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
Austin, Texas, USA
mostafijur.rahman, mmunir, radum@utexas.edu

Abstract

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. To address this concern, we introduce EMCAD, a new efficient multi-scale convolutional attention decoder, designed to optimize both performance and computational efficiency. EMCAD leverages a unique multi-scale depth-wise convolution block, significantly enhancing feature maps through multi-scale convolutions. EMCAD also employs channel, spatial, and grouped (large-kernel) gated attention mechanisms, which are highly effective at capturing intricate spatial relationships while focusing on salient regions. By employing group and depth-wise convolution, EMCAD is very efficient and scales well (e.g., only 1.91M parameters and 0.381G FLOPs are needed when using a standard encoder). Our rigorous evaluations across 12 datasets that belong to six medical image segmentation tasks reveal that EMCAD achieves state-of-the-art (SOTA) performance with 79.4% and 80.3% reduction in #Params and #FLOPs, respectively. Moreover, EMCAD's adaptability to different encoders and versatility across segmentation tasks further establish EMCAD as a promising tool, advancing the field towards more efficient and accurate medical image analysis. Our implementation is available at https://github.com/SLDGroup/EMCAD.

1. Introduction

In the realm of medical diagnostics and therapeutic strategies, automated segmentation of medical images is vital, as it classifies pixels to identify critical regions such as lesions, tumors, or entire organs. A variety of U-shaped convolutional neural network (CNN) architectures [20, 24, 37, 41, 44, 62], notably UNet [44], UNet++ [62], UNet3+ [24], and nnU-Net [19], have become standard techniques for this purpose, achieving high-quality, high-resolution segmentation output. Attention mechanisms [12, 17, 20, 41, 57] have also been integrated into these models to enhance feature maps and improve pixel-level classification. Although attention-based models have shown improved performance, they still face significant challenges due to the computationally expensive convolutional blocks that are typically used in conjunction with attention mechanisms.

Recently, vision transformers [18] have shown promise in medical image segmentation tasks [5, 8, 17, 42, 43, 52, 54, 61] by capturing long-range dependencies among pixels through Self-attention (SA) mechanisms. Hierarchical vision transformers like Swin [34], PVT [55, 56], MaxViT [49], MERIT [43], ConvFormer [33], and MetaFormer [59] have been introduced to further improve the performance in this field. While the SA excels at capturing global information, it is less adept at understanding the local spatial context [13, 28]. To address this limitation, some approaches have integrated local convolutional attention within the decoders to better grasp spatial details. Nevertheless, these methods can still be computationally demanding because they frequently employ costly convolutional blocks. This limits their applicability to real-world scenarios where computational resources are restricted.

To address the aforementioned limitations, we introduce EMCAD, an efficient multi-scale convolutional attention decoding using a new multi-scale depth-wise convolution block. More precisely, EMCAD enhances the feature maps via efficient multi-scale convolutions, while incorporating complex spatial relationships and local attention through the use of channel, spatial, and grouped (large-kernel) gated attention mechanisms. Our contributions are as follows:

- **New Efficient Multi-scale Convolutional Decoder:** We introduce an efficient multi-scale cascaded fully-convolutional attention decoder (EMCAD) for 2D medical image segmentation; this takes the multi-stage features of vision encoders and progressively enhances the multi-scale and multi-resolution spatial representations. EMCAD has only 0.506M parameters and 0.11G FLOPs for a tiny encoder with #channels = [32, 64, 160, 256],

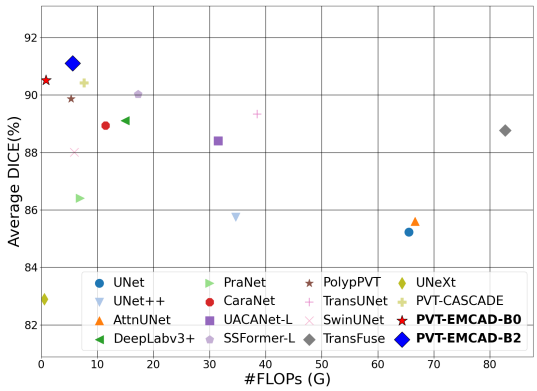

Figure 1. Average DICE scores vs. #FLOPs for different methods over 10 binary medical image segmentation datasets. As shown, our approaches (PVT-EMCAD-B0 and PVT-EMCAD-B2) have the lowest #FLOPs, yet the highest DICE scores.

while it has 1.91M parameters and 0.381G FLOPs for a standard encoder with #channels = [64, 128, 320, 512].

- **Efficient Multi-scale Convolutional Attention Module:** We introduce MSCAM, a new efficient multi-scale convolutional attention module that performs depth-wise convolutions at multiple scales; this refines the feature maps produced by vision encoders and enables capturing multi-scale salient features by suppressing irrelevant regions. The use of depth-wise convolutions makes MSCAM very efficient.
- **Large-kernel Grouped Attention Gate:** We introduce a new grouped attention gate to fuse refined features with the features from skip connections. By using larger kernel (3×3) group convolutions instead of point-wise convolutions in the design, we capture salient features in a larger local context with less computation.
- **Improved Performance:** We empirically show that EMCAD can be used with *any* hierarchical vision encoder (e.g., PVTv2-B0, PVTv2-B2 [56]), while significantly improving the performance of 2D medical image segmentation. EMCAD produces better results than SOTA methods with a significantly lower computational cost (as shown in Figure 1) on 12 medical image segmentation benchmarks that belong to six different tasks.

The remaining of this paper is organized as follows: Section 2 summarizes related work. Section 3 describes the proposed method. Section 4 explains our experimental setup and results on 12 medical image segmentation benchmarks. Section 5 covers different ablation experiments. Lastly, Section 6 concludes the paper.

2. Related Work

2.1. Vision encoders

Convolutional Neural Networks (CNNs) [21–23, 32, 35, 45–48] have been foundational as encoders due to their proficiency in handling spatial relationships in images. More precisely, AlexNet [32] and VGG [46] pave the way, leveraging deep layers of convolutions to extract features progressively. GoogleNet [47] introduces the inception module, allowing more efficient computation of representations across various scales. ResNet [21] introduces residual connections, enabling the training of networks with substantially more layers by addressing the vanishing gradients problem. MobileNets [22, 45] bring CNNs to mobile devices through lightweight, depth-wise separable convolutions. EfficientNet [48] introduces a scalable architectural design to CNNs with compound scaling. Although CNNs are pivotal for many vision applications, they generally lack the ability to capture long-range dependencies within images due to their inherent local receptive fields.

Recently, Vision Transformers (ViTs), pioneered by Dosovitskiy et al. [18], enabled the learning of long-range relationships among pixels using Self-attention (SA). Since then, ViTs have been enhanced by integrating CNN features [49, 56], developing novel self-attention (SA) blocks [34, 49], and introducing new architectural designs [55, 58]. The Swin Transformer [34] incorporates a sliding window attention mechanism, while SegFormer [58] leverages Mix-FFN blocks for hierarchical structures. PVT [55] uses spatial reduction attention, refined in PVTv2 [56] with overlapping patch embedding and a linear complexity attention layer. MaxViT [49] introduces a multi-axis self-attention to form a hierarchical CNN-transformer encoder. Although ViTs address the CNNs limitation in capturing long-range pixel dependencies [21–23, 32, 35, 45–48], they face challenges in capturing the local spatial relationships among pixels. In this paper, we aim to overcome these limitations by introducing a new multi-scale cascaded attention decoder that refines feature maps and incorporates local attention using a multi-scale convolutional attention module.

2.2. Medical image segmentation

Medical image segmentation involves pixel-wise classification to identify various anatomical structures like lesions, tumors, or organs within different imaging modalities such as endoscopy, MRI, or CT scans [8]. U-shaped networks [7, 19, 24, 26, 37, 41, 44, 62] are particularly favored due to their simple but effective encoder-decoder design. The UNet [44] pioneered this approach with its use of skip connections to fuse features at different resolution stages. UNet++ [62] evolves this design by incorporating nested encoder-decoder pathways with dense skip connections. Expanding on these ideas, UNet 3+ [24] introduces comprehensive skip pathways that facilitate full-scale feature integration. Further advancement comes with DC-UNet [37], which integrates a multi-resolution convolution scheme and residual paths into its skip connections. The DeepLab series, including DeepLabv3 [10] and DeepLabv3+ [11], in-

troduce atrous convolutions and spatial pyramid pooling to handle multi-scale information. SegNet [2] uses pooling indices to upsample feature maps, preserving the boundary details. nnU-Net [19] automatically configures hyperparameters based on the specific dataset characteristics, using standard 2D and 3D UNets. Collectively, these U-shaped models have become a benchmark for success in the domain of medical image segmentation.

Recently, vision transformers have emerged as a formidable force in medical image segmentation, harnessing the ability to capture pixel relationships at global scales [5, 8, 17, 42, 43, 52, 58, 61]. TransUNet [8] presents a novel blend of CNNs for local feature extraction and transformers for global context, enhancing both local and global feature capture. Swin-Unet [5] extends this by incorporating Swin Transformer blocks [34] into a U-shaped model for both encoding and decoding processes. Building on these concepts, MERIT [43] introduces a multi-scale hierarchical transformer, which employs SA across different window sizes, thus enhancing the model capacity to capture multi-scale features critical for medical image segmentation.

The integration of attention mechanisms has been investigated within CNNs [20, 41] and transformer-based systems [17] for enhancing medical image segmentation. PraNet [20] employs a reverse attention strategy for feature refinement. PolypPVT [17] leverages PVTv2 [56] as its backbone encoder and incorporates CBAM [57] within its decoding stages. The CASCADE [42] presents a novel cascaded decoder, combining channel [23] and spatial [9] attention to refine features at multiple stages, extracted from a transformer encoder, culminating in high-resolution segmentation outputs. While CASCADE achieves notable performance in segmenting medical images by integrating local and global insights from transformers, it is computationally inefficient due to the use of triple 3×3 convolution layers at each decoder stage. In addition to this, it uses single-scale convolutions during decoding. Our new proposal involves the adoption of multi-scale depth-wise convolutions to mitigate these constraints.

3. Methodology

In this section, we first introduce our new EMCAD decoder and then explain two transformer-based architectures (i.e., PVT-EMCAD-B0 and PVT-EMCAD-B2) incorporating our proposed decoder.

3.1. Efficient multi-scale convolutional attention decoding (EMCAD)

In this section, we introduce our efficient multi-scale convolutional decoding (EMCAD) to process the multi-stage features extracted from pretrained hierarchical vision encoders for high-resolution semantic segmentation. As shown in Figure 2(b), EMCAD consits of efficient multi-scale convolutional attention modules (MSCAMs) to robustly enhance the feature maps, large-kernel grouped attention gates (LGAGs) to refine feature maps fusing with the skip connection via gated attention mechanism, efficient up-convolution blocks (EUCBs) for up-sampling followed by enhancement of feature maps, and segmentation heads (SHs) to produce the segmentation outputs.

More specifically, we use four MSCAMs to refine pyramid features (i.e., $X1$, $X2$, $X3$, $X4$ in Figure 2) extracted from the four stages of the encoder. After each MSCAM, we use an SH to produce a segmentation map of that stage. Subsequently, we upscale the refined feature maps using EUCBs and add them to the outputs from the corresponding LGAGs. Finally, we add four different segmentation maps to produce the final segmentation output. Different modules of our decoder are described next.

3.1.1 Large-kernel grouped attention gate (LGAG)

We introduce a new *large-kernel grouped attention gate* (LGAG) to progressively combine feature maps with attention coefficients, which are learned by the network to allow higher activation of relevant features and suppression of irrelevant ones. This process employs a gating signal derived from higher-level features to control the flow of information across different stages of the network, thus enhancing its precision for medical image segmentation. Unlike Attention UNet [41] which uses 1×1 convolution to process gating signal g (features from skip connections) and input feature map x (upsampled features), in our $q_{att}(.)$ function, we process g and x by applying separate 3×3 group convolutions $GC_g(.)$ and $GC_x(.)$, respectively. These convolved features are then normalized using batch normalization ($BN(.)$) [27] and merged through element-wise addition. The resultant feature map is activated through a ReLU ($R(.)$) layer [39]. Afterward, we apply a 1×1 convolution ($C(.)$) followed by $BN(.)$ layer to get a single channel feature map. We then pass the resultant single-channel feature map through a Sigmoid ($\sigma(.)$) activation function to yield the attention coefficients. The output of this transformation is used to scale the input feature x through element-wise multiplication, producing the attention-gated feature $LGAG(g, x)$. The $LGAG(\cdot)$ (Figure 2(g)) can be formulated as in Equations 1 and 2:

$$q_{att}(g, x) = R(BN(GC_g(g) + BN(GC_x(x)))) \quad (1)$$

$$LGAG(g, x) = x \circledast \sigma(BN(C(q_{att}(g, x)))) \quad (2)$$

Due to using 3×3 kernel group convolutions in $q_{att}(.)$, our LGAG captures comparatively larger spatial contexts with less computational cost.

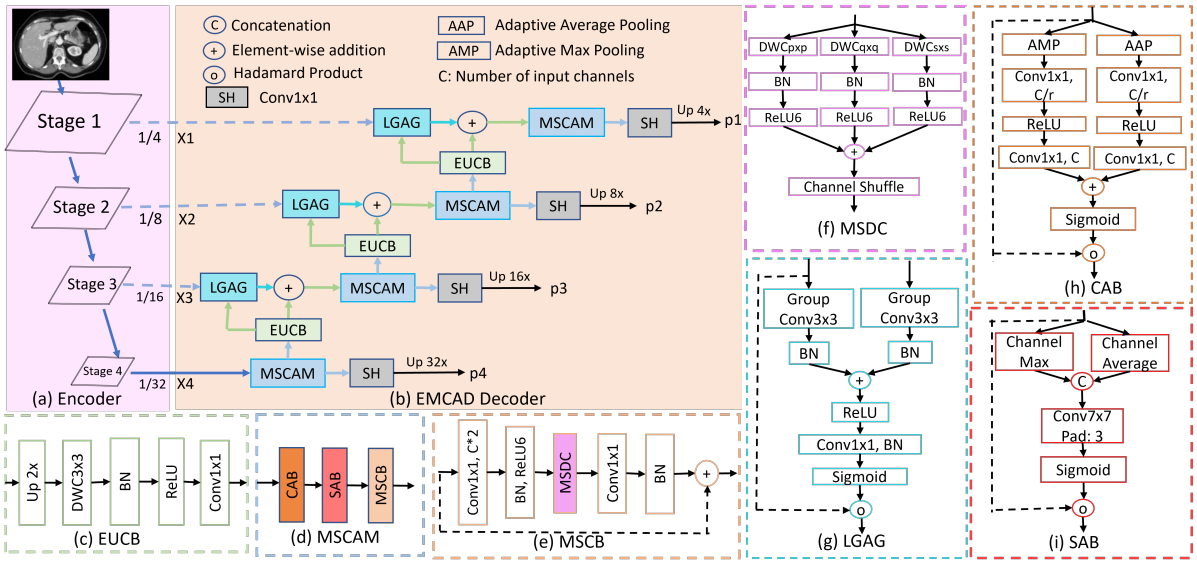

Figure 2. Hierarchical encoder with newly proposed EMCAD decoder architecture. (a) CNN or transformer encoder with four hierarchical stages, (b) EMCAD decoder, (c) Efficient up-convolution block (EUCB), (d) Multi-scale convolutional attention module (MSCAM), (e) Multi-scale convolution block (MSCB), (f) Multi-scale (parallel) depth-wise convolution (MSDC), (g) Large-kernel grouped attention gate (LGAG), (h) Channel attention block (CAB), and (i) Spatial attention block (SAB). X1, X2, X3, and X4 are the features from the four stages of the hierarchical encoder. p1, p2, p3, and p4 are output segmentation maps from four stages of our decoder.

3.1.2 Multi-scale convolutional attention module (MSCAM)

We introduce an efficient multi-scale convolutional attention module to refine the feature maps. MSCAM consists of a channel attention block ($CAB(\cdot)$) to put emphasis on pertinent channels, a spatial attention block [9] ($SAB(\cdot)$) to capture the local contextual information, and an efficient multi-scale convolution block ($MSCB(.)$) to enhance the feature maps preserving contextual relationships. The $MSCAM(.)$ (Figure 2(d)) is given in Equation 3:

$$MSCAM(x) = MSCB(SAB(CAB(x))) \quad (3)$$

where x is the input tensor. Due to using depth-wise convolution in multiple scales, our MSCAM is more effective with significantly lower computational cost than the convolutional attention module (CAM) proposed in [42].

Multi-scale Convolution Block (MSCB): We introduce an efficient multi-scale convolution block to enhance the features generated by our cascaded expanding path. In our MSCB, we follow the design of the inverted residual block (IRB) of MobileNetV2 [45]. However, unlike IRB, our MSCB performs depth-wise convolution at multiple scales and uses channel_shuffle [60] to shuffle channels across groups. More specifically, in our MSCB, we first expand the number of channels (i.e., expansion_factor = 2) using a point-wise (1×1) convolution layers $PWC_1(\cdot)$ followed by a batch normalization layer $BN(\cdot)$ and a ReLU6 [31] activation layer $R6(.)$. We then use a multi-scale depth-wise convolution $MSDC(.)$ to capture both multi-scale and multi-resolution contexts. As depth-wise convolution overlooks the relationships among channels, we use a channel_shuffle operation to incorporate relationships among channels. Afterward, we use another point-wise convolution $PWC_2(.)$ followed by a $BN(.)$ to transform back the original #channels, which also encodes dependency among channels. The $MSCB(\cdot)$ (Figure 2(e)) is formulated as in Equation 4:

$$MSCB(x) = BN(PWC_2(CS(MSDC(R6(BN(PWC_1(x))))))) \quad (4)$$

where parallel $MSDC(.)$ (Figure 2(f)) for different kernel sizes (KS) can be formulated using Equation 5:

$$MSDC(x) = \sum_{ks \in KS} DWCB_{ks}(x) \quad (5)$$

where $DWCB_{ks}(x) = R6(BN(DWC_{ks}(x)))$. Here, $DWC_{ks}(.)$ is a depth-wise convolution with the kernel size ks. $BN(.)$ and $R6(.)$ are batch normalization and ReLU6 activation, respectively. Additionally, our sequential $MSDC(.)$ uses the recursively updated input x, where the input x is residually connected to the previous $DWCB_{ks}(.)$ for better regularization as in Equation 6:

$$x = x + DWCB_{ks}(x) \quad (6)$$

Channel Attention Block (CAB): We use channel attention block to assign different levels of importance to each channel, thus emphasizing more relevant features while suppressing less useful ones. Basically, the CAB identifies *which feature maps* to focus on (and then refine them). Following [57], in CAB, we first apply the adaptive maximum pooling ($P_m(\cdot)$) and adaptive average pooling ($P_a(\cdot)$) to the spatial dimensions (i.e., height and width) to extract the most significant feature of the entire feature map per

channel. Then, for each pooled feature map, we reduce the number of channels $r = 1/16$ times separately using a point-wise convolution ($C_1(\cdot)$) followed by a ReLU activation (R). Afterward, we recover the original channels using another point-wise convolution ($C_2(\cdot)$). We then add both recovered feature maps and apply Sigmoid (σ) activation to estimate attention weights. Finally, we incorporate these weights to input x using the Hadamard product ($\circledast$). The $CAB(\cdot)$ (Figure 2(h)) is defined using Equation 7:

$$CAB(x) = \sigma(C_2(R(C_1(P_m(x)))) + C_2(R(C_1(P_a(x))))) \circledast x \quad (7)$$

Spatial Attention Block (SAB): We use spatial attention to mimic the attentional processes of the human brain by focusing on specific parts of an input image. Basically, the SAB determines *where* to focus in a feature map; then it enhances those features. This process enhances the model's ability to recognize and respond to relevant spatial features, which is crucial for image segmentation where the context and location of objects significantly influence the output. In SAB, we first pool maximum ($Ch_{max}(\cdot)$) and average ($Ch_{avg}(\cdot)$) values along the channel dimension to pay attention to local features. Then, we use a large kernel (i.e., 7×7 as in [17]) convolution layer to enhance local contextual relationships among features. Afterward, we apply the Sigmoid activation (σ) to calculate attention weights. Finally, we feed these weights to the input x (using Hadamard product ($\circledast$) to attend information in a more targeted way. The $SAB(.)$ (Figure 2(i)) is defined using Equation 8:

$$SAB(x) = \sigma(LKC([Ch_{max}(x), Ch_{avg}(x)])) \circledast x \quad (8)$$

3.1.3 Efficient up-convolution block (EUCB)

We use an efficient up-convolution block to progressively upsample the feature maps of the current stage to match the dimension and resolution of the feature maps from the next skip connection. The EUCB first uses an UpSampling $Up(\cdot)$ with scale-factor 2 to upscale the feature maps. Then, it enhances the upscaled feature maps by applying a 3×3 depth-wise convolution $DWC(\cdot)$ followed by a $BN(\cdot)$ and a $ReLU(.)$ activation. Finally, a 1×1 convolution $C_{1\times1}(.)$ is used to reduce the #channels to match with the next stage. The $EUCB(\cdot)$ (Figure 2(c)) is formulated as in Equation 9:

$$EUCB(x) = C_{1\times1}(ReLU(BN(DWC(Up(x))))) \quad (9)$$

Due to using depth-wise convolution instead of 3×3 convolution, our EUCB is *very efficient*.

3.1.4 Segmentation head (SH)

We use segmentation heads to produce the segmentation outputs from the refined feature maps of four stages of the decoder. The SH layer applies a 1×1 convolution $Conv_{1\times1}(\cdot)$ to the refined feature maps having ch_i channels (ch_i is the #channels in the feature map of stage i) and produces output with #channels equal to #classes in target dataset for multi-class but 1 channel for binary segmentation. The $SH(\cdot)$ is formulated as in Equation 10:

$$SH(x) = Conv_{1\times1}(x) \quad (10)$$

3.2. Overall architecture

To show the generalization, effectiveness, and ability to process multi-scale features for medical image segmentation, we integrate our EMCAD decoder alongside tiny (PVTv2-B0) and standard (PVTv2-B2) networks of PVTv2 [56]. However, our decoder is adaptable and seamlessly compatible with other hierarchical backbone networks.

PVTv2 differs from conventional transformer patch embedding modules by applying convolutional operations for consistent spatial information capture. Using PVTv2-b0 (Tiny) and PVTv2-b2 (Standard) encoders [56], we develop the PVT-EMCAD-B0 and PVT-EMCAD-B2 architectures. To adopt PVTv2, 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 EMCAD decoder as shown in Figure 2(a-b). EMCAD then processes them and produces four segmentation maps that correspond to the four stages of the encoder network.

3.3. Multi-stage loss and outputs aggregation

Our EMCAD decoder's four segmentation heads produce four prediction maps p_1, p_2, p_3, and p_4 across its stages.

Loss aggregation: We adopt a combinatorial approach to loss combination called MUTATION, inspired by the work of MERIT [43] for multi-class segmentation. This involves calculating the loss for all possible combinations of predictions derived from 4 heads, totaling $2^4 - 1 = 15$ unique predictions, and then summing these losses. We focus on minimizing this cumulative combinatorial loss during the training process. For binary segmentation, we optimize the additive loss like [42] with an additional term $\mathcal{L}_{p_1+p_2+p_3+p_4}$ as in Equation 11:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{p_1} + \beta\mathcal{L}_{p_2} + \gamma\mathcal{L}_{p_3} + \zeta\mathcal{L}_{p_4} + \delta\mathcal{L}_{p_1+p_2+p_3+p_4} \quad (11)$$

where $\mathcal{L}_{p_1}$, $\mathcal{L}_{p_2}$, $\mathcal{L}_{p_3}$, and $\mathcal{L}_{p_4}$ are the losses of each individual prediction maps. $\alpha = \beta = \gamma = \zeta = \delta = 1.0$ are the weights assigned to each loss.

Output segmentation maps aggregation: We consider the prediction map, p_4, from the last stage of our decoder as the final segmentation map. Then, we obtain the final segmentation output by employing a $Sigmoid$ function for binary or a $Softmax$ function for multi-class segmentation.

4. Experiments

In this section, we present the details of our implementation followed by a comparative analysis of our PVT-EMCAD-B0 and PVT-EMCAD-B2 against SOTA methods. Datasets and evaluation metrics are in Supplementary Section 7.

Methods	#Params	#FLOPs	Polyp					Skin Lesion		Cell		BUSI	Avg.
			Clinic	Colon	ETIS	Kvasir	BKAI	ISIC17	ISIC18	DSB18	EM		
UNet [44]	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
UNet++ [62]	9.16M	34.65G	92.17	87.88	77.40	83.36	84.07	82.98	87.46	91.97	95.48	74.76	85.75
AttnUNet [41]	34.88M	66.64G	92.20	86.46	76.84	83.49	84.07	83.66	87.05	92.22	**95.55**	74.48	85.60
DeepLabv3+ [10]	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 [20]	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
CaraNet [38]	46.64M	11.48G	94.08	91.19	90.25	89.74	89.71	85.02	90.18	89.15	92.78	77.34	88.94
UACANet-L [30]	69.16M	31.51G	94.16	91.02	89.77	90.17	90.35	83.72	89.76	88.86	89.28	76.96	88.41
SSFormer-L [54]	66.22M	17.28G	94.18	92.11	90.16	91.47	91.14	85.28	90.25	92.03	94.95	78.76	90.03
PolypPVT [17]	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 [8]	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 [5]	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 [61]	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 [50]	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 [42]	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**

Table 1. Results of binary medical image segmentation (i.e., polyp, skin lesion, cell, and breast cancer). We reproduce the results of SOTA methods using their publicly available implementation with our train-val-test splits of 80:10:10. #FLOPs of all the methods are reported for 256×256 inputs, except Swin-UNet (224×224). All results are averaged over five runs. Best results are shown in bold.

Architectures	Average			Aorta	GB	KL	KR	Liver	PC	SP	SM
	DICE↑	HD95↓	mIoU↑								
UNet [44]	70.11	44.69	59.39	84.00	56.70	72.41	62.64	86.98	48.73	81.48	67.96
AttnUNet [41]	71.70	34.47	61.38	82.61	61.94	76.07	70.42	87.54	46.70	80.67	67.66
R50+UNet [8]	74.68	36.87	–	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92
R50+AttnUNet [8]	75.57	36.97	–	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
SSFormer [54]	78.01	25.72	67.23	82.78	63.74	80.72	78.11	93.53	61.53	87.07	76.61
PolypPVT [17]	78.08	25.61	67.43	82.34	66.14	81.21	73.78	94.37	59.34	88.05	79.4
TransUNet [8]	77.61	26.9	67.32	86.56	60.43	80.54	78.53	94.33	58.47	87.06	75.00
SwinUNet [5]	77.58	27.32	66.88	81.76	65.95	82.32	79.22	93.73	53.81	88.04	75.79
MT-UNet [53]	78.59	26.59	–	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
MISSFormer [25]	81.96	18.20	–	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
PVT-CASCADE [42]	81.06	20.23	70.88	83.01	**70.59**	82.23	80.37	94.08	64.43	90.1	83.69
TransCASCADE [42]	82.68	17.34	73.48	86.63	68.48	87.66	**84.56**	94.43	65.33	90.79	83.52
PVT-EMCAD-B0 (**Ours**)	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 (**Ours**)	**83.63**	**15.68**	**74.65**	**88.14**	68.87	**88.08**	84.10	**95.26**	**68.51**	92.17	**83.92**

Table 2. Results of abdomen organ segmentation on Synapse Multi-organ dataset. DICE scores are reported for individual organs. Results of UNet, AttnUNet, PolypPVT, SSFormerPVT, TransUNet, and SwinUNet are taken from [42]. ↑ (↓) denotes the higher (lower) the better. '–' means missing data from the source. EMCAD results are averaged over five runs. Best results are shown in bold.

4.1. Implementation details

We implement our network and conduct experiments using Pytorch 1.11.0 on a single NVIDIA RTX A6000 GPU with 48GB of memory. We utilize ImageNet [16] pre-trained PVTv2-b0 and PVTv2-b2 [56] as encoders. In the MSDC of our decoder, we set the multi-scale kernels $[1, 3, 5]$ through an ablation study. We use the parallel arrangement of depth-wise convolutions in all experiments. Our models are trained using the AdamW optimizer [36] with a learning rate and weight decay of $1e-4$. We generally train for 200 epochs with a batch size of 16, except for Synapse multi-organ (300 epochs, batch size 6) and ACDC cardiac organ (400 epochs, batch size 12), saving the best model based on the DICE score. We resize images to 352×352 and use a multi-scale $\{0.75, 1.0, 1.25\}$ training strategy with a gradient clip limit of 0.5 for ClinicDB [3], Kvasir [29], ColonDB [51], ETIS [51], BKAI [40], ISIC17 [15], and ISIC18 [15], while we resize images to 256×256 for BUSI [1], EM [6], and DSB18 [4]. For Synapse and ACDC datasets, images are resized to 224×224, with random rotation and flipping augmentations, optimizing a combined Cross-entropy (0.3) and DICE (0.7) loss. For binary segmentation, we utilize the combined weighted BinaryCrossEntropy (BCE) and weighted IoU loss function.

4.2. Results

We compare our architectures (i.e., PVT-EMCAD-B0 and PVT-EMCAD-B2) with SOTA CNN and transformer-based

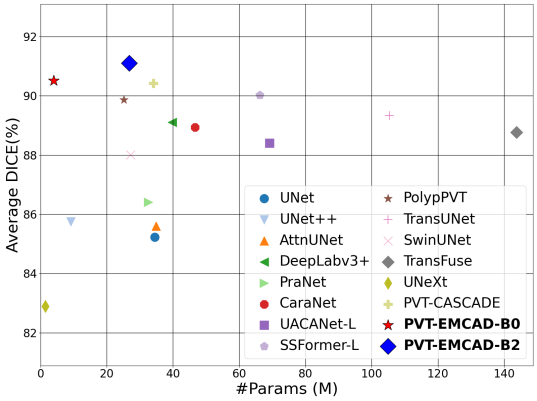

Figure 3. Average DICE scores vs. #Params for different methods over 10 binary medical image segmentation datasets. As shown, our proposed approaches (PVT-EMCAD-B0 and PVT-EMCAD-B2) have the fewest parameters, yet the highest DICE scores.

Methods	Avg. DICE	RV	Myo	LV
R50+UNet [8]	87.55	87.10	80.63	94.92
R50+AttnUNet [8]	86.75	87.58	79.20	93.47
ViT+CUP [8]	81.45	81.46	70.71	92.18
R50+ViT+CUP [8]	87.57	86.07	81.88	94.75
TransUNet [8]	89.71	86.67	87.27	95.18
SwinUNet [5]	88.07	85.77	84.42	94.03
MT-UNet [53]	90.43	86.64	89.04	95.62
MISSFormer [25]	90.86	89.55	88.04	94.99
PVT-CASCADE [42]	91.46	89.97	88.9	95.50
TransCASCADE [42]	91.63	90.25	89.14	95.50
Cascaded MERIT [43]	91.85	90.23	89.53	95.80
PVT-EMCAD-B0 (**Ours**)	91.34±0.2	89.37	88.99	95.65
PVT-EMCAD-B2 (**Ours**)	**92.12**±0.2	**90.65**	**89.68**	**96.02**

Table 3. Results of cardiac organ segmentation on ACDC dataset. DICE scores (%) are reported for individual organs. We get the results of SwinUNet from [42]. Best results are shown in bold.

segmentation methods on 12 datasets that belong to six medical image segmentation tasks. Qualitative results are in the Supplementary Section 7.3.

4.2.1 Results of binary medical image segmentation

Results for different methods on 10 binary medical image segmentation datasets are shown in Table 1 and Figure 1. Our PVT-EMCAD-B2 attains the highest average DICE score (91.10%) with only 26.76M parameters and 5.6G FLOPs. The multi-scale depth-wise convolution in our EMCAD decoder, combined with the transformer encoder, contributes to these performance gains.

Polyp segmentation: Table 1 reveals that our PVT-EMCAD-B2 surpasses all SOTA methods in five polyp segmentation datasets. PVT-EMCAD-B2 achieves DICE score improvements of 1.08%, 0.78%, 2.36%, 1.19%, and 1.79% over PolypPVT in ClinicDB, ColonDB, ETIS, Kvasir, and BKAI-IGI, despite having slightly more parameters and FLOPs. The smallest model UNeXt, exhibits the worst performance in all five polyp segmentation datasets. Our smaller model with only 3.92M parameters and 0.84G FLOPs also outperforms all the methods except PVT-CASCADE (in Kvasir and BKAI-IGH) and SSFormer-L (in ColonDB), which achieve the best performance among SOTA methods. In conclusion, our PVT-EMCAD-B2 achieves the new SOTA results in these five polyp segmentation datasets.

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**

Table 4. Effect of different components of EMCAD with PVTv2-b2 encoder on Synapse multi-organ dataset. #FLOPs are reported for input resolution of 224×224 and 256×256. All results are averaged over five runs. Best results are shown in bold.

Skin lesion segmentation: Table 1 shows PVT-EMCAD-B2's strong performance on ISIC17 and ISIC18 skin lesion segmentation datasets, achieving DICE scores of 85.95% and 90.96%, surpassing DeepLabV3+ by 2.11% and 2.32%. It also beats the nearest method PVT-CASCADE by 0.45% and 0.55% in ISIC17 and ISIC18, respectively, though our decoder is significantly more efficient than CASCADE. Our PVT-EMCAD-B0 also shows huge potential in point care applications like skin lesion segmentation with only 3.92M parameters and 0.84G FLOPs.

Cell segmentation: To evaluate our method's effectiveness in biological imaging, we use DSB18 [4] for cell nuclei and EM [6] for cell structure segmentation. As Table 1 indicates, our PVT-EMCAD-B2 sets a SOTA benchmark in cell nuclei segmentation on DSB18, outperforming DeepLabv3+, TransFuse, and PVT-CASCADE. On the EM dataset, PVT-EMCAD-B2 secures the second-best DICE score (95.53%), offering significantly lower computational costs than the top-performing AttnUNet (95.55%).

Breast cancer segmentation: We conduct experiments on the BUSI dataset for breast cancer segmentation in ultrasound images. Our PVT-EMCAD-B2 achieves the SOTA DICE score (80.25%) on this dataset. Furthermore, our PVT-EMCAD-B0 outperforms the computationally similar method UNeXt by a notable margin of 5.54%.

4.2.2 Results of abdomen organ segmentation

Table 2 shows that our PVT-EMCAD-B2 excels in abdomen organ segmentation on the Synapse multi-organ dataset, achieving the highest average DICE score of 83.63% and surpassing all SOTA CNN- and transformer-based methods. It outperforms PVT-CASCADE by 2.57% in DICE score and 4.55 in HD95 distance, indicating superior organ boundary location. Our EMCAD decoder boosts individ-

Conv. kernels	[1]	[3]	[5]	[1, 3]	[3, 3]	[1, 3, 5]	[3, 3, 3]	[3, 5, 7]	[1, 3, 5, 7]	[1, 3, 5, 7, 9]
Synapse	82.43	82.79	82.74	82.98	82.81	**83.63**	82.92	83.11	83.57	83.34
ClinicDB	94.81	94.90	94.98	95.13	95.06	**95.21**	95.15	95.03	95.18	95.07

Table 5. Effect of multi-scale kernels in the depth-wise convolution of MSDC on ClinicDB and Synapse multi-organ datasets. We use the PVTv2-b2 encoder for these experiments. All results are averaged over five runs. Best results are highlighted in bold.

Encoders	Decoders	#FLOPs(G)	#Params(M)	DICE (%)
PVTv2-B0	CASCADE	0.439	2.32	80.54
PVTv2-B0	EMCAD (**Ours**)	**0.110**	**0.507**	**81.97**
PVTv2-B2	CASCADE	1.93	9.27	82.78
PVTv2-B2	EMCAD (**Ours**)	**0.381**	**1.91**	**83.63**

Table 6. Comparison with the baseline decoder on Synapse Multi-organ dataset. We only report the #FLOPs (with input resolution of 224×224) and the #parameters of the decoders. All the results are averaged over five runs. Best results are shown in bold.

ual organ segmentation, significantly outperforming SOTA methods on six of eight organs.

4.2.3 Results of cardiac organ segmentation

Table 3 shows the DICE scores of our PVT-EMCAD-B2 and PVT-EMCAD-B0 along with other SOTA methods, on the MRI images of the ACDC dataset for cardiac organ segmentation. Our PVT-EMCAD-B2 achieves the highest average DICE score of 92.12%, thus improving about 0.27% over Cascaded MERIT though our network has significantly lower computational cost. Besides, PVT-EMCAD-B2 has better DICE scores in all three organ segmentations.

5. Ablation Studies

In this section, we conduct ablation studies to explore different aspects of our architectures and the experimental framework. More ablations are in Supplementary Section 8.

5.1. Effect of different components of EMCAD

We conduct a set of experiments on the Synapse multi-organ dataset to understand the effect of different components of our EMCAD decoder. We start with only the encoder and add different modules such as Cascaded structure, LGAG, and MSCAM to understand their effect. Table 4 exhibits that the cascaded structure of the decoder helps to improve performance over the non-cascaded one. The incorporation of LGAG and MSCAM improves performance, however, MSCAM proves to be more effective. When both the LGAG and MSCAM modules are used together, it produces the best DICE score of 83.63%. It is also evident that there is about 3.53% improvement in the DICE score with an additional 0.381G FLOPs and 1.91M parameters.

5.2. Effect of multi-scale kernels in MSCAM

We have conducted another set of experiments on Synapse multi-organ and ClinicDB datasets to understand the effect of different multi-scale kernels used for depth-wise convolutions in MSDC. Table 5 reports these results which show that performance improves from 1×1 to 3×3 kernel. When 1×1 kernel is used together with 3×3 it improves more than when using them alone. However, when two 3×3 kernels are used together, performance drops. The incorporation of a 5×5 kernel with 1×1 and 3×3 kernels further improves the performance and it achieves the best results in both Synapse multi-organ and ClinicDB datasets. If we add additional larger kernels (e.g., 7×7, 9×9), the performance of both datasets drops. Based on these empirical observations, we choose $[1, 3, 5]$ kernels in all our experiments.

5.3. Comparison with the baseline decoder

In Table 6, we report the experimental results with the computational complexity of our EMCAD decoder and a baseline decoder, namely CASCADE. From Table 6, we can see that our EMCAD decoder with PVTv2-b2 requires 80.3% fewer FLOPs and 79.4% fewer parameters to outperform (by 0.85%) the respective CASCADE decoder. Similarly, our EMCAD decoder with PVTv2-B0 achieves 1.43% better DICE score than the CASCADE decoder with 78.1% fewer parameters and 74.9% fewer FLOPs.

6. Conclusions

In this paper, we have presented EMCAD, a new and efficient multi-scale convolutional attention decoder designed for multi-stage feature aggregation and refinement in medical image segmentation. EMCAD employs a multi-scale depth-wise convolution block, which is key for capturing diverse scale information within feature maps, a critical factor for precision in medical image segmentation. This design choice, using depth-wise convolutions instead of standard 3×3 convolution blocks, makes EMCAD notably efficient.

Our experiments reveal that EMCAD surpasses the recent CASCADE decoder in DICE scores with 79.4% fewer parameters and 80.3% less FLOPs. Our extensive experiments also confirm EMCAD's superior performance compared to SOTA methods across 12 public datasets covering six different 2D medical image segmentation tasks. EMCAD's compatibility with smaller encoders makes it an excellent fit for point-of-care applications while maintaining high performance. We anticipate that our EMCAD decoder will be a valuable asset in enhancing a variety of medical image segmentation and semantic segmentation tasks.

Acknowledgements: This work is supported in part by the NSF grant CNS 2007284, and in part by the iMAGiNE Consortium (https://imagine.utexas.edu/).

References

[1] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020. 6, 1

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(12):2481–2495, 2017. 3

[3] 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. *Comput. Med. Imaging Graph.*, 43:99–111, 2015. 6, 1

[4] Juan C Caicedo, Allen Goodman, Kyle W Karhohs, Beth A Cimini, Jeanelle Ackerman, Marzieh Haghighi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, et al. Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nature methods*, 16(12):1247–1253, 2019. 6, 7, 1

[5] 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. 1, 3, 6, 7

[6] Albert Cardona, Stephan Saalfeld, Stephan Preibisch, Benjamin Schmid, Anchi Cheng, Jim Pulokas, Pavel Tomancak, and Volker Hartenstein. An integrated micro-and macroarchitectural analysis of the drosophila brain by computer-assisted serial section electron microscopy. *PLoS biology*, 8(10):e1000502, 2010. 6, 7, 1

[7] Gongping Chen, Lei Li, Yu Dai, Jianxun Zhang, and Moi Hoon Yap. Aau-net: an adaptive attention u-net for breast lesions segmentation in ultrasound images. *IEEE Trans. Med. Imaging*, 2022. 2

[8] 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. 1, 2, 3, 6, 7

[9] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5659–5667, 2017. 3, 4

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2017. 2, 6

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Eur. Conf. Comput. Vis.*, pages 801–818, 2018. 2

[12] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *Eur. Conf. Comput. Vis.*, pages 234–250, 2018. 1

[13] 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. 1

[14] 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. 1

[15] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *IEEE Int. Symp. Biomed. Imaging*, pages 168–172. IEEE, 2018. 6, 1

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255. Ieee, 2009. 6

[17] 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. 1, 3, 5, 6

[18] 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. 1, 2

[19] Isensee et al. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 1, 2, 3

[20] 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 *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pages 263–273. Springer, 2020. 1, 3, 6

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 2

[22] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 2

[23] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7132–7141, 2018. 2, 3

[24] 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 *ICASSP*, pages 1055–1059. IEEE, 2020. 1, 2

[25] Xiaohong Huang, Zhifang Deng, Dandan Li, and Xueguang Yuan. Missformer: An effective medical image segmentation transformer. *arXiv preprint arXiv:2109.07162*, 2021. 6, 7

[26] Nabil Ibtehaz and Daisuke Kihara. Acc-unet: A completely convolutional unet model for the 2020s. In *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pages 692–702. Springer, 2023. 2

[27] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Int. Conf. Mach. Learn.*, pages 448–456. pmlr, 2015. 3

[28] Md Amirul Islam, Sen Jia, and Neil DB Bruce. How much position information do convolutional neural networks encode? *arXiv preprint arXiv:2001.08248*, 2020. 1

[29] 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 *Int. Conf. Multimedia Model.*, pages 451–462. Springer, 2020. 6, 1

[30] Taehun Kim, Hyemin Lee, and Daijin Kim. Uacanet: Uncertainty augmented context attention for polyp segmentation. In *ACM Int. Conf. Multimedia*, pages 2167–2175, 2021. 6

[31] Alex Krizhevsky and Geoff Hinton. Convolutional deep belief networks on cifar-10. *Unpublished manuscript*, 40(7): 1–9, 2010. 4

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inform. Process. Syst.*, 25, 2012. 2

[33] Xian Lin, Zengqiang Yan, Xianbo Deng, Chuansheng Zheng, and Li Yu. Convformer: Plug-and-play cnn-style transformers for improving medical image segmentation. In *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pages 642–651. Springer, 2023. 1

[34] 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 *Int. Conf. Comput. Vis.*, pages 10012–10022, 2021. 1, 2, 3

[35] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11976–11986, 2022. 2

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[37] Ange Lou, Shuyue Guan, and Murray Loew. Dc-unet: rethinking the u-net architecture with dual channel efficient cnn for medical image segmentation. In *Med. Imaging 2021: Image Process.*, pages 758–768. SPIE, 2021. 1, 2

[38] Ange Lou, Shuyue Guan, Hanseok Ko, and Murray H Loew. Caranet: context axial reverse attention network for segmentation of small medical objects. In *Med. Imaging 2022: Image Process.*, pages 81–92. SPIE, 2022. 6

[39] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Int. Conf. Mach. Learn.*, pages 807–814, 2010. 3

[40] Phan Ngoc Lan, Nguyen Sy An, Dao Viet Hang, Dao Van Long, Tran Quang Trung, Nguyen Thi Thuy, and Dinh Viet Sang. Neounet: Towards accurate colon polyp segmentation and neoplasm detection. In *Adv. Vis. Comput. – Int. Symp.*, pages 15–28. Springer, 2021. 6, 1

[41] 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. 1, 2, 3, 6

[42] Md Mostafijur Rahman and Radu Marculescu. Medical image segmentation via cascaded attention decoding. In *IEEE/CVF Winter Conf. Appl. Comput. Vis.*, pages 6222–6231, 2023. 1, 3, 4, 5, 6, 7

[43] Md Mostafijur Rahman and Radu Marculescu. Multi-scale hierarchical vision transformer with cascaded attention decoding for medical image segmentation. In *Med. Imaging Deep Learn.*, 2023. 1, 3, 5, 7

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pages 234–241. Springer, 2015. 1, 2, 6

[45] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4510–4520, 2018. 2, 4

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[47] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1–9, 2015. 2

[48] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Int. Conf. Mach. Learn.*, pages 6105–6114. PMLR, 2019. 2

[49] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *Eur. Conf. Comput. Vis.*, pages 459–479. Springer, 2022. 1, 2

[50] Jeya Maria Jose Valanarasu and Vishal M Patel. Unext: Mlp-based rapid medical image segmentation network. In *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pages 23–33. Springer, 2022. 6, 1

[51] 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. *J. Healthc. Eng.*, 2017, 2017. 6, 1

[52] 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 *AAAI*, pages 2441–2449, 2022. 1, 3

[53] 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 *ICASSP*, pages 2390–2394. IEEE, 2022. 6, 7

[54] 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. 1, 6

[55] 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 *Int. Conf. Comput. Vis.*, pages 568–578, 2021. 1, 2

[56] 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. *Comput. Vis. Media*, 8(3):415–424, 2022. 1, 2, 3, 5, 6

[57] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Eur. Conf. Comput. Vis.*, pages 3–19, 2018. 1, 3, 4

[58] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inform. Process. Syst.*, 34:12077–12090, 2021. 2, 3

[59] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10819–10829, 2022. 1

[60] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6848–6856, 2018. 4

[61] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pages 14–24. Springer, 2021. 1, 3, 6

[62] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, pages 3–11. Springer, 2018. 1, 2, 6