

# MK-UNet: Multi-kernel Lightweight CNN for Medical Image Segmentation

Md Mostafijur Rahman  
 The University of Texas at Austin  
 Austin, Texas  
 mostafijur.rahman@utexas.edu

Radu Marculescu  
 The University of Texas at Austin  
 Austin, Texas  
 radum@utexas.edu

## Abstract

*In this paper, we introduce MK-UNet, a paradigm shift towards ultra-lightweight, multi-kernel U-shaped CNNs tailored for medical image segmentation. Central to MK-UNet is the multi-kernel depth-wise convolution block (MKDC) we design to adeptly process images through multiple kernels, while capturing complex multi-resolution spatial relationships. MK-UNet also emphasizes the images salient features through sophisticated attention mechanisms, including channel, spatial, and grouped gated attention. Our MK-UNet network, with a modest computational footprint of only 0.316M parameters and 0.314G FLOPs, represents not only a remarkably lightweight, but also significantly improved segmentation solution that provides higher accuracy over state-of-the-art (SOTA) methods across six binary medical imaging benchmarks. Specifically, MK-UNet outperforms TransUNet in DICE score with nearly  $333\times$  and  $123\times$  fewer parameters and FLOPs, respectively. Similarly, when compared against UNeXt, MK-UNet exhibits superior segmentation performance, improving the DICE score up to 6.7% margins while operating with  $4.7\times$  fewer #Params. Our MK-UNet also outperforms other recent lightweight networks, such as MedT, CMUNeXt, EGE-UNet, and Rolling-UNet, with much lower computational resources. This leap in performance, coupled with drastic computational gains, positions MK-UNet as an unparalleled solution for real-time, high-fidelity medical diagnostics in resource-limited settings, such as point-of-care devices. Our implementation is available at <https://github.com/SLDGroup/MK-UNet>.*

## 1. Introduction

The field of medical image segmentation has experienced transformative growth through the development of U-shaped convolutional neural network (CNN) architectures [12, 26, 31, 44] such as UNet [31], UNet++ [44], AttentionUNet [26], PraNet [12], UACANet [18], and DeepLabv3+ [8]. These models excel at segmenting medical images, en-

abling precise segmentation of critical tumors, lesions, or polyps. Attention mechanisms [12, 26, 41] integrated into these architectures help refine feature maps, thus enhancing pixel-level classification. However, the substantial computational demands of these models, including those with attention mechanisms, limit their applicability in resource-constrained environments such as point-of-care diagnostics.

The introduction of vision transformers [4, 6, 27–30, 39], including TransUNet [6], SwinUNet [4] and MedT [39], marked a shift towards leveraging self-attention to capture long-range dependencies within images for a comprehensive global view. However, transformers tend to neglect crucial local spatial relationships among pixels which are essential for precise segmentation. Moreover, transformers usually have high memory and computational demands for calculation and fusing attention with convolutional mechanisms, which limits their practical deployment.

In recent years, a good number of lightweight architectures such as UNeXt [38], CMUNeXt [37], MALUNet [32], EGE-UNet [33], and Rolling-UNet [21], bridge this gap by combining the strengths of CNNs and multi-layer perceptron (MLP). However, most of these architectures are designed for less complex or easy-to-segment applications such as skin lesions, breast cancer in ultrasound, and microscopic cell nuclei/structure segmentation. Consequently, these architectures show poor performance in challenging applications like polyp segmentation due to the high variability in the shape, size, and texture of polyps.

To address these computational and precision challenges, we introduce MK-UNet, a significant breakthrough in image segmentation, which leverages the benefits of a multi-kernel perspective (i.e.,  $k_1 = k_2$  or  $k_1 \neq k_2$  for  $k_1, k_2 \in \text{Kernels}$ ) and depth-wise convolutions to address the computational complexity and challenges inherent in existing CNN- and transformer-based models. Depth-wise convolutions drastically reduce the computational load, making the network more efficient without sacrificing the ability to capture detailed features within the image. Additionally, our multi-kernel property enables the model to effectively handle feature representations at *same*

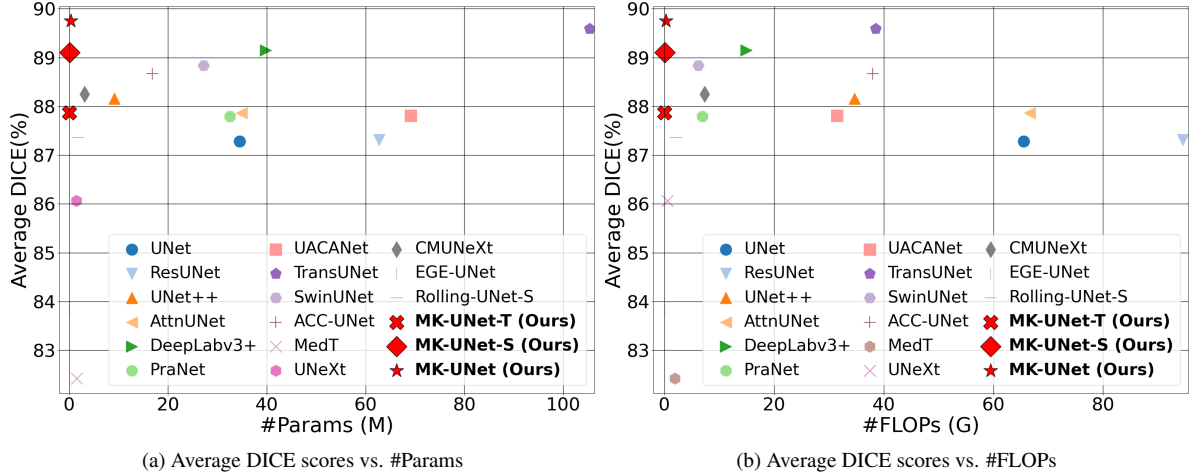


Figure 1. Comparison of MK-UNet against different SOTA methods over six binary medical image segmentation datasets. As shown, our MK-UNet has the second lowest #Params and #FLOPs (behind EGE-UNet [33]), yet the highest DICE scores. Though EGE-UNet has the lowest #Params and #FLOPs among existing methods, it has the lowest DICE score as well.

or varying receptive fields, thus allowing for a more robust and comprehensive analysis of complex images in diverse applications. Moreover, by incorporating sophisticated convolutional multi-focal (channel and spatial) attention mechanisms *only* in our decoder further refines the feature maps by capturing the image salient features. We note that our network is effective for segmentation in both scenarios, whether the regions of interest vary significantly in size and shape or remain relatively uniform. By integrating these new ideas, MK-UNet achieves a fine balance between computational efficiency and segmentation accuracy, thus offering an ultra-lightweight model that not only surpasses the performance of heavyweight counterparts (in DICE scores), but it does so with significantly fewer #Params and #FLOPs. Our contributions are as follows:

- **New Lightweight Multi-kernel UNet:** We propose a new end-to-end U-shaped network, MK-UNet, for medical image segmentation, which encodes an image using lightweight multi-kernel convolutions to capture multi-resolution spatial representations. MK-UNet also progressively refines the multi-resolution spatial representations using multi-kernel convolutional attention. Of note, our MK-UNet-T (tiny) has only 0.027M and 0.062G #Params and #FLOPs, respectively, yet provides SOTA performance. Moreover, MK-UNet (standard) has only 0.316M #Params and 0.314G #FLOPs. The extremely low model size (#Params) and computations (#FLOPs) make our MK-UNet easy to deploy in point-of-care diagnostics or resource-constraint environments (e.g., mobile or edge devices).
- **Lightweight Multi-kernel Inverted Residual:** We introduce MKIR, a new Multi-Kernel Inverted Residual block that performs depth-wise convolutions with multiple ker-

nels (i.e.,  $k_1 = k_2$  or  $k_1 \neq k_2$ , for  $k_1, k_2 \in \text{Kernels}$ ). Our encoder extracts features using the MKIR block; this choice is motivated by the need to efficiently process and encode diverse and complex structures in medical images, thus providing a rich representation with minimal computational costs.

- **Lightweight Multi-kernel Inverted Residual Attention:** We propose Multi-Kernel Inverted Residual Attention (MKIRA), a new block to refine and enhance multi-scale salient features by suppressing irrelevant regions. In our decoder, MKIRA enhances features discrimination by focusing on key feature channels and highlighting the important spatial regions in an image. This ensures that the decoder can reconstruct precise and accurate segmentation maps by focusing only on the most critical aspects of the encoded features.
- **Improved Performance across Various Tasks and Benchmarks:** We empirically show that MK-UNet significantly improves the performance of medical image segmentation compared to SOTA methods with a significantly lower computational cost (as shown in Fig. 1) on six binary medical image segmentation benchmarks that belong to four different tasks.

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

## 2. Related Work

### 2.1. Convolutional Neural Networks (CNNs)

The advent of CNNs marks a significant shift in medical image segmentation [8, 12, 18, 26, 31, 44]. Pioneering work

such as Fully Convolutional Networks (FCNs) [23] laid the foundation for end-to-end segmentation models. FCNs replace fully connected layers with convolutional layers, thus enabling pixel-wise predictions and efficient learning of spatial hierarchies in images. U-Net [31] became a cornerstone in medical image segmentation due to its encoder-decoder architecture with skip connections. This design effectively combines high-resolution features from the encoder with context information from the decoder, hence leading to precise segmentation even with limited training data. The sophisticated design of U-shaped architecture for pixel-level segmentation tasks motivates us to choose the U-shaped design in our proposed network.

U-Net’s success has inspired numerous variants and improvements. Inspired by residual learning in ResNet [13], ResUNet [43] uses residual blocks to facilitate gradient flow and improve convergence, addressing the vanishing gradient problem in deep networks. Zhou et al. [44] introduce UNet++, which uses dense nested skip connections to further enhance the feature propagation and improve the segmentation accuracy. AttnUNet [26] incorporates attention mechanisms to focus on relevant regions in the feature maps, improving segmentation performance by suppressing irrelevant background noise. Fan et al. [12] introduce PraNet for precise polyp segmentation that employs parallel reverse attention and edge-guidance to refine segmentation boundaries. UACANet [18] uses uncertainty-aware mechanisms to improve the reliability and robustness of segmentation outcomes. DeepLabv3+ [8] integrates atrous convolutions and spatial pyramid pooling to capture multi-scale context information. ACC-UNet [16] employs adaptive context capture mechanisms to dynamically adjust the receptive fields based on the input image.

## 2.2. Vision Transformers

Vision Transformers (ViTs) [11, 22] have emerged as a powerful alternative to CNNs, offering a new paradigm for medical image analysis tasks by leveraging the self-attention mechanism [4, 6, 27–30, 39]. By combining the strengths of CNNs for local feature extraction and Transformers for capturing long-range dependencies, TransUNet [6] achieves superior performance in medical image segmentation. SwinUNet [4] is introduced based on the Swin Transformer [22] architecture, which utilizes shifted windows to achieve hierarchical feature representation, enabling efficient computation. MedT [39], a lightweight Transformer model specifically designed for medical image segmentation, which employs gated axial attention mechanisms to focus on relevant regions and reduce computational complexity. Rahman et al. introduce CASCADE [27], a cascaded-attention decoding network using standard convolutions. Recently, EMCAD [30] introduces a depth-wise convolutions-based multi-scale decoder. Although,

CASCADE and EMCAD perform well in medical image segmentation, their segmentation accuracy and computational complexity solely depend on the strength and complexity of the exiting pretrained transformer encoder they use, thus making them less suitable for resource-constrained settings. In contrast, we propose to design an extremely efficient (ultra-lightweight) end-to-end (both encoder and decoder) architecture using the multi-kernel trick (where,  $k_1 = k_2$  or  $k_1 \neq k_2$  for  $k_1, k_2 \in \text{Kernels}$ ) with depthwise convolutions.

## 2.3. Lightweight CNNs

Recent efforts have focused on making CNNs more efficient for real-time and resource-constrained environments. MobileNets [14] and EfficientNets [36] introduce depthwise separable convolutions and compound scaling, respectively, to create lightweight models with competitive performance. Additionally, several novel lightweight architectures have been developed to further enhance the efficiency of medical image segmentation [21, 33, 37, 38]. UNeXt [38] uses hybrid convolutional and transformer blocks to capture local and global features efficiently, improving segmentation accuracy while maintaining computational efficiency. CMUNeXt [37] combines convolutional and multi-scale features to enhance segmentation performance. EGE-UNet [33] integrates edge-guided mechanisms to refine segmentation boundaries. Rolling-UNet [21] incorporates rolling convolutional blocks to enhance the model’s ability to capture long-range dependencies.

## 3. Method

We describe next our core building blocks. Then, we introduce our MK-UNet architecture by integrating these blocks into the baseline UNeXt [38] (Fig. 2a in green box).

### 3.1. Multi-kernel Inverted Residual (MKIR)

We first introduce the multi-kernel inverted residual (*MKIR*) block to generate and refine feature maps (Fig. 2c). By utilizing multiple (same or different) kernel sizes, *MKIR* allows for better understanding of both fine-grained details and broader contexts, thereby enabling a comprehensive representation of the input. As shown in Fig. 2c, the process begins by expanding the #channels (i.e.,  $\text{expansion\_factor} = 2$ ) through point-wise convolution  $PWC_1$ , batch normalization  $BN$  [17], and  $ReLU6$  activation [19]. This is followed by multi-kernel depth-wise convolution  $MKDC$  for capturing application-specific complex spatial contexts. A subsequent point-wise convolution  $PWC_2$  and  $BN$  restore the original #channels. The MKIR (Eq. 1) significantly reduces the computational cost while ensuring rich feature representation:

$$MKIR(x) = BN(PWC_2(MKDC(ReLU6(BN(PWC_1(x))))) \quad (1)$$

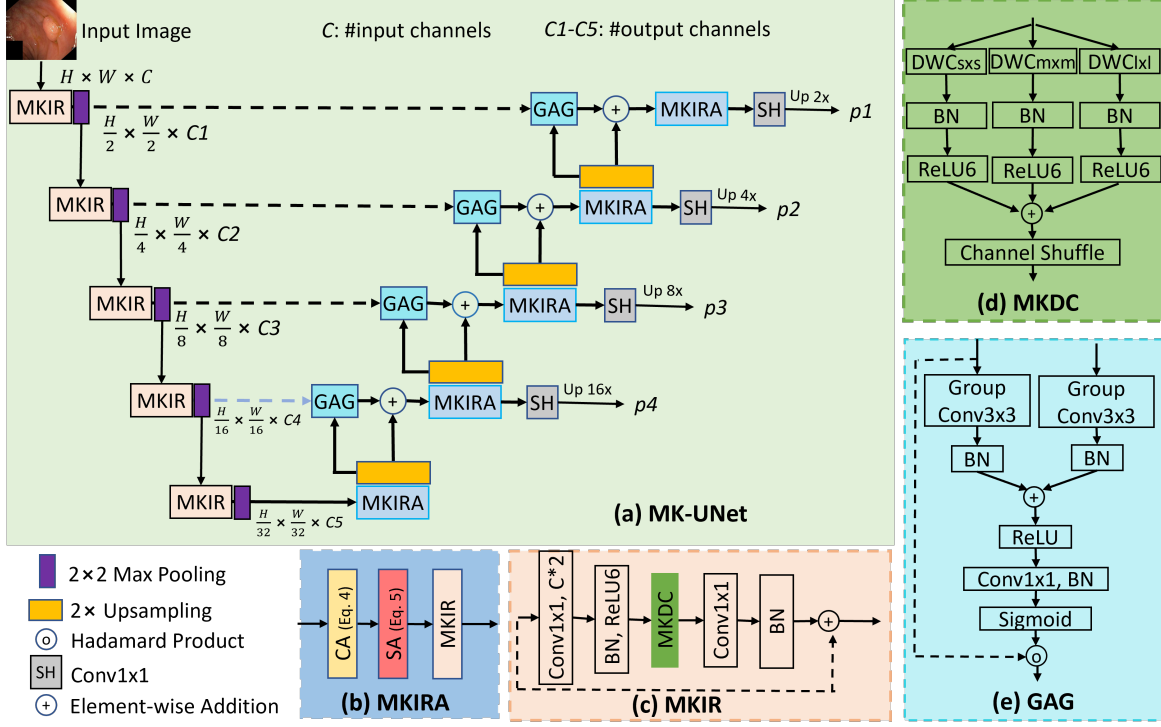


Figure 2. The proposed network diagram. (a) Five-stage MK-UNet network (b) Multi-kernel inverted residual attention (MKIRA), (c) Multi-kernel inverted residual (MKIR), (d) Multi-kernel (parallel) depth-wise convolution (MKDC), (e) Grouped attention gate (GAG). p1, p2, p3, and p4 are output segmentation maps. **Note:** Channel attention (CA), Spatial attention (SA).

where  $MKDC$  for multiple kernels ( $K$ ) is defined in Eq. 2 and Fig. 2d:

$$MKDC(x) = CS(\sum_{k \in K} DWCB_k(x)) \quad (2)$$

where  $DWCB_k(x) = ReLU6(BN(DWC_k(x)))$ . Here,  $DWC_k(\cdot)$  is a depth-wise convolution with the kernel  $k \times k$ . To address the channel independence in depth-wise convolution, a channel shuffle ( $CS$ ) is used to ensure the inter-channel information flow. Our  $MKDC(\cdot)$  differs from  $MSCB$  (in EMCAD [30]) in their core theoretical concepts. Our multi-kernel trick supports both  $k_1 = k_2$  (same-size kernels) and  $k_1 \neq k_2$  (different-size kernels) for  $k_1, k_2 \in K$  versus conventional multi-scale (only  $k_1 \neq k_2$ ) designs [20, 30, 35], thus allowing adaptable context extraction. This conceptual distinction allows MK-UNet to adapt kernel sizes based on application-specific needs (e.g., large kernels for large objects, small kernels for small objects, or mixed for both objects segmentation).

### 3.2. Multi-kernel Inverted Residual Attention (MKIRA)

We present a lightweight multi-kernel inverted residual attention, MKIRA, to refine the feature maps ( $x$ ). MKIRA uses channel attention (CA) [15] to focus on relevant channels, a spatial attention (SA) [7] to capture the local context,

and a multi-kernel inverted residual (MKIR) to enrich the feature maps while capturing contextual relationships. The MKIRA (Fig. 2b) is given in Eq. 3:

$$MKIRA(x) = MKIR(SA(CA(x))) \quad (3)$$

**Channel Attention (CA):** We use CA [15] to enhance relevant features by applying adaptive max (AMP) and average pooling (AAP) to condense the spatial information [27]; this is followed by channel reduction ( $r = 16$ ) through point-wise convolution  $PWC_1$  with  $ReLU$  ( $R$ ) activation [25] and expansion through  $PWC_2$ . The *Sigmoid* ( $\sigma$ ) activation generates attention weights, which are then applied to the input via the Hadamard product ( $\otimes$ ), while focusing on crucial feature maps. The CA is defined in Eq. 4:

$$CA(x) = \sigma(PWC_2(R(PWC_1(AMP(x)))) + PWC_2(R(PWC_1(AAP(x))))) \otimes x \quad (4)$$

**Spatial Attention (SA):** We use SA [7] to focus on specific image regions to highlight key features, crucial for accurate segmentation. The SA aggregates maximum ( $Ch_{max}$ ) and average ( $Ch_{avg}$ ) channel values to highlight local details, then it employs a large-kernel ( $7 \times 7$ ) convolution ( $LKC$ ) to strengthen the contextual connections. The *Sigmoid* ( $\sigma$ ) activation derives attention weights, applied to the input  $x$  via the Hadamard product ( $\otimes$ ), ensuring targeted refinement. The SA is derived in Eq. 5:

$$SA(x) = \sigma(LKC([Ch_{max}(x), Ch_{avg}(x)])) \otimes x \quad (5)$$



### 3.3. Grouped attention gate (GAG)

We use a grouped attention gate (GAG, Fig. 2e) that mixes the feature maps with the attention coefficients for enhancing the relevant features and suppressing the irrelevant ones. By utilizing a gating signal from higher-resolution features, GAG directs the information flow, thus improving medical image segmentation accuracy. Unlike Attention UNet [26], which processes signals with  $1 \times 1$  convolution, our method applies  $3 \times 3$  group convolutions (GC) to both gating ( $g$ ) and input ( $x$ ) feature maps separately [30]. After convolution, the features undergo batch normalization (BN) and get combined via addition, followed by *ReLU* ( $R$ ) activation. Subsequently, a  $1 \times 1$  convolution and batch normalization (BN) produce a unified feature map which, after the *Sigmoid* activation ( $\sigma$ ), generates the attention coefficients. These coefficients adjust the input feature  $x$ , and create an attention-enhanced output. GAG is defined in Eq.s 6:

$$GAG(g, x) = x \otimes \sigma(BN(Conv(R(BN(GC_g(g) + BN(GC_x(x))))))) \quad (6)$$

### 3.4. Multi-kernel UNet (MK-UNet)

Our MK-UNet employs a multi-kernel approach across five encoding and decoding stages to generate high-resolution segmentation maps, as depicted in Fig. 2a. Each encoding stage uses a multi-kernel inverted residual (MKIR) block to produce  $C_i$  feature maps, followed by max pooling for downsampling while retaining crucial information. The output from the final encoding stage passes through a multi-kernel inverted residual attention (MKIRA) block in the decoder’s initial stage, significantly refining the feature maps. These are then upsampled using bilinear interpolation for subsequent decoding stages. Decoder stages integrate skip-connections with refined features using a grouped attention gate (GAG) followed by additive aggregation. The resultant feature maps are refined through the MKIRA block and upsampled (only bilinear  $2\times$ , no convolutions) to align with the later stages.

The segmentation heads (SHs) at the last four stages output the segmentation maps  $p_4, p_3, p_2$ , and  $p_1$ . We consider the feature map  $p_1$  as final prediction and obtain the final segmentation output by employing a *Sigmoid* for binary segmentation or a *Softmax* activation for multi-class segmentation. We optimize the loss of only final prediction  $p_1$  for all binary segmentation tasks. However, we recommend using deep supervision for multi-class segmentation.

## 4. Experiments and Results

### 4.1. Datasets

We evaluate the MK-UNet’s efficacy across six datasets covering four segmentation tasks, including breast cancer (BUSI [1], 647 images: 437 benign and 210 malignant), polyp (ClinicDB [2] with 612 images, and ColonDB [40]

with 379 images), skin lesion (ISIC18 [9], 2,594 images), and cell nuclei/structure segmentation (DSB18 [3] with 670 images, and EM [5] with 30 images). These datasets, collected from various imaging centers, offer a broad diversity in image characteristics, ensuring a comprehensive evaluation. An 80:10:10 train-val-test split was applied across all datasets and the DICE score of testset is reported.

### 4.2. Implementation details

Our networks are developed and evaluated using Pytorch 1.11.0, operating on a single NVIDIA RTX A6000 GPU equipped with 48GB of RAM. We utilize multi-scale kernels [1, 3, 5] within our MKDC, based on an ablation study. The architecture employs a series of parallel depth-wise convolutions in the MK-UNet network, standardizing on channel configurations of [16, 32, 64, 96, 160] across all experiments, unless specified otherwise. Model optimization is achieved via the AdamW [24] optimizer with both learning rate and weight decay set to  $1e-4$ . Training spans over 200 epochs with batches of 16, during which we save the model achieving the highest DICE score.

Image dimensions are set to  $256 \times 256$  pixels for BUSI [1], ISIC18 [10], EM [5], and DSB18 [3] datasets, while for ClinicDB [2] and ColonDB [40], the resolution is adjusted to  $352 \times 352$  pixels. We utilize a multi-scale training approach, with scales of  $\{0.75, 1.0, 1.25\}$ , and enforce gradient clipping at 0.5. We do not apply any form of augmentation and use a hybrid loss function that combines (1:1) weighted BinaryCrossEntropy (BCE) with a weighted Intersection over Union (IoU) loss.

### 4.3. Results

Table 1 and Fig. 1 compare our MK-UNet with SOTA CNNs and transformers on six datasets of four binary medical segmentation tasks. Our MK-UNet achieves the top average DICE score of 89.75%, while maintaining an ultra-lightweight footprint of only 0.316M #Params and 0.314G #FLOPs. Our MK-UNet-T with 0.027M #Params and 0.062G #FLOPs, outperforms the existing most tiny model EGE-UNet [33] by on an average 5.93% DICE score over six datasets. The multi-kernel depth-wise convolutions within our network, alongside local attention mechanisms, play a crucial role in these strong results. Our MK-UNet’s performance on different datasets highlights its superior ability to balance accuracy with computational efficiency, setting a new benchmark for point-of-care diagnostics. The quantitative results of four different tasks are described next.

#### 4.3.1. Breast cancer segmentation

We focus on breast cancer segmentation in ultrasound images using the BUSI dataset. Our MK-UNet model notably outperforms existing methods, achieving the highest DICE score of 78.04% with remarkably low computa-

Network	Pretrain	#Params (M)	FLOPs (G)	Throughput (/s)	BUSI	ISIC18	Polyp		Cell		Avg.
							Clinic	Colon	DSB18	EM	
UNet [31]	No	34.53M	65.53G	119.03	74.04	86.67	91.43	83.95	92.23	95.36	87.28
UNet++ [44]	No	9.16M	34.65G	136.98	74.76	87.46	91.52	87.88	91.97	95.38	88.16
AttnUNet [26]	No	34.88M	66.64G	108.68	74.48	87.05	91.50	86.46	92.22	95.45	87.86
DeepLabv3+ [8]	Yes	39.76M	14.92G	128.20	76.81	88.64	92.46	89.86	92.14	94.96	89.15
PraNet [12]	Yes	32.55M	6.93G	64.10	75.14	88.46	91.71	89.16	89.89	92.37	87.79
UACANet [18]	Yes	69.16M	31.51G	43.28	76.96	88.72	93.29	89.76	88.86	89.28	87.81
TransUNet [6]	Yes	105.32M	38.52G	65.35	78.01	89.04	93.18	89.97	92.04	95.27	89.59
SwinUNet [4]	Yes	27.17M	6.2G	80.64	77.38	88.66	92.42	89.07	91.03	94.47	88.84
Rolling-UNet-S [21]	No	1.78M	2.1G	57.14	76.38	87.35	90.23	82.48	92.50	95.23	87.36
MedT [39]	No	1.57M	1.95G	8.40	69.23	86.78	83.44	68.90	92.28	93.87	82.42
UNeXt [38]	No	1.47M	0.57G	172.41	74.71	87.78	90.20	83.84	86.01	93.81	86.06
CMUNeXt [37]	No	0.418M	1.09G	<b>175.43</b>	77.34	87.51	92.82	83.85	92.58	95.38	88.25
EGE-UNet [33]	No	0.054M	0.072G	101.01	71.34	86.95	84.76	76.03	90.10	93.76	83.82
UltraLight_VMUNet [42]	No	0.050M	<b>0.060G</b>	98.03	72.31	87.85	87.11	80.06	91.88	93.96	85.53
<b>MK-UNet-T (Ours)</b>	No	<b>0.027M</b>	0.062G	140.85	75.64	88.19	91.26	85.03	92.38	94.69	87.87
<b>MK-UNet-S (Ours)</b>	No	0.093M	0.125G	139.42	77.26	88.57	92.31	88.78	92.45	95.22	89.10
<b>MK-UNet (Ours)</b>	No	0.316M	0.314G	138.89	78.04	88.74	93.48	90.01	92.71	95.52	89.75
<b>MK-UNet-M (Ours)</b>	No	1.15M	0.951G	133.73	78.27	89.08	93.67	90.27	92.74	95.62	89.94
<b>MK-UNet-L (Ours)</b>	No	3.76M	3.19G	122.62	<b>79.02</b>	<b>89.25</b>	<b>93.85</b>	<b>91.82</b>	<b>92.80</b>	<b>95.67</b>	<b>90.40</b>

Table 1. Results of binary (breast cancer, skin lesion, polyp, and cell) segmentation. We reproduce the results of SOTA methods using their publicly available implementations with our 80:10:10 train-val-test splits. FLOPs of all methods are reported for  $256 \times 256$  inputs. The FLOPs of all methods for polyp segmentation with  $352 \times 352$  inputs will be higher. We report DICE scores (%) averaging over five runs, thus having 1-4% standard deviations.  $[C1, C2, C3, C4, C5] = \text{MK-UNet-T [4, 8, 16, 24, 32], MK-UNet-S [8, 16, 32, 48, 80], MK-UNet [16, 32, 64, 96, 160], MK-UNet-M [32, 64, 128, 192, 320], and MK-UNet-L [64, 128, 256, 384, 512]}$ . Best results are shown in **bold**.

tional requirements. This achievement underscores the efficiency and effectiveness of our approach, particularly when compared against more complex models. For instance, TransUNet, despite having substantially more parameters (105.32M) and FLOPs (38.52G), achieves a lower DICE score of 78.01%. Similarly, ACC-UNet, with its 16.8M parameters and 38.0G FLOPs, closely follows with a DICE score of 77.02%. Additionally, compared to the computationally comparable model UNeXt, our MK-UNet demonstrates a significant improvement, surpassing it by a margin of 3.33% with  $4.7\times$  fewer #Params.

#### 4.3.2. Skin lesion segmentation

On the ISIC18 dataset for skin lesion segmentation, our MK-UNet network exhibits a commendable performance with a DICE score of 88.74%. This result not only showcases our model’s robustness, but also its capability to handle the complex variations present in skin lesion images. We note that TransUNet leads with a marginally higher score of 89.04% but with a significantly higher cost ( $333\times$  and  $123\times$  larger #Params and #FLOPs, respectively). Our MK-UNet achieves near-top performance while maintaining minimal computational requirements—merely 0.316M #Params and 0.314G #FLOPs.

#### 4.3.3. Polyp segmentation

In polyp segmentation on Clinic and Colon datasets, our MK-UNet model excels with leading scores of 93.48% and 90.01%, respectively, and with a remarkably low computational footprint. This performance surpasses notable competitors like DeepLabv3+ and ACC-UNet, which, de-

spite their higher resource consumption, do not match MK-UNet’s precision. Specifically, DeepLabv3+ reaches 92.46% on Clinic and 89.86% on Colon with  $126\times$  and  $47.5\times$  larger parameters and FLOPs, respectively, while ACC-UNet closely follows but still falls short.

#### 4.3.4. Microscopic cell nuclei/structure segmentation

For cell nuclei/structure segmentation on the DSB18 and EM datasets, our MK-UNet network demonstrates exceptional accuracy, achieving DICE scores of 92.71% and 95.52%, respectively. In contrast, other models like TransUNet and UNeXt, despite their heavy-weight design and higher computational demands, do not surpass MK-UNet’s DICE score. For instance, TransUNet achieves lower scores of 92.04% on DSB18 and 95.27% on EM, while UNeXt falls 6.70% behind MK-UNet.

### 4.4. Qualitative results

In Figure 3, we report the segmentation maps of breast tumors, skin lesions, polyps, and cell segmentation for representative test images. In breast tumor segmentation, UNet, UNet++, and UNeXt show greater false segmentation, while TransUNet and our MK-UNet produce near-perfect segmentation maps. Similarly, in skin lesion segmentation, UNet, ResUNet, UNet++, AttnUNet, DeepLabV3+, PraNet, SwinUNet, and UNeXt miss part of the lesion (in red rectangular box). However, UACANet, TransUNet, ACC-UNet, and our MK-UNet can segment that challenging region well. Our MK-UNet can also segment the polyp correctly, while all other methods incorrectly segment an-

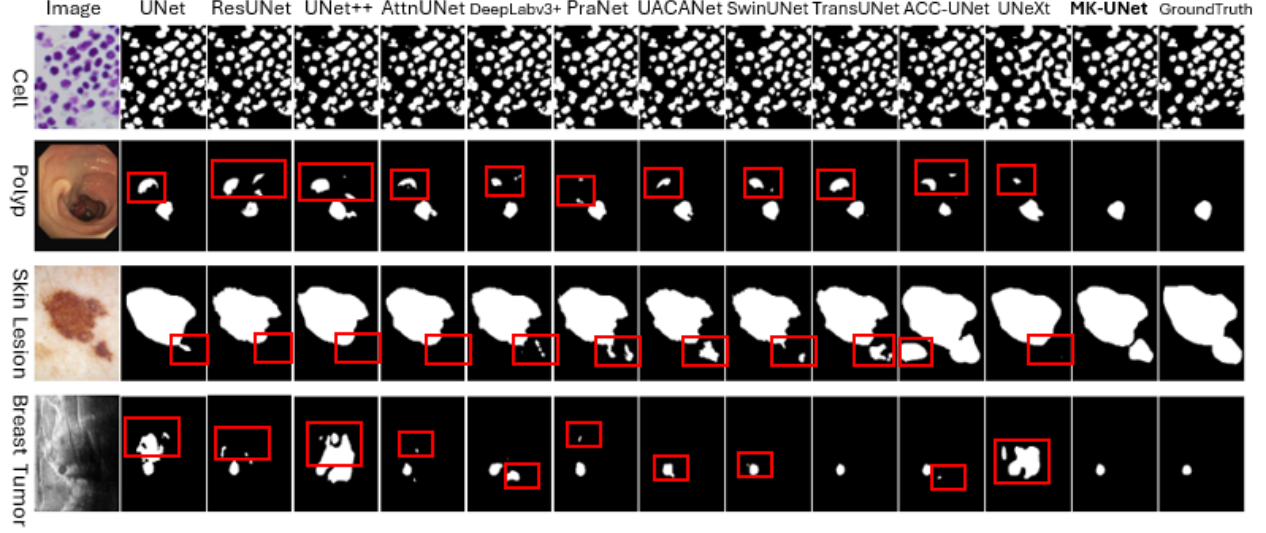


Figure 3. Qualitative results of our MK-UNet and SOTA methods. The incorrect segmented regions by different methods are highlighted using the red rectangular box.

Network	#Params	#FLOPs	BUSI	Clinic	Colon	ISIC18	DSB18	EM
Baseline UNeXt	1.47M	0.57G	74.71	90.20	83.84	87.78	86.01	93.81
Redesigned Mobile UNet	0.271M	0.230G	72.41	90.90	84.15	87.20	90.52	94.87
MKIR	0.306M	0.300G	74.74	92.63	86.46	88.22	92.40	95.31
MKIR + GAG	0.310M	0.311G	74.98	91.97	86.56	88.34	92.67	95.48
MKIR + MKIRA	0.311M	0.303G	76.61	92.64	89.40	88.56	92.64	95.37
<b>MKIR + GAG + MKIRA (Ours)</b>	0.316M	0.314G	<b>78.04</b>	<b>93.48</b>	<b>90.01</b>	<b>88.64</b>	<b>92.71</b>	<b>95.52</b>

Table 2. Effect of different components of MK-UNet with #channels = [16, 32, 64, 96, 160] and [1, 3, 5] kernels. UNeXt has #channels = [16, 32, 128, 160, 256]. We design Mobile UNet following the structure of UNeXt network. However, we use the #channels = [16, 32, 64, 96, 160] and kernel size of [3] with the original inverted residual block (IRB) in the Mobile UNet. We report the DICE scores (%) averaging over five runs. Best results are shown in **bold**.

other region as a polyp. In general, our MK-UNet produces the best overlapping segmentation map in all four tasks. The reason behind this well-rounded performance by our MK-UNet with a very low computational budget is the use of multi-kernel depth-wise convolutions along with gated and local attention mechanisms.

## 5. Ablation Study

This section describes three critical ablation studies. More ablation results are given in the Appendix.

### 5.1. Effect of different components

Table 2 presents the performance of various configurations within the MK-UNet network across six medical image segmentation datasets, highlighting the impact of integrating different components like MKIR, GAG, and MKIRA. The comparison spans models from UNeXt to the advanced MKIR + GAG + MKIRA variant, revealing a progressive improvement in the DICE scores with the addition of each component. Notably, the multi-kernel trick, implemented through MKIR (in the encoder) and MKIRA (in the decoder) blocks, is the most critical component for improv-

ing the segmentation accuracy, increasing the DICE score from 72.41% to 76.61% on the BUSI dataset. This indicates the significant contribution of the multi-kernel approach to feature extraction and refinement. However, when we integrate all proposed modules (MKIR + GAG + MKIRA), our model achieves the highest overall DICE score of 78.04% on the same dataset, with minimal computational resources (0.316M #Params and 0.314G #FLOPs). This exhibits the efficacy of combining multi-kernel convolution with attention mechanisms within MK-UNet.

### 5.2. Effect of multiple kernels

Table 3 evaluates the influence of different convolutional kernel combinations on the performance of MKDC within the MK-UNet network, specifically for the BUSI dataset. By experimenting with a variety of kernel sizes ranging from 1 to 3, 5, 7, it becomes evident that a mix of 1, 3, 5 kernels stands out by achieving the best DICE score of 78.04% with a moderate increase in computational resources (0.316M #Params and 0.314G #FLOPs). This finding highlights the effectiveness of a multi-scale kernel approach in capturing diverse feature representations, thus significantly improving segmentation accuracy without a sub-

Convolution kernels	#Params(M)	FLOPs(G)	DICE	Convolution kernels	#Params(M)	FLOPs(G)	DICE
$1 \times 1$	0.272	0.220	70.83	$5 \times 5$	0.299	0.276	76.81
$1 \times 1, 1 \times 1$	0.275	0.229	71.11	$1 \times 1, 5 \times 5$	0.303	0.286	77.05
$3 \times 3$	0.281	0.239	76.42	$3 \times 3, 3 \times 3, 3 \times 3$	0.306	0.295	76.86
$1 \times 1, 3 \times 3$	0.284	0.248	76.81	$3 \times 3, 5 \times 5$	0.312	0.304	77.62
$1 \times 1, 1 \times 1, 3 \times 3$	0.288	0.257	77.08	<b><math>1 \times 1, 3 \times 3, 5 \times 5</math></b>	0.316	0.314	<b>78.04</b>
$3 \times 3, 3 \times 3$	0.294	0.267	76.83	$5 \times 5, 5 \times 5$	0.331	0.342	77.88
$1 \times 1, 3 \times 3, 3 \times 3$	0.297	0.276	77.26	$5 \times 5, 5 \times 5, 5 \times 5$	0.362	0.408	77.80

Table 3. Effect of multiple kernels in the depth-wise convolution of MKDC on BUSI dataset. The results for kernels beyond  $7 \times 7$  are not reported as the performance does not scale proportionally with the computational cost of larger kernels. We use the MK-UNet network with #channels= [16, 32, 64, 96, 160] for these experiments and report the FLOPs for  $256 \times 256$  inputs. We report the DICE scores (%) averaging over five runs. Best results are highlighted in **bold**.

Blocks	#Params	#FLOPs	BUSI	Clinic	Colon	ISIC18	DSB18	EM
IRB	0.271M	0.230G	72.41	90.90	84.15	87.20	90.52	94.87
MKIR (Ours)	0.306M	0.300G	<b>74.74</b>	<b>92.63</b>	<b>86.46</b>	<b>88.22</b>	<b>92.40</b>	<b>95.31</b>

Table 4. Original Inverted Residual Block (IRB) [34] vs our Multi-Kernel Inverted Residual (MKIR) with #channels = [16, 32, 64, 96, 160]. We use the kernel size of [3] and [1, 3, 5] for IRB and MKIR, respectively. We report the DICE scores (%) averaging over five runs. Best results are shown in **bold**.

Encoder	Decoder	#Params	#FLOPs	BUSI	Clinic	Colon	ISIC18	DSB18	EM
MKIRA	MKIRA	0.321M	0.346G	77.28	92.81	89.63	88.61	92.65	95.43
(Ours) MKIR	MKIRA	0.316M	0.314G	<b>78.04</b>	<b>93.48</b>	<b>90.01</b>	<b>88.64</b>	<b>92.71</b>	<b>95.52</b>

Table 5. Effect of MKIRA in the encoder and decoder of MK-UNet with #channels = [16, 32, 64, 96, 160] and [1, 3, 5] kernels. We report the DICE scores (%) averaging over five runs. Best results are shown in **bold**.

stantial rise in computational demands. Drawing from these empirical findings, we opt for the kernel combination of [1, 3, 5] across all our experiments.

### 5.3. Effectiveness of our MKIR over IRB [34]

Table 4 reports the results of the original IRB of MobileNetV2 [34] and our proposed MKIR block. It can be concluded from the table that our MKIR significantly outperforms (up to 2.33%) IRB in all the datasets with only an additional 0.035M #Params and 0.07G #FLOPs. The use of lightweight convolutions with multiple kernels contributes to these performance improvements with nominal additional computational resources.

### 5.4. Effectiveness of MKIR vs. MKIRA in Encoder

Table 5 demonstrates that employing MKIR in the encoder and MKIRA in the decoder yields superior performance across all datasets. Specifically, this configuration achieves the best average DICE scores of 78.04% (BUSI), 93.48% (Clinic), 90.01% (Colon), 88.64% (ISIC18), 92.71% (DSB18), and 95.52% (EM). The MKIR block in the encoder effectively extracts complex features by leveraging multiple kernels to capture a diverse range of spatial patterns and global contexts without the need for localized attention, which is more computationally intensive. Since the encoder primarily focuses on feature extraction, this design helps preserve critical details while maintaining lightweightness. In contrast, localized attention is crucial

in the decoder to facilitate precise reconstruction. The MKIRA in the decoder attends to key spatial regions, enabling effective feature refinement. This complementary setup leads to an optimal balance between performance and computational cost, as evidenced by the superior results achieved with only 0.316M #Params and 0.314G #FLOPs.

## 6. Conclusion

In this paper, we have presented MK-UNet, a significant advancement in the realm of medical image segmentation. MK-UNet addresses the long-standing challenge of balancing high segmentation accuracy with computational efficiency. By utilizing depth-wise convolution and multi-kernel processing, MK-UNet outperforms models like TransUNet and UNeXt with a fraction of their computational overhead (nearly  $333\times$  lower #Params and  $123\times$  less complexity than TransUNet, and  $4.7\times$  fewer #Params compared to UNeXt). This significant reduction in computational resources, coupled with improved performance, makes MK-UNet an optimal choice for deployment in environments where computational resources are limited, such as in point-of-care devices.

The current evaluations of our proposed MK-UNet architecture are limited to binary medical segmentation tasks. In the future, we will extend experiments on multi-class and 3D segmentation tasks.



## Acknowledgments

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. 5
- [2] 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. 5
- [3] 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. 5
- [4] 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. In *Proceedings of the European Conference on Computer Vision Workshops*, pages 205–218. Springer, 2022. 1, 3, 6
- [5] 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. 5
- [6] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, page 103280, 2024. 1, 3, 6
- [7] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 5659–5667, 2017. 4
- [8] 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 Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017. 1, 2, 3, 6
- [9] 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
- [10] 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 International Symposium on Biomedical Imaging*, pages 168–172. IEEE, 2018. 5
- [11] 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. 3
- [12] 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*, pages 263–273. Springer, 2020. 1, 2, 3, 6
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3
- [14] 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. 3
- [15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 4
- [16] Nabil Ibtehaz and Daisuke Kihara. Acc-unet: A completely convolutional unet model for the 2020s. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 692–702. Springer, 2023. 3
- [17] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. pmlr, 2015. 3
- [18] Taehun Kim, Hyemin Lee, and Daijin Kim. Uacanet: Uncertainty augmented context attention for polyp segmentation. In *ACM International Conference on Multimedia*, pages 2167–2175, 2021. 1, 2, 3, 6
- [19] Alex Krizhevsky and Geoff Hinton. Convolutional deep belief networks on cifar-10. *Unpublished manuscript*, 40(7): 1–9, 2010. 3
- [20] Weifeng Lin, Ziheng Wu, Jiayu Chen, Jun Huang, and Lianwen Jin. Scale-aware modulation meet transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6015–6026, 2023. 4
- [21] Yutong Liu, Haijiang Zhu, Mengting Liu, Huaiyuan Yu, Zihan Chen, and Jie Gao. Rolling-unet: Revitalizing mlp’s ability to efficiently extract long-distance dependencies for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3819–3827, 2024. 1, 3, 6

- [22] 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*, pages 10012–10022, 2021. 3
- [23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 3
- [24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5
- [25] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning*, pages 807–814, 2010. 4
- [26] 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. In *Medical Imaging with Deep Learning*, 2018. 1, 2, 3, 5, 6
- [27] 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*, pages 6222–6231, 2023. 1, 3, 4
- [28] 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.
- [29] Md Mostafijur Rahman and Radu Marculescu. G-cascade: Efficient cascaded graph convolutional decoding for 2d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7728–7737, 2024.
- [30] Md Mostafijur Rahman, Mustafa Munir, and Radu Marculescu. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11769–11779, 2024. 1, 3, 4, 5
- [31] 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*, pages 234–241. Springer, 2015. 1, 2, 3, 6
- [32] Jiacheng Ruan, Suncheng Xiang, Mingye Xie, Ting Liu, and Yuzhuo Fu. Malunet: A multi-attention and light-weight unet for skin lesion segmentation. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1150–1156. IEEE, 2022. 1
- [33] Jiacheng Ruan, Mingye Xie, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. Ege-unet: an efficient group enhanced unet for skin lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 481–490. Springer, 2023. 1, 2, 3, 5, 6
- [34] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 8, 1
- [35] Hyunseok Seo, Seohee So, Sojin Yun, Seokjun Lee, and Jiseong Barg. Spatial feature conservation networks (sfcons) for dilated convolutions to improve breast cancer segmentation from dce-mri. In *International Workshop on Applications of Medical AI*, pages 118–127. Springer, 2022. 4
- [36] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 3
- [37] Fenghe Tang, Jianrui Ding, Quan Quan, Lingtao Wang, Chunping Ning, and S Kevin Zhou. Cmunext: An efficient medical image segmentation network based on large kernel and skip fusion. In *IEEE International Symposium on Biomedical Imaging*, pages 1–5. IEEE, 2024. 1, 3, 6
- [38] Jeya Maria Jose Valanarasu and Vishal M Patel. Unext: Mlp-based rapid medical image segmentation network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 23–33. Springer, 2022. 1, 3, 6
- [39] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 36–46. Springer, 2021. 1, 3, 6
- [40] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of Healthcare Engineering*, 2017, 2017. 5
- [41] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *International Conference on Machine Learning*, pages 3–19, 2018. 1
- [42] Renkai Wu, Yinghao Liu, Pengchen Liang, and Qing Chang. Ultralight vm-unet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation. *arXiv preprint arXiv:2403.20035*, 2024. 6
- [43] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018. 3
- [44] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *International Workshop on Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018. 1, 2, 3, 6