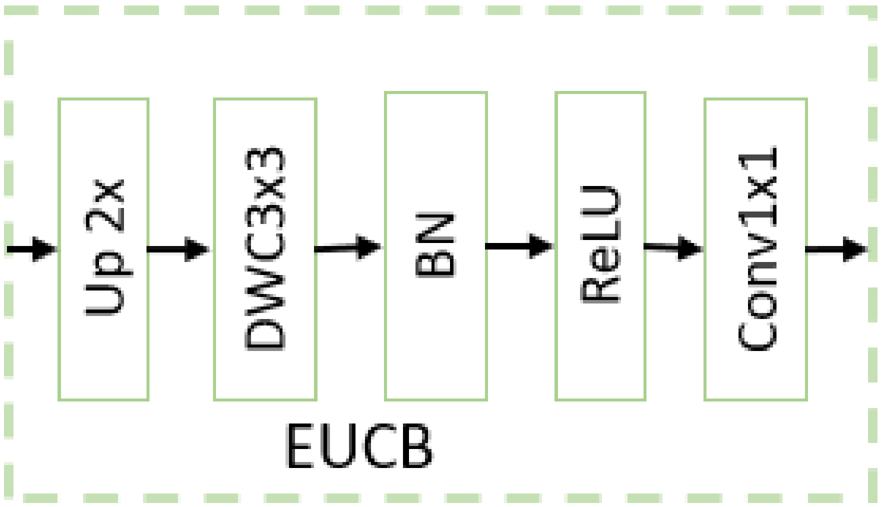
Large-kernel Grouped Attention Gate (LGAG)

- Fuse refined features with the features from skip connections.
- Uses larger kernel (3×3) group convolutions instead of point-wise convolutions.
- Captures salient features in a larger local context with less computation.

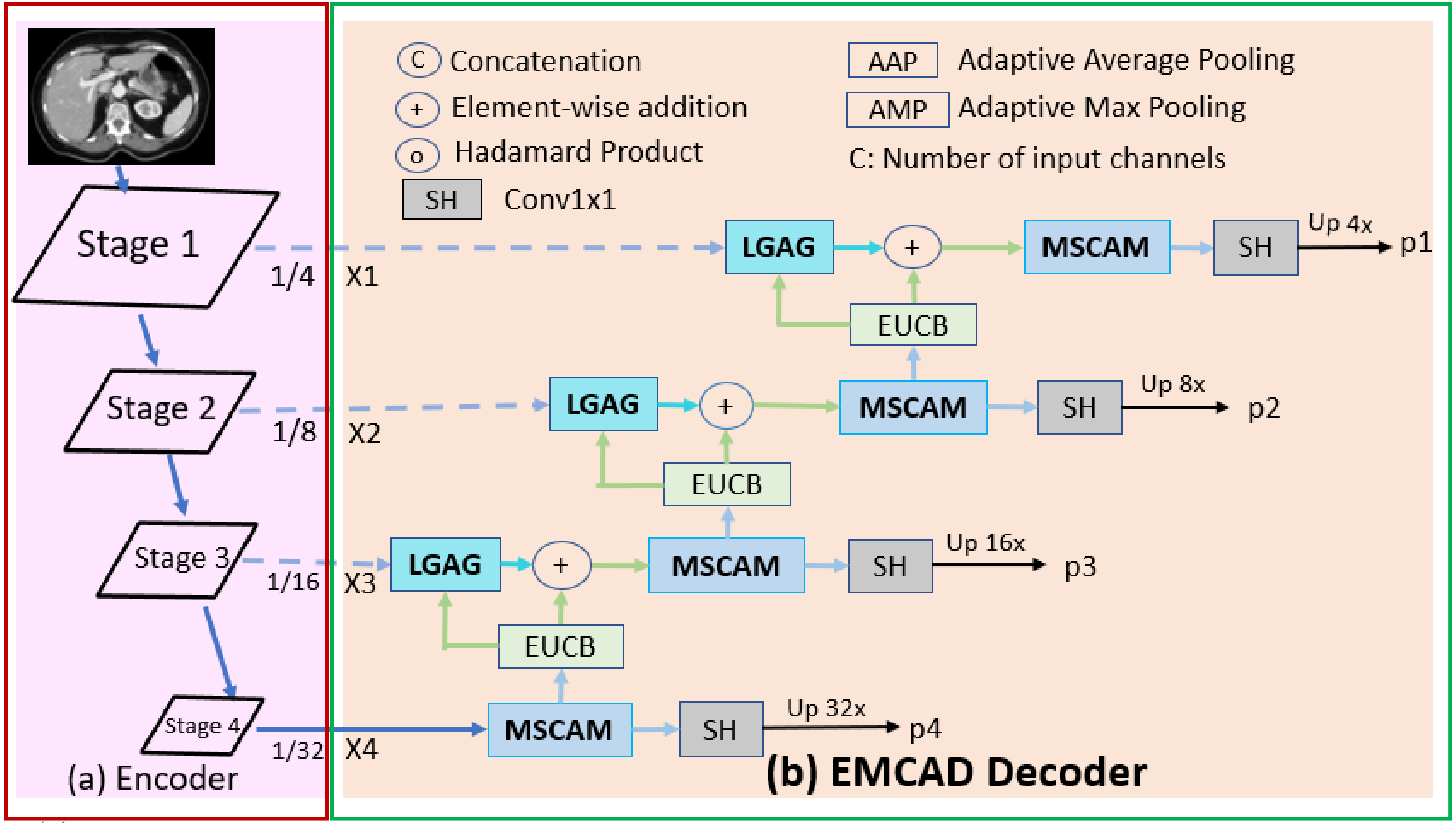


Efficient up-convolution block (EUCB)

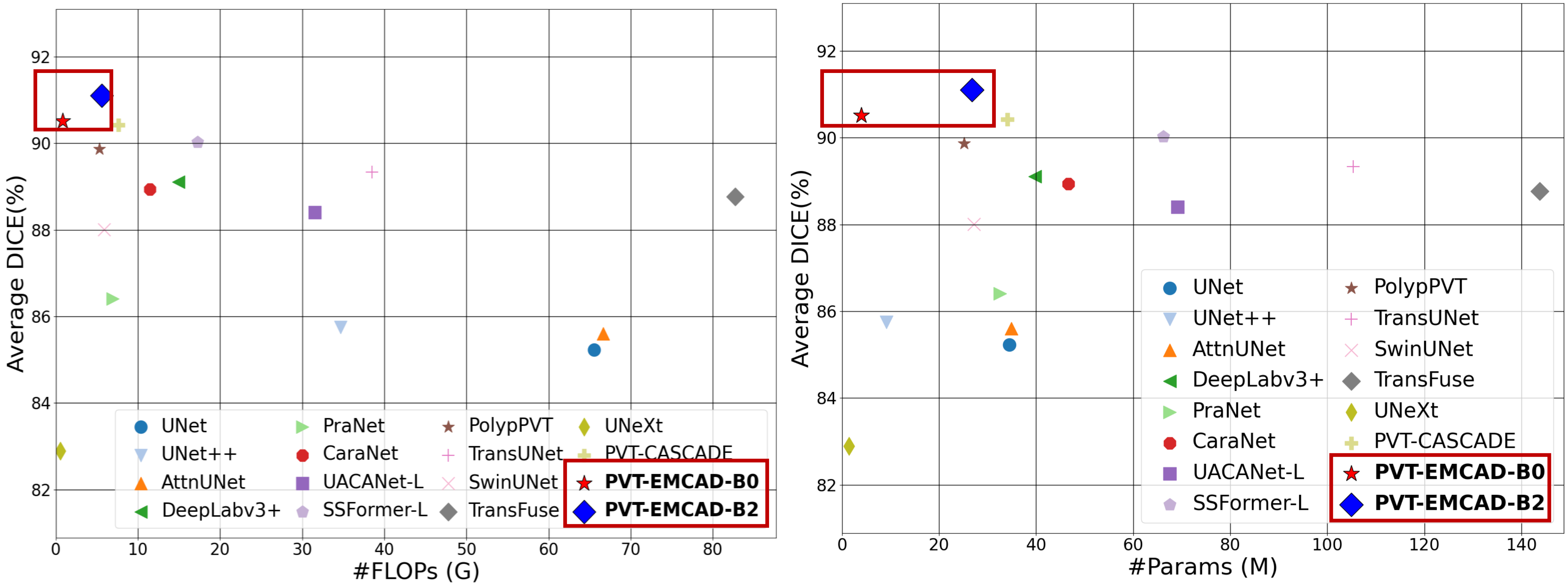
- Uses depth-wise convolutions followed by a point-wise convolution to reduce computational costs.
- Progressively upsamples the feature maps of the current stage to match the dimension and resolution of the feature maps from the next skip connection.



EMCAD Architecture



Experimental Results Summary



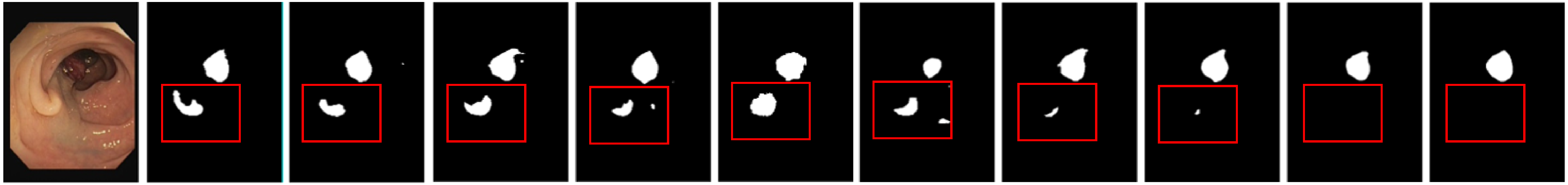
Average DICE scores vs. #FLOPs or #Params for different methods over 10 binary medical image segmentation datasets. As shown, our approaches (PVT-EMCAD-B0 and PVT-EMCAD-B2) have the lowest #FLOPs and #Params, yet the highest DICE scores.

Quantitative Results

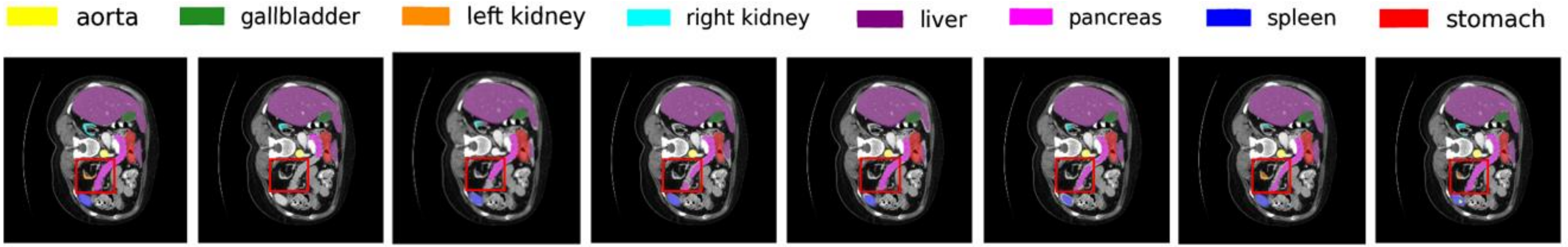
Methods	#Params	#FLOPs	Polyp					Skin Lesion		Cell		BUSI	Avg.
			Clinic	Colon	ETIS	Kvasir	BKAI	ISIC17	ISIC18	DSB18	EM		
UNet	34.53M	65.53G	92.11	83.95	76.85	82.87	85.05	83.07	86.67	92.23	95.46	74.04	85.23
DeepLabv3+	39.76M	14.92G	93.24	91.92	90.73	89.06	89.74	83.84	88.64	92.14	94.96	76.81	89.11
PraNet	32.55M	6.93G	91.71	89.16	83.84	84.82	85.56	83.03	88.56	89.89	92.37	75.14	86.41
PolypPVT	25.11M	5.30G	94.13	91.53	89.93	91.56	91.17	85.56	90.36	90.69	94.40	79.35	89.87
TransUNet	105.32M	38.52G	93.90	91.63	87.79	91.08	89.17	85.00	89.16	92.04	95.27	78.30	89.33
SwinUNet	27.17M	6.2G	92.42	89.27	85.10	89.59	87.61	83.97	89.26	91.03	94.47	77.38	88.01
TransFuse	143.74M	82.71G	93.62	90.35	86.91	90.24	87.47	84.89	89.62	90.85	94.35	79.36	88.77
UNeXt	1.47M	0.57G	90.20	83.84	74.03	77.88	77.93	82.74	87.78	86.01	93.81	74.71	82.89
PVT-CASCADE	34.12M	7.62G	94.53	91.60	91.03	92.05	92.14	85.50	90.41	92.35	95.42	79.21	90.42
PVT-EMCAD-B0 (Ours)	3.92M	0.84G	94.60	91.71	91.65	91.95	91.30	85.67	90.70	92.46	95.35	79.80	90.52
PVT-EMCAD-B2 (Ours)	26.76M	5.6G	**95.21**	**92.31**	**92.29**	**92.75**	**92.96**	**85.95**	**90.96**	**92.74**	95.53	**80.25**	**91.10**

Outperforms closest method by 0.68% with much lower #Params and #FLOPs.

Qualitative Results



(a) Image (b) UNet (c) DeepLabV3+ (d) PolypPVT (e) TransUNet (f) SwinUNet (g) UNeXt (h) PVT-CASCADE (i) PVT-EMCAD-B0 (j) PVT-EMCAD-B2 (k) GroundTruth



(a) Ground Truth (b) TransUNet (c) SwinUNet (d) MISSFormer (e) PVT-CASCADE (f) TransCASCADE (g) PVT-EMCAD-B0 (h) PVT-EMCAD-B2

The segmentation maps generated by our EMCAD have strong similarities with the GroundTruth (GT).

Major Ablation Results

Components			#FLOPs(G)		#Params	Avg
Cascaded	LGAG	MSCAM	224	256	(M)	DICE
No	No	No	0	0	0	80.10±0.2
Yes	No	No	0.100	0.131	0.224	81.08±0.2
Yes	Yes	No	0.108	0.141	0.235	81.92±0.2
Yes	No	Yes	0.373	0.487	1.898	82.86±0.3
Yes	Yes	Yes	0.381	0.498	1.91	**83.63±0.3**

Conv. kernels	[1]	[3]	[5]	[1, 3]	[3, 3]
Synapse	82.43	82.79	82.74	82.98	82.81
ClinicDB	94.81	94.90	94.98	95.13	95.06

Conv. kernels	[1, 3, 5]	[3, 3, 3]	[3, 5, 7]	[1, 3, 5, 7]	[1, 3, 5, 7, 9]
Synapse	**83.63**	82.92	83.11	83.57	83.34
ClinicDB	**95.21**	95.15	95.03	95.18	95.07

Takeaways

- Our multi-scale depth-wise convolutions make EMCAD more efficient and effective (1.91M #Params and 0.498G #FLOPs) compared to SOTA models.
- PVT-EMCAD-B2 outperforms SOTA models on 12 datasets that belong to six different tasks with 79.4% and 80.3% reduction in #Params and #FLOPs, respectively.
- Please read our paper for detailed information and visit https://github.com/SLDGroup/EMCAD for our implementation in Pytorch.

Thank You.