

Modality-agnostic Domain Generalizable Medical Image Segmentation by Multi-Frequency in Multi-Scale Attention

Ju-Hyeon Nam¹ Nur Suriza Syazwany¹ Su Jung Kim¹ Sang-Chul Lee^{1,2}

Department of Electrical and Computer Engineering, Inha University¹, DeepCardio²

{jhnam0514, surizasyazwany, sk2266}@inha.edu sclee@{inha.ac.kr, deepcardio.com}

Abstract

Generalizability in deep neural networks plays a pivotal role in medical image segmentation. However, deep learning-based medical image analyses tend to overlook the importance of frequency variance, which is critical element for achieving a model that is both modality-agnostic and domain-generalizable. Additionally, various models fail to account for the potential information loss that can arise from multi-task learning under deep supervision, a factor that can impair the model's representation ability. To address these challenges, we propose a Modality-agnostic Domain Generalizable Network (MADGNet) for medical image segmentation, which comprises two key components: a Multi-Frequency in Multi-Scale Attention (MFMSA) block and Ensemble Sub-Decoding Module (E-SDM). The MFMSA block refines the process of spatial feature extraction, particularly in capturing boundary features, by incorporating multi-frequency and multi-scale features, thereby offering informative cues for tissue outline and anatomical structures. Moreover, we propose E-SDM to mitigate information loss in multi-task learning with deep supervision, especially during substantial upsampling from low resolution. We evaluate the segmentation performance of MADGNet across six modalities and fifteen datasets. Through extensive experiments, we demonstrate that MADGNet consistently outperforms state-of-the-art models across various modalities, showcasing superior segmentation performance. This affirms MADGNet as a robust solution for medical image segmentation that excels in diverse imaging scenarios. Our MADGNet code is available in [GitHub Link](#).

1. Introduction

Various types of cancers continue to pose a substantial threat to human life, contributing significantly to global mortality rates. In this context, the importance of medical image analysis becomes evident, serving as a linchpin in the

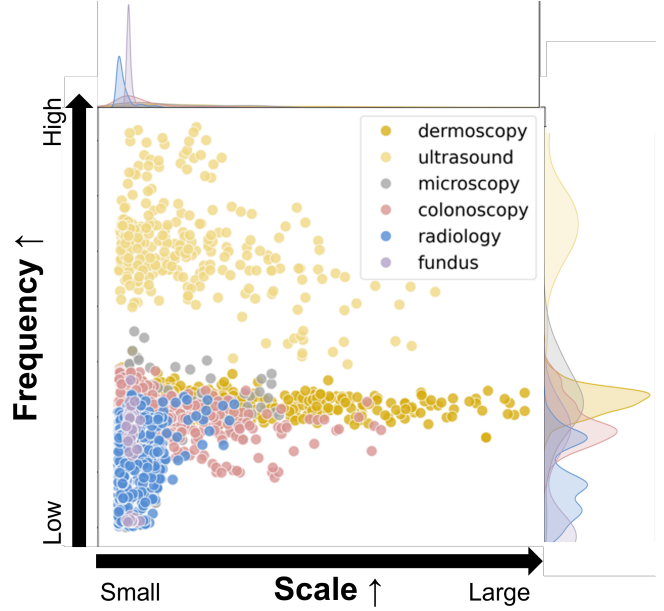


Figure 1. Scale vs Frequency distribution per modality. The scale denotes the size of lesions, measured as the ratio of foreground pixels to the total number of pixels. Frequency is calculated by the power spectrum ratio of the high-frequency and full-frequency. We observe that the frequency variance is higher than the scale, which mainly focuses on solving the various sizes in other methods.

early detection of malignant tumors or abnormal cells, ultimately contributing to extending patients' lives [10]. However, the manual analysis of noisy and blurred medical images presents notable challenges, rendering it susceptible to human errors [9]. In response to these issues, computer-aided diagnosis using traditional segmentation algorithms [24, 33, 46, 59] has garnered attention from clinical experts. Nevertheless, these algorithms still lack generalizability to new patient cases due to issues such as uneven intensity distribution, unexpected artifacts, and severe noise in medical images [50].

Recent advancements in deep learning [34, 60] have paved the way for their application in medical image segmentation. In particular, UNet [51] has emerged as a fundamental approach for medical image segmentation, owing to its effective utilization of skip connections within its U-shaped architecture. An example of this evolution is UNet++ [73], which employs nested UNet and deep supervision with skip connections to reduce the semantic gap between the encoder and decoder. With the recent surge in attention mechanisms [27, 47, 62, 64], the evolution of UNet has witnessed the integration of attention, extracting discriminative features from noisy medical images, as seen in Attention UNet [44] and Focus UNet [69]. Nevertheless, owing to disparities in medical image acquisition methods (modality) and the limited accessibility of datasets by patient privacy, these models often lack generalizability for unseen clinical settings, rendering practical implementation challenging. Therefore, exploring the question, “*How a well-trained model in a modality can be generalized into other modalities (modality-agnostic) and be applicable in unseen clinical settings (domain generalizability)?*” is a critical task to address these challenges.

Plotting the scale and frequency distribution graph (Fig. 1) unveiled a significant finding: medical images exhibit distinct distributions in both dimensions. While acknowledging the inherent interdependence between multi-frequency and multi-scale, Fig. 1 emphasizes that they still reveal distinctive information. Multi-frequency, in particular, displays more variance, highlighting unique and valuable insights. This challenges the prevailing focus on either multi-scale [18, 22, 56] or multi-frequency information [4, 68] and underscores the necessity of recognizing the independent and complementary nature of both dimensions in medical image segmentation. This discovery prompted us to efficiently utilize both dimensions, recognizing that even with interdependence, each dimension contributes distinct and valuable information.

To address the aforementioned problem, we propose a novel attention mechanism called Multi-Frequency in Multi-Scale Attention (MFMSA) block. This block employs multi-frequency channel attention (MFCA) with 2D Discrete Cosine Transform (2D DCT) [2] to produce a channel attention map by extracting frequency statistics. Subsequently, multi-scale spatial attention (MSSA) is applied to extract discriminative boundary features and aggregate them from each scale. By effectively leveraging the sequential attention method to suppress the influence of noisy channels via MFCA and extracting discriminative feature maps for boundaries across various scales through MSSA, the MFMSA block’s dual utilization enables the comprehensive capture of a broad spectrum of information from medical images. In more detail through multi-frequency analysis, it extract diverse features across dif-

ferent frequency bands while simultaneously performing multi-scale analysis to capture intricate details and broader structural information, ensuring a comprehensive understanding of the image content. Additionally, we introduce a Ensemble Sub-Decoding Module (E-SDM) to prevent information loss caused by drastic upsampling during multi-task learning with deep supervision. The resulting model, MADGNet, which mainly comprises the MFMSA block and E-SDM, achieved the highest segmentation performance in various modalities and clinical settings. The contributions of this paper can be summarized as follows:

- We propose **MADGNet** for medical image segmentation in various modalities (**modality-agnostic**) and unseen clinical settings (**domain generalizability**) by efficiently fusing multi-frequency and multi-scale features.
- Our novel MFMSA block is an efficient solution for integrating multi-scale and multi-frequency information, both of which are crucial for generalizing across various modalities and unseen clinical settings. E-SDM prevents information loss caused by drastic upsampling during multi-task learning with deep supervision in an ensemble manner.
- We evaluated MADGNet on fifteen different datasets with six different modalities and demonstrated that our model achieves the highest segmentation performance on various datasets. Furthermore, we showed that MADGNet can be utilized for multi-label medical image segmentation.

2. Related Works

Recently, the significance of multi-scale information in the context of semantic segmentation has garnered increasing attention. For instance, DCSAUNet [66] retains the main features of the input image while combining feature maps with various receptive fields, resulting in enhanced segmentation quality. Simultaneously, several works [8, 36, 37, 58, 67] have demonstrated the effectiveness of enlarging the receptive field to capture multi-scale objects, often in conjunction with attention mechanisms. MS-CNN [31] improves context alignment through attention layers and multi-scale features, whereas DMSANet [52] proposes a lightweight module that integrates local and global features with spatial and channel attention.

In parallel, research efforts have surged to integrate multi-frequency techniques and attention mechanisms, aiming to enhance the extraction of local and global context for fine to coarse information by implementing various frequency transformation methods [2, 19, 29]. In particular, 2D DCT is widely used in computer vision to extract the frequency statistic feature for increasing the representation power due to its compression ability. For instance, FCANet [49] proposed a generalized SE block [27] based on 2D DCT basis functions. However, current medical image seg-

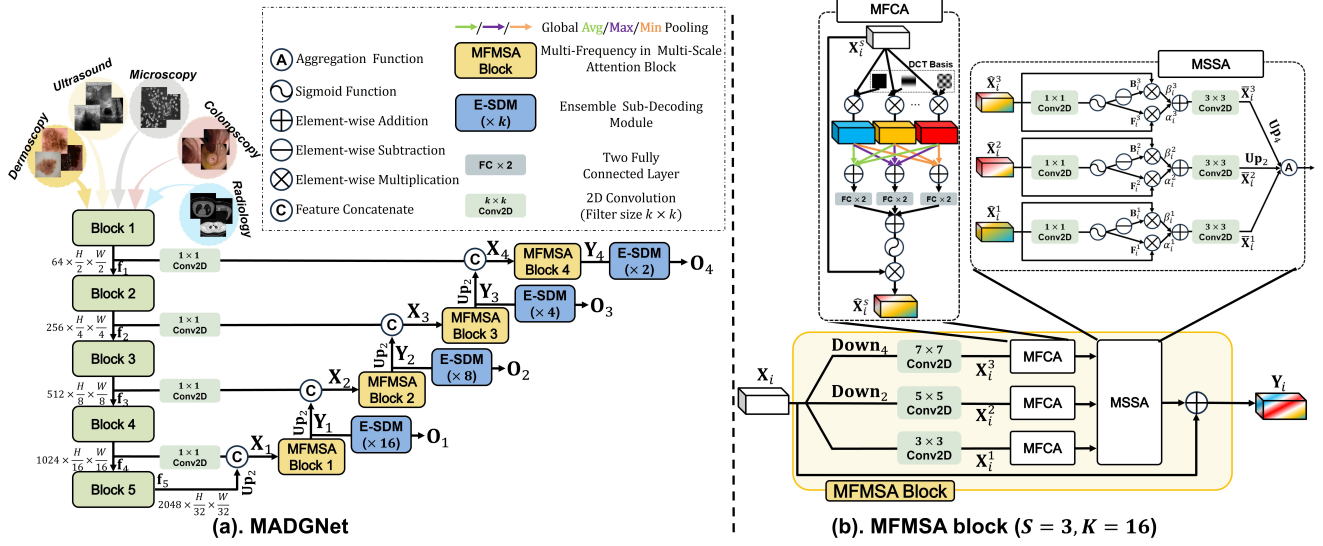


Figure 2. (a) The overall architecture of the proposed MADGNet mainly comprises MFMSA block and E-SDM (See Fig. 3). (b) MFMSA block contains S scale branches ($S = 3$ in this figure) where the s -th branch input feature map are downsampled into η^{s-1} ($\eta = \frac{1}{2}$ in this figure). As our MFMSA block considers two dimensions (scale and frequency), MADGNet achieves the highest performance in various modalities and other clinical settings. Additionally, since E-SDM predicts a core task from sub-tasks, the final output is more accurate than when processed parallelly.

mentation methods tend to focus on multi-scale or multi-frequency aspects, ignoring the potential benefits of combining both. To address this limitation, we integrate multi-frequency information with multi-scale features, thereby improving the model’s ability to detect subtle variations in lesion characteristics, which is crucial for accurate medical image segmentation.

Additionally, multi-task learning and deep supervision are recognized strategies for enhancing the representation power of models by training various tasks at low levels. These strategies are widely used in medical image segmentation, which demands dense predictions [11, 16, 17, 28, 42, 73]. However, when training multiple tasks with deep supervision, a challenge arises during upsampling of a low-resolution feature map to a high resolution, potentially leading to information loss. This effect is particularly pronounced when predicting detailed boundaries, negatively impacting model training. To address this challenge, we propose a multi-task learning scheme with deep supervision in an ensemble manner.

3. Method

3.1. Multi-Frequency in Multi-Scale Attention Block

Motivation: The human visual system seamlessly integrates multi-scale and multi-frequency information to accurately interpret the environment [39, 43]. Our approach mirrors a comprehensive analysis of visual information across di-

verse scales and frequencies designed to enhance tasks in the medical domain, motivated by Fig. 1 and the operational algorithm of the human visual system. This process encompasses the wide variation in lesion sizes in medical images, necessitating multi-scale features for precisely segmenting regions such as tumors, polyps, and cells. Moreover, as medical images exhibit higher frequency variance than scale due to modality characteristics, facilitating multi-frequency information is vital for crafting an effective medical image segmentation model. Propelled by these insights, we propose the innovative *Multi-Frequency in Multi-Scale Attention (MFMSA) block*, efficiently integrating multi-scale and multi-frequency information to effectively address a critical aspect often neglected in prior approaches. The overall architecture of the MADGNet and MFMSA block is illustrated on Fig. 2. And, the MFMSA block can be divided into three steps: 1) *Scale Decomposition*, 2) *MFCA*, and 3) *MSSA*.

Feature Extraction. We use a pre-trained ResNeSt [70] consisting of split-attention residual blocks to extract feature maps from input images due to enhanced feature representations. We also conducted experiments on various backbone (ResNet [25], Res2Net [20], Vision Transformer [13]), in Appendix 9.1. Let f_i be the feature maps from the i -th stage encoder block for $i = 1, 2, 3, 4, 5$. As the number of channels in each stage influences a decoder’s complexity, we reduce the number of channels into C_e using 2D convolution. Furthermore, during decoding to restore the features to input image resolutions, we fuse two features from the

encoder and previous decoder blocks as follows:

$$\mathbf{X}_i = \text{Cat}(\text{Conv2D}_1(\mathbf{f}_{5-i}), \text{Up}_2(\mathbf{Y}_{i-1})), \quad (1)$$

where $\text{Conv2D}_k(\cdot)$, $\text{Cat}(\cdot, \cdot)$ and $\text{Up}_m(\cdot)$ denote the 2D convolution with a kernel size of k , concatenation between feature maps along the channel dimension and a upsampling with scale factor 2^{m-1} , respectively. And, $\mathbf{Y}_0 = \mathbf{f}_5$. For convenience, we assume that $C_i = C_e + C_{i-1} = C$, $H_i = H$, $W_i = W$.

Scale Decomposition. To produce multi-scale features from input feature map $\mathbf{X}_i$, we assume a total of S branches operating on S different scales. For each s -th scale branch ($1 \leq s \leq S$), we reduce the resolution and channels of input feature map $\mathbf{X}_i$ with a channel and resolution reduction ratio $\gamma \in (0, 1)$, respectively, for computational efficiency. Thus, each input feature map at the s -th scale branch can be written as follows:

$$\mathbf{X}_i^s = \text{Conv2D}_{2s+1}(\text{Down}_s(\mathbf{X}_i)) \in \mathbb{R}^{C_s \times H_s \times W_s}, \quad (2)$$

where $\text{Down}_s(\cdot)$ is a downsampling with scale factor 2^{s-1} . Additionally, we denote the number of channels, height, and width at the s -th scale branch as $C_s = \max(\frac{C}{\gamma^{s-1}}, C_{\min})$, $H_s = \max(\frac{H}{2^{s-1}}, H_{\min})$, and $W_s = \max(\frac{W}{2^{s-1}}, W_{\min})$, respectively.

Multi-Frequency Channel Attention (MFCA). Recently, the compression ability of 2D DCT by expressing an image into a weighted sum of basis images produced by cosine functions oscillating at different frequencies has gained attention for feature extraction in the frequency domain [21, 49, 53]. Features at each scale branch can be characterized using 2D DCT with basis images $\mathbf{D}$ as follows:

$$\mathbf{X}_i^{s,k} = \sum_{h=0}^{H_s-1} \sum_{w=0}^{W_s-1} (\mathbf{X}_i^s)_{:,h,w} \mathbf{D}_{h,w}^{u_k,v_k}, \quad (3)$$

where (u_k, v_k) is the frequency indices corresponding to $\mathbf{X}_i^{s,k}$. Moreover, 2D DCT basis images at the s -th scale branch are defined as $\mathbf{D}_{h,w}^{u_k,v_k} = \cos(\frac{\pi h}{H_s}(u_k + \frac{1}{2})) \cos(\frac{\pi w}{W_s}(v_k + \frac{1}{2}))$ with a top- K selection strategy [49]. Subsequently, each $\mathbf{X}_i^{s,k}$ is compressed into $\mathbf{Z}_{\text{avg}}$, $\mathbf{Z}_{\text{max}}$, and $\mathbf{Z}_{\text{min}}$ using Global Average Pooling, Global Max Pooling, and Global Min Pooling and aggregates each statistic of frequency to produce channel attention map at the s -th scale branch by using two fully-connected layers $\mathbf{W}_1 \in \mathbb{R}^{C_s \times \frac{C_s}{r}}$ and $\mathbf{W}_2 \in \mathbb{R}^{\frac{C_s}{r} \times C_s}$ with reduction ratio r , as follows:

$$\mathbf{M}_i^s = \sigma\left(\sum_{d \in \{\text{avg}, \text{max}, \text{min}\}} \mathbf{W}_2(\delta(\mathbf{W}_1 \mathbf{Z}_d))\right) \in \mathbb{R}^{C_s}, \quad (4)$$

where δ and σ denote ReLU and Sigmoid activation functions, respectively. Finally, we recalibrate the feature map $\mathbf{X}_i^s$ using $\mathbf{M}_i^s$ at the s -th scale branch as $\hat{\mathbf{X}}_i^s = \mathbf{X}_i^s \times \mathbf{M}_i^s$. **Multi-Scale Spatial Attention (MSSA).** The channel-recalibrated feature map $\hat{\mathbf{X}}_i^s$ is used to determine discriminative boundary cues with various scales in the spatial domain. At this stage, we introduce the two learnable parameters (α_i^s and β_i^s) for each scale branch to control the information flow between foreground and background, respectively, as follows:

$$\bar{\mathbf{X}}_i^s = \text{Conv2D}_3\left(\alpha_i^s(\hat{\mathbf{X}}_i^s \times \mathbf{F}_i^s) + \beta_i^s(\hat{\mathbf{X}}_i^s \times \mathbf{B}_i^s)\right), \quad (5)$$

where $\mathbf{F}_i^s = \sigma(\text{Conv2D}_1(\hat{\mathbf{X}}_i^s))$ is a foreground attention map. Accordingly, the background attention map can be derived via elementary-wise subtraction between $\mathbf{F}_i^s$ and a matrix filled with one, that is, $\mathbf{B}_i^s = 1 - \mathbf{F}_i^s$. At that point, we restore the number channel at the s -th scale branch C_s in to C . Finally, to aggregate each refined feature from different scale branches, we apply upsampling to match the resolutions. For more stable training, we apply residual connection after the spatially refined feature map as follows:

$$\mathbf{Y}_i = \mathbf{X}_i + \mathbf{A}(\bar{\mathbf{X}}_i^1, \text{Up}_2(\bar{\mathbf{X}}_i^2), \dots, \text{Up}_S(\bar{\mathbf{X}}_i^S)), \quad (6)$$

where $\mathbf{A}(\cdot)$ is the feature aggregation function. Subsequently, $\mathbf{Y}_i$ is concatenated with feature map from encoder and E-SDM to apply stable deep supervision. As the channel attention is initially applied in the MFCA, the channel information is enriched for representation, reducing the effect of noisy contained medical images. Subsequently, more discriminative feature maps for boundary cues robust to various scale and noise are produced using the MSSA. This dual process allows our model understand the subtle anatomical differences between various modalities and complex characteristics of irregular lesions in noisy medical images.

Parameter Analysis. To further analyze the MFMSA block, we compare its efficiency with the most simple variant of our block, a single-frequency in single-scale block. A theoretical proof of the number of parameters for each step is reported in Appendix 7.1. Consequently, the ratio of the number of parameters to the original scale branch at the s -th scale branch is $\frac{p_s}{p} = \gamma^{s-1}$, where $p_1 = p$. Based on this result, we conclude that the MFMSA block contains more $\sum_{s=1}^S \frac{p_s}{p} = \frac{1-\gamma^S}{1-\gamma}$ than the single-frequency in single-scale block.

3.2. Ensemble Multi-task Learning with Deep Supervision

Motivation: Multi-task learning with deep supervision is a well-known training strategy for enhancing the model rep-

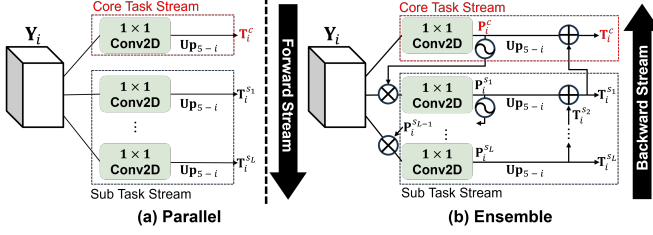


Figure 3. Comparison of multi-task learning method between (a) parallel and (b) ensemble manners.

resentation power and preventing gradient vanishing problems; this is achieved through co-training with core task and other L sub-tasks $\mathbf{O}_i = \{\mathbf{T}_i^c, \mathbf{T}_i^{s1}, \dots, \mathbf{T}_i^{sL}\}$ at each i -th decoder stage [16, 17, 72]. However, low-resolution feature maps must be upsampled into a high resolution to calculate the loss function between the ground truth and prediction. This drastic upsampling interferes with the model’s representation ability due to information loss for predicting detailed boundaries and structures.

To solve this problem, we propose a *ensemble sub-decoding module (E-SDM)*, a novel training strategy for multi-task learning with deep supervision. Fig. 3 illustrates the difference between (a) parallel and (b) ensemble manners. The main idea is to cascadingly supplement the information loss by incorporating sub-task predictions after upsampling, thereby improving the core task prediction.

Forward Stream. During the forward stream, core and sub-task pseudo predictions $\{\mathbf{P}_i^c, \mathbf{P}_i^{s1}, \dots, \mathbf{P}_i^{sL}\}$ are produced at the i -th decoder stage as follows:

$$\begin{cases} \mathbf{P}_i^c = \text{Conv2D}_1(\mathbf{Y}_i) \\ \mathbf{P}_i^{sl} = \text{Conv2D}_1(\mathbf{Y}_i \times \sigma(\mathbf{P}_i^{s_{l-1}})) \text{ for } 1 \leq l \leq L \end{cases} \quad (7)$$

where $\mathbf{P}_i^{s0} = \mathbf{P}_i^c$. This stream ensures that the following sub-task prediction focuses on the region through a spatial attention mechanism, starting from the core pseudo prediction $\mathbf{P}_i^c$.

Backward Stream. After producing the L -th sub-task pseudo prediction $\mathbf{P}_i^{sL}$, to obtain the final core task prediction $\mathbf{T}_i^c$, we apply the backward stream as follows:

$$\begin{cases} \mathbf{T}_i^{sL} = \text{Up}_{5-i}(\mathbf{P}_i^{sL}) \\ \mathbf{T}_i^{sl} = \text{Up}_{5-i}(\mathbf{P}_i^{sl}) + \mathbf{T}_i^{s_{l+1}} \text{ for } 1 \leq l \leq L-1 \end{cases}, \quad (8)$$

where $\mathbf{T}_i^{s0} = \mathbf{T}_i^c$. To further analyze the Eq 8, we can recursively rewrite from core task $\mathbf{T}_i^c$ as follows:

Algorithm 1 Ensemble Sub-Decoding Module for Multi-task Learning with Deep Supervision

Input: Refined feature map $\mathbf{Y}_i$ from i -th MFMSA block

Output: Core task prediction $\mathbf{T}_i^c$ and sub-task predictions $\{\mathbf{T}_i^{s1}, \dots, \mathbf{T}_i^{sL}\}$ at i -th decoder

- 1: $\mathbf{P}_i^c = \text{Conv2D}_1(\mathbf{Y}_i)$
- 2: **for** $l = 1, 2, \dots, L$ **do**
- 3: $\mathbf{P}_i^{sl} = \text{Conv2D}_1(\mathbf{Y}_i \times \sigma(\mathbf{P}_i^{s_{l-1}}))$.
- 4: **end for**
- 5: $\mathbf{T}_i^{sL} = \text{Up}_{5-i}(\mathbf{P}_i^{sL})$
- 6: **for** $l = L-1, \dots, 0$ **do**
- 7: $\mathbf{T}_i^{sl} = \text{Up}_{5-i}(\mathbf{P}_i^{sl}) + \mathbf{T}_i^{s_{l+1}}$
- 8: **end for**
- 9: **return** $\mathbf{O}_i = \{\mathbf{T}_i^c, \mathbf{T}_i^{s1}, \dots, \mathbf{T}_i^{sL}\}$

$$\begin{aligned} \mathbf{T}_i^c &= \mathbf{T}_i^{s0} = \text{Up}_{5-i}(\mathbf{P}_i^{s0}) + \mathbf{T}_i^{s1} \\ &= [\text{Up}_{5-i}(\mathbf{P}_i^{s0}) + \text{Up}_{5-i}(\mathbf{P}_i^{s1})] + \mathbf{T}_i^{s2} = \dots \\ &= \sum_{l=0}^L \text{Up}_{5-i}(\mathbf{P}_i^{sl}) \end{aligned} \quad (9)$$

Consequently, E-SDM can be interpreted as an ensemble of predictions between different tasks for describing the same legion. Algorithm 1 describes the detailed training algorithm for E-SDM.

Loss Function. The loss function for multi-task learning with deep supervision in a ensemble manner is the same as that in a parallel manner, as follows:

$$\mathcal{L}_{total} = \sum_{i=1}^4 \sum_{t \in \{c, s_1, \dots, s_L\}} \lambda_t \mathcal{L}_t(\mathbf{G}^t, \text{Up}_{5-i}(\mathbf{T}_i^t)) \quad (10)$$

where $\mathbf{G}^t$ and $\mathbf{T}_i^t$ are the ground truth and predictions for task t and i -th decoder, respectively. Additionally, $\mathcal{L}_t$ and λ_t are the loss function and ratio for the task t , respectively.

For multi-task learning with deep supervision, we defined a core task as region R and two sub-tasks as boundary B and distance map D . The loss function for region prediction is defined as $\mathcal{L}_R = \mathcal{L}_{IoU}^w + \mathcal{L}_{bce}^w$, where $\mathcal{L}_{IoU}^w$ and $\mathcal{L}_{bce}^w$ are the weighted IoU and bce loss functions, respectively. Additionally, we defined boundary and distance map loss function $\mathcal{L}_B$ and $\mathcal{L}_D$ as bce and mse loss, respectively.

4. Experiment Results

4.1. Experiment Settings

In this study, we evaluated the models on seven medical image segmentation datasets with different modalities, including Dermoscopy [23], Radiology [32], Ultrasound [3],

Method	Dermoscopy		Radiology		Ultrasound		Microscopy		Colonoscopy				P-value
	ISIC2018 [23]		COVID19-1 [32]		BUSI [3]		DSB2018 [6]		CVC-ClinicDB [5]		Kvasir [30]		
	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	
UNet [51]	87.3 (0.8)	80.2 (0.7)	47.7 (0.6)	38.6 (0.6)	69.5 (0.3)	60.2 (0.2)	91.1 (0.2)	84.3 (0.3)	76.5 (0.8)	69.1 (0.9)	80.5 (0.3)	72.6 (0.4)	5.2E-06
AttUNet [44]	87.8 (0.1)	80.5 (0.1)	57.5 (0.2)	48.4 (0.2)	71.3 (0.4)	62.3 (0.6)	91.6 (0.1)	85.0 (0.1)	80.1 (0.6)	74.2 (0.5)	83.9 (0.1)	77.1 (0.1)	4.1E-06
UNet++ [73]	87.3 (0.2)	80.2 (0.1)	65.6 (0.7)	57.1 (0.8)	72.4 (0.1)	62.5 (0.2)	91.6 (0.1)	85.0 (0.1)	79.7 (0.2)	73.6 (0.4)	84.3 (0.3)	77.4 (0.2)	7.5E-07
CENet [22]	89.1 (0.2)	82.1 (0.1)	76.3 (0.4)	69.2 (0.5)	79.7 (0.6)	71.5 (0.5)	91.3 (0.1)	84.6 (0.1)	89.3 (0.3)	83.4 (0.2)	89.5 (0.7)	83.9 (0.7)	1.0E-05
TransUNet [7]	87.3 (0.2)	81.2 (0.8)	75.6 (0.4)	68.8 (0.2)	75.5 (0.5)	68.4 (0.1)	91.8 (0.3)	85.2 (0.2)	87.4 (0.2)	82.9 (0.1)	86.4 (0.4)	81.3 (0.4)	9.9E-08
FRCUNet [4]	88.9 (0.1)	83.1 (0.2)	77.3 (0.3)	70.4 (0.2)	81.2 (0.2)	73.3 (0.3)	90.8 (0.3)	83.8 (0.4)	91.8 (0.2)	87.0 (0.2)	88.8 (0.4)	83.5 (0.6)	6.6E-02
MSRFNet [56]	88.2 (0.2)	81.3 (0.2)	75.2 (0.4)	68.0 (0.4)	76.6 (0.7)	68.1 (0.7)	91.9 (0.1)	85.3 (0.1)	83.2 (0.9)	76.5 (1.1)	86.1 (0.5)	79.3 (0.4)	8.8E-07
HiFormer [26]	88.7 (0.5)	81.9 (0.5)	72.9 (1.4)	63.3 (1.5)	79.3 (0.2)	70.8 (0.1)	90.7 (0.2)	83.8 (0.4)	89.1 (0.6)	83.7 (0.6)	88.1 (1.0)	82.3 (1.2)	1.8E-05
DCSAUNet [66]	89.0 (0.3)	82.0 (0.3)	75.3 (0.4)	68.2 (0.4)	73.7 (0.5)	65.0 (0.5)	91.1 (0.2)	84.4 (0.2)	80.5 (1.2)	73.7 (1.1)	82.6 (0.5)	75.2 (0.5)	6.2E-07
M2SNet [72]	89.2 (0.2)	83.4 (0.2)	81.7 (0.4)	74.7 (0.5)	80.4 (0.8)	72.5 (0.7)	91.6 (0.2)	85.1 (0.3)	92.8 (0.8)	88.2 (0.8)	90.2 (0.5)	85.1 (0.6)	2.0E-05
SFSSNet	88.8 (0.3)	81.9 (0.2)	80.3 (0.8)	73.0 (0.7)	66.1 (0.6)	59.3 (0.8)	91.5 (0.2)	84.0 (0.2)	90.7 (0.4)	83.0 (0.7)	88.1 (0.6)	82.2 (0.7)	2.2E-06
MFSSNet	88.5 (0.2)	81.8 (0.2)	80.4 (0.7)	73.1 (0.4)	81.0 (0.1)	73.2 (0.2)	91.6 (0.1)	85.1 (0.2)	92.3 (0.5)	87.7 (0.5)	89.9 (0.6)	84.7 (0.7)	5.1E-07
SFMSNet	89.2 (0.3)	82.5 (0.3)	81.4 (0.3)	74.5 (0.3)	80.8 (0.4)	73.0 (0.3)	91.5 (0.2)	84.9 (0.4)	92.3 (0.3)	88.0 (0.3)	89.0 (0.6)	84.1 (0.5)	1.4E-04
MADGNet	90.2 (0.1)	83.7 (0.2)	83.7 (0.2)	76.8 (0.2)	81.3 (0.4)	73.4 (0.5)	92.0 (0.0)	85.5 (0.1)	93.9 (0.6)	89.5 (0.5)	90.7 (0.8)	85.3 (0.8)	-

Table 1. Segmentation results on five different modalities with *seen* clinical settings. We also provide one tailed *t*-Test results (*P*-value) compared to our method and other methods. (·) denotes the standard deviations of multiple experiment results.

Method	Dermoscopy		Radiology		Ultrasound		Microscopy		Colonoscopy				P-value		
	PH2 [41]		COVID19-2 [1]		STU [74]		MonuSeg2018 [12]		CVC-300 [61]		CVC-ColonDB [57]			ETIS [54]	
	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU		DSC	mIoU
UNet [51]	90.3 (0.1)	83.5 (0.1)	47.1 (0.7)	37.7 (0.6)	71.6 (1.0)	61.6 (0.7)	29.2 (5.1)	18.9 (3.5)	66.1 (2.3)	58.5 (2.1)	56.8 (1.3)	49.0 (1.2)	41.6 (1.1)	35.4 (1.0)	1.1E-09
AttUNet [44]	89.9 (0.2)	82.6 (0.3)	43.7 (0.8)	35.2 (0.8)	77.0 (1.6)	68.0 (1.7)	39.0 (3.1)	26.5 (2.4)	63.0 (0.3)	57.2 (0.4)	56.8 (1.6)	50.0 (1.5)	38.4 (0.3)	33.5 (0.1)	6.7E-09
UNet++ [73]	88.0 (0.3)	80.1 (0.3)	50.5 (3.8)	40.9 (3.7)	77.3 (0.4)	67.8 (0.3)	25.4 (0.8)	15.3 (0.5)	64.3 (2.2)	58.4 (2.0)	57.5 (0.4)	50.2 (0.4)	39.1 (2.4)	34.0 (2.1)	1.0E-05
CENet [22]	90.5 (0.1)	83.3 (0.1)	60.1 (0.3)	49.9 (0.3)	86.0 (0.7)	77.2 (0.9)	27.7 (1.5)	16.9 (1.0)	85.4 (1.6)	78.2 (1.4)	65.9 (1.6)	59.2 (0.1)	57.0 (3.4)	51.4 (0.5)	4.5E-06
TransUNet [7]	89.5 (0.3)	82.1 (0.4)	56.9 (1.0)	48.0 (0.7)	41.4 (9.5)	32.1 (4.2)	15.9 (8.5)	9.6 (5.5)	85.0 (0.6)	77.3 (0.3)	63.7 (0.1)	58.4 (0.3)	50.1 (0.5)	44.0 (2.3)	1.6E-06
FRCUNet [4]	90.6 (0.1)	83.4 (0.2)	62.9 (1.1)	52.7 (0.9)	86.5 (2.3)	77.2 (2.7)	26.1 (5.6)	16.8 (4.3)	86.7 (0.7)	79.4 (0.3)	69.1 (1.0)	62.6 (0.9)	65.1 (1.0)	58.4 (0.5)	2.3E-05
MSRFNet [56]	90.5 (0.3)	83.5 (0.3)	58.3 (0.8)	48.4 (0.6)	84.0 (5.5)	75.2 (8.2)	9.1 (1.0)	5.3 (0.7)	72.3 (2.2)	65.4 (2.2)	61.5 (1.0)	54.8 (0.8)	38.3 (0.6)	33.7 (0.7)	1.0E-07
HiFormer [26]	86.9 (1.6)	79.1 (1.8)	54.1 (1.0)	44.5 (0.8)	80.7 (2.9)	71.3 (3.2)	21.9 (8.9)	13.2 (5.7)	84.7 (1.1)	77.5 (1.1)	67.6 (1.4)	60.5 (1.3)	56.7 (3.2)	50.1 (3.3)	2.5E-07
DCSAUNet [66]	89.0 (0.4)	81.5 (0.3)	52.4 (1.2)	44.0 (0.7)	86.1 (0.5)	76.5 (0.8)	4.3 (0.3)	2.4 (0.9)	68.9 (4.0)	59.8 (3.9)	57.8 (0.4)	49.3 (0.4)	42.9 (3.0)	36.1 (2.9)	1.3E-07
M2SNet [72]	90.7 (0.3)	83.5 (0.5)	68.6 (0.1)	58.9 (0.2)	79.4 (0.7)	69.3 (0.6)	36.3 (0.9)	23.1 (0.8)	89.9 (0.2)	83.2 (0.3)	75.8 (0.7)	68.5 (0.5)	74.9 (1.3)	67.8 (1.4)	4.9E-02
SFSSNet	89.8 (0.2)	82.2 (0.4)	65.1 (1.6)	55.5 (1.3)	59.1 (0.3)	49.3 (0.7)	21.5 (7.2)	14.3 (5.0)	81.7 (0.3)	74.7 (0.4)	65.6 (0.4)	58.4 (0.5)	56.4 (0.7)	49.4 (0.4)	2.0E-07
MFSSNet	90.2 (0.8)	83.3 (0.9)	67.6 (0.5)	57.9 (0.3)	66.1 (0.8)	59.3 (0.2)	30.1 (7.5)	20.5 (5.5)	83.3 (1.4)	76.1 (1.2)	66.0 (0.7)	59.1 (0.8)	59.3 (0.2)	52.6 (0.6)	3.9E-04
SFMSNet	90.8 (0.3)	83.9 (0.5)	67.7 (1.1)	58.0 (1.3)	84.5 (0.2)	74.3 (0.1)	28.1 (9.9)	18.2 (7.1)	84.2 (1.2)	78.1 (1.0)	75.9 (0.8)	68.3 (0.8)	68.9 (0.3)	62.7 (0.4)	7.9E-03
MADGNet	91.3 (0.1)	84.6 (0.1)	72.2 (0.3)	62.6 (0.3)	88.4 (1.0)	79.9 (1.5)	46.7 (4.3)	32.0 (2.9)	87.4 (0.4)	79.9 (0.4)	77.5 (1.1)	69.7 (1.2)	77.0 (0.3)	69.7 (0.5)	-

Table 2. Segmentation results on five different modalities with *unseen* clinical settings. We also provide one tailed *t*-Test results (*P*-value) compared to our method and other methods. (·) denotes the standard deviations of multiple experiment results.

Microscopy [6], Colonoscopy [5, 30], and Fundus Imaging [45], to validate the modality-agnostic ability. Moreover, we evaluated the domain generalizability on eight external datasets with different modalities, including Dermoscopy [41], Radiology [1], Ultrasound [74], Microscopy [12], Colonoscopy [54, 57, 61], and Fundus Imaging [55]. To evaluate the performance of each model, we selected two metrics, DSC and mIoU, which are widely used in medical image segmentation. We compared the **MADGNet (Ours)** with ten medical image segmentation models, including UNet [51], Attention UNet (AttUNet [44]), UNet++ [73], CENet [22], TransUNet [7], FRCUNet [4], MSRFNet [56], HiFormer [26], DCSAUNet [66], and M2SNet [72]. We report the *mean* performance of three trials for all results. **Red** and **Blue** are the first and second best performance results, respectively. Due to the page limit, we present the detailed dataset description and results using other metrics [14, 15, 40] in Appendix 6 and 12.

4.2. Implementation Details

Training Settings. We implemented MADGNet on a single NVIDIA RTX 3090 Ti in Pytorch 1.11 [48]. We started

with an initial learning rate of 10^{-4} using the Adam optimizer and reduced the parameters of each model to 10^{-6} using a cosine annealing learning rate scheduler [38]. We optimized each model with a batch size of 16 and epochs of 50, 100, 100, 100, and 200 for Colonoscopy, Dermoscopy, Microscopy, Ultrasound, and Radiology modalities, respectively. During training, we used horizontal/vertical flipping, with a probability of 50%, and rotation between -5° and 5° based on the multi-scale training strategy, which is widely used in medical image segmentation [16, 71, 72]. At this stage, because images in each dataset have different resolutions, all images were resized to 352×352 .

Hyperparameters of MADGNet. Key hyperparameters for MADGNet on all datasets were set to $C_e = 64$, $C_{\min} = 32$, $H_{\min} = W_{\min} = 8$, $\gamma = \frac{1}{2}$, $r = 16$ for efficiency and $S = 3$ and $K = 16$ for combining multi-scale and multi-frequency features. The aggregation function $\mathbf{A}(\cdot)$ averaged refined feature maps from each scale branch. To reduce the sensitivity of hyperparameters, the loss ratio between tasks was fixed as $\lambda_t = 1$ for all tasks $t \in \{R, D, B\}$. In this paper, the core task indicates the region prediction, while sub-tasks indicate the distance map and boundary pre-

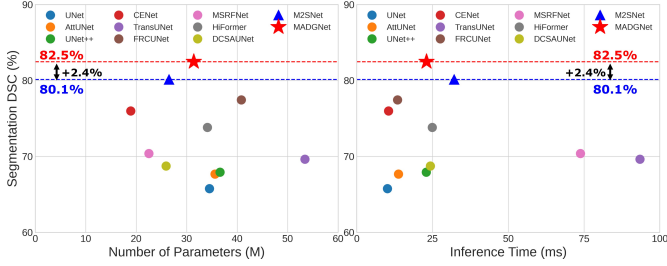


Figure 4. Comparison of parameters (M) and inference speed (ms) vs segmentation performance (DSC) on average for all datasets.

dictions. In Appendix 9.2, we provide the experiment results on various hyperparameter settings.

4.3. Comparison with State-of-the-art models

Binary Segmentation. As listed in Tab. 1 and 2, MADGNet achieved the highest segmentation performance on various modalities and clinical settings. In particular, compared to M2SNet, which achieved the second highest segmentation performance in most modalities in Tab. 1, MADGNet improved the DSC and mIoU by 1.1% and 1.0% on average, respectively. Additionally, when compared to FRCUNet, which uses multi-frequency attention, MADGNet improved the both DSC and mIoU by 2.0% on average. To evaluate the domain generalizability of the model trained in each modality, we tested each model on external dataset from unseen clinical settings. As a result, MADGNet improved its performance by a higher margin than M2SNet, except for CVC-300. In addition, when compared FRCUNet and MADGNet on BUSI, the performance gap was tight. Nevertheless, MADGNet exhibits significant improvement of 1.9% and 2.7% on DSC and mIoU on STU, respectively. These results indicate that other models, which do not consider scale and frequency dimensions simultaneously, cannot comprehend intricate anatomical knowledge. Fig. 4 indicates that MADGNet contains almost 31M parameters with reasonable inference speed (0.024 sec/image), which has obvious advantages in terms of computational efficiency.

Fig. 5 illustrates the qualitative results on various modalities. The UNet relies solely on skip connection and progressive decoding; therefore, it produces noisy predictions. Such unreliable predictions are also observed in other models, such as AttUNet and UNet++, which do not consider multi-scale and multi-frequency features. And the other two transformer-based models, TransUNet and HiFormer, cannot segment in low-contrast Dermoscopy due to lack of inductive bias. As CENet, MSRFNet, DCSAUNet, and M2SNet utilize multi-scale information, they depict detailed boundary prediction; however, they struggle with severe noise in images from Radiology, Colonoscopy, and Ultrasound. Additionally, FRCUNet depicts detailed bound-

Method	REFUGE [45] (<i>seen</i>)		Drishti-GS [55] (<i>unseen</i>)	
	OD	OC	OD	OC
UNet [51]	79.9 (0.9)	79.2 (0.8)	62.2 (1.1)	38.6 (1.2)
AttUNet [44]	80.8 (0.1)	79.4 (0.2)	72.6 (2.6)	73.8 (1.4)
UNet++ [73]	80.6 (0.1)	79.5 (0.2)	76.1 (1.4)	71.3 (1.4)
CENet [22]	80.4 (0.1)	74.1 (0.4)	87.6 (0.6)	78.8 (1.2)
TransUNet [7]	81.3 (0.2)	40.7 (0.2)	76.0 (3.2)	38.0 (3.1)
FRCUNet [4]	84.1 (1.3)	45.2 (0.1)	87.2 (1.1)	44.4 (0.1)
MSRFNet [56]	83.6 (0.6)	81.3 (0.7)	71.3 (0.8)	32.3 (0.9)
HiFormer [26]	80.5 (0.1)	75.3 (0.4)	79.9 (1.3)	68.6 (2.5)
DCSAUNet [66]	81.2 (0.2)	59.6 (0.5)	53.3 (3.7)	28.3 (3.6)
M2SNet [72]	81.0 (1.1)	60.1 (1.2)	84.5 (3.3)	69.4 (1.2)
MADGNet	84.9 (0.5)	78.5 (0.4)	88.8 (0.3)	83.6 (0.4)

Table 3. Segmentation results between methods on two Fundus Image segmentation datasets (REFUGE [45] and Drishti-GS [55]).

ary on Ultrasound, which contains high-frequency images; but, has difficulties in predicting in Radiology and Dermoscopy. Despite this severe noise and lesions of various sizes, MADGNet successfully depicts detailed boundaries for all modalities due to the dual utilization of multi-scale and multi-frequency information.

Multi-label Segmentation. For medical image analysis, certain medical image segmentation datasets contain multi-label objects that need to be segmented. To satisfy this demand, we evaluated all models on two Fundus Imaging datasets, including REFUGE [45] (*seen*) and Drishti-GS [55] (*unseen*), with different labels, Optic Disk (OD) and Optic Cup (OC). Except for data augmentation, we trained all models for 200 epochs with the same setting presented in Section 4.2. We cropped 512×512 ROIs centering OD and applied data augmentations, which are the same settings as those reported in the literature [63].

As listed in Tab. 3, compared to CENet, MADGNet improved the average DSC of OD and OC by 2.9% and 1.9%, respectively. This result indicates that our model can be generalized to a multi-label segmentation task compared to other methods. The last row in Fig. 5 presents the qualitative results for each method.

4.4. Ablation Study on MADGNet

To demonstrate the effectiveness of MADGNet, we conducted ablation studies on the MSMFA block and E-SDM.

Ablation Study on MFMSA Block. To demonstrate that multi-scale and multi-frequency features are crucial in building the medical image segmentation model, we compared three variant models of MADGNet; Single-Frequency in Single-Scale Network (SFSSNet), Multi-Frequency in Single-Scale Network (MFSSNet), Single-Frequency in Multi-Scale Network (SFMSNet). These variants differ in (S, K) ; SFSSNet, MFSSNet, and SFMSNet were set to $(1, 1)$, $(1, 16)$, and $(3, 1)$, respectively. As listed in Tab. 1 and 2, as SFSSNet suffers from capturing the various sizes and detailed features in medical images, this approach has demonstrated unsatisfactory results in multiple clinical settings and modalities. The performance gain is

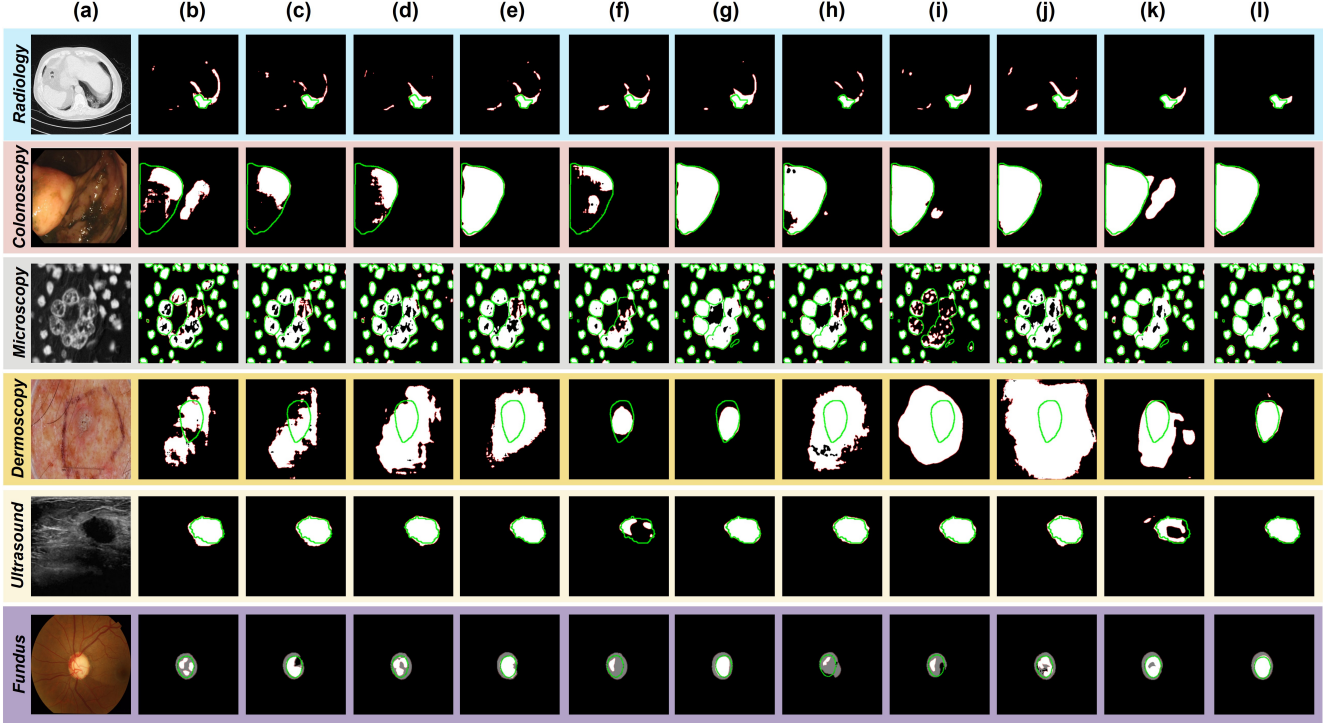


Figure 5. Qualitative comparison of other methods and MADGNet. (a) Input images. (b) UNet [51]. (c) AttUNet [44]. (d) UNet++ [73]. (e) CENet [22]. (f) TransUNet [7]. (g) FRCUNet [4]. (h) MSRFNet [56]. (i) HiFormer [26]. (j) DCSAUNet [66]. (k) M2SNet [72]. (l) MADGNet (Ours). **Green** and **Red** lines denote the boundaries of the ground truth and prediction, respectively.

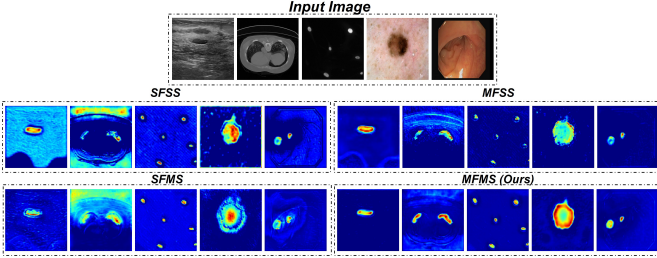


Figure 6. Feature visualization of SFSS, MFSS, SFMS, **MFMS**.

achieved by utilizing multi-scale and multi-frequency features when these features are applied separately. This experimental evidence demonstrates the critical significance of both features in medical image segmentation, as illustrated in Fig. 1. Moreover, we observed that using both multi-scale and multi-frequency significantly improves the feature representation power and extracts enhanced boundary features, as illustrated in Fig. 6. Consequently, MADGNet simultaneously employs both features, resulting in significantly better performance across various modalities and clinical settings due to the enriched feature representation robust to noise and scale.

Ablation Study on E-SDM. We performed ablation studies

DS	Flow	Task	Seen		Unseen	
			DSC	mIoU	DSC	mIoU
✗	-	R	90.8	85.7	75.2	68.2
	Parallel	$R \& D \& B$	91.5	86.6	76.2	69.9
✓	Parallel	$R \& D \& B$	90.8	85.9	73.7	66.8
	Ensemble	$R \rightarrow D \rightarrow B$	91.4	86.5	77.5	70.0
	Ensemble	$R \leftrightarrow D \leftrightarrow B$	92.0	87.3	80.9	73.3

Table 4. Ablation study of E-SDM on the *seen* ([5, 30]) and *unseen* ([54, 57, 61]) datasets on Colonoscopy. DS denotes Deep Supervision. R , D , B are region, distance map, and boundary task, respectively. $\rightarrow$ and $\leftrightarrow$ denote E-SDM without and with backward stream, respectively.

to demonstrate the effectiveness of E-SDM on Colonoscopy images. As listed in Tab. 4, our experimentation with multi-task learning and deep supervision in a parallel manner revealed a potential issue of performance degradation. This issue can be addressed by tackling the information loss due to drastic upsampling during deep supervision training. For this reason, when we trained a model in an ensemble manner, we observed that E-SDM had higher DSC of 1.2% and 7.2% on seen and unseen datasets, comparing two decoding flows (Parallel and Ensemble) with deep supervision, respectively. By using the ensemble method, we can overcome these challenges and obtain better results by leveraging an ensemble approach that combines the predictions

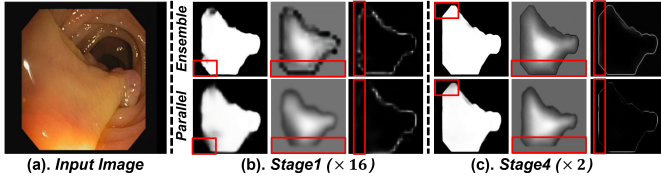


Figure 7. Qualitative results between ensemble and parallel manners. (a) Input Image, (b) and (c) Predictions from Stage1 ($\times 16$ Up) and Stage4 ($\times 2$ Up). First and second rows in (b) and (c) are predictions with **ensemble (Ours)** and parallel manners.

from different tasks. This process enables us to identify complex boundaries B and structures D of the lesion R , resulting in higher performance compared to the parallel approach. Fig. 7 illustrates the qualitative results of ensemble and parallel manners. E-SDM predict more reliably than in a parallel manner, even though it exhibits a high upsampling rate ($\times 16$). The ensemble method also depicts the detailed boundary and distance map predictions than in a parallel manner. Finally, we examined which forward and backward streams had a larger impact on segmentation performance. Our study showed that the backward stream is crucial for maintaining the overall output quality as it compensates for the information loss caused by upsampling. While the forward stream only identifies areas requiring emphasis, the backward stream preserves critical information.

5. Conclusion

Based on the outcomes of extensive experiments on various modalities and clinical settings, we can summarize the effectiveness of MADGNet into three key aspects: 1) For the design of medical image segmentation, multi-scale and multi-frequency features prove pivotal. 2) MFCA’s adeptness at extracting discriminative features from noisy medical image feature maps translates into the acquisition of more sophisticated boundary cues through the MSSA. 3) E-SDM compensates for information loss caused by drastic upsampling when performing multi-task learning with deep supervision through an ensemble of tasks. Furthermore, our method demonstrated outstanding performance in segmenting OD and OC not only in binary classification but also in multi-label segmentation. In conclusion, we propose a novel medical image segmentation model called MADGNet, which can be utilized in various modalities and clinical settings. It includes two key components: the MFMSA block and E-SDM, which extract distinctive features and compensate for information loss during multi-task learning with deep supervision. Through rigorous experiment, we discovered that MADGNet is a highly effective model that surpasses other state-of-the-art options regarding segmentation performance. Furthermore, we will focus

on increasing the memory-efficiency for real clinical use.

Acknowledgements

This work was supported in part by the National Research Foundation of Korea (NRF) under Grant NRF-2021R1A2C2010893 and in part by Institute of Information and communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-00155915, Artificial Intelligence Convergence Innovation Human Resources Development (Inha University)).

References

- [1] Covid19 dataset. <https://www.kaggle.com/datasets/piyushsamant11/pidata-new-names>. 6, 1, 3, 4, 5, 7
- [2] Nasir Ahmed, T. Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1): 90–93, 1974. 2
- [3] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020. 5, 6, 1, 3, 4, 8, 11
- [4] Reza Azad, Afshin Bozorgpour, Maryam Asadi-Aghbolaghi, Dorit Merhof, and Sergio Escalera. Deep frequency recalibration u-net for medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3274–3283, 2021. 2, 6, 7, 8, 1, 5, 9, 10, 11
- [5] 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. 6, 8, 1, 3, 4, 10, 11
- [6] 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, 1, 3, 4, 9, 11
- [7] 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. 6, 7, 8, 1, 5, 9, 10, 11
- [8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 2
- [9] Xu Chen, Bryan M Williams, Srinivasa R Vallabhaneni, Gabriela Czanner, Rachel Williams, and Yalin Zheng. Learning active contour models for medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11632–11640, 2019. 1

- [10] Alan S Coates, Eric P Winer, Aron Goldhirsch, Richard D Gelber, Michael Gnant, M Piccart-Gebhart, Beat Thürlimann, H-J Senn, Panel Members, Fabrice André, et al. Tailoring therapies—improving the management of early breast cancer: St gallen international expert consensus on the primary therapy of early breast cancer 2015. *Annals of oncology*, 26(8):1533–1546, 2015. 1
- [11] Runmin Cong, Yumo Zhang, Ning Yang, Haisheng Li, Xueqi Zhang, Ruochen Li, Zewen Chen, Yao Zhao, and Sam Kwong. Boundary guided semantic learning for real-time covid-19 lung infection segmentation system. *IEEE Transactions on Consumer Electronics*, 68(4):376–386, 2022. 3
- [12] Tuan Le Dinh, Seong-Geun Kwon, Suk-Hwan Lee, and Ki-Ryong Kwon. Breast tumor cell nuclei segmentation in histopathology images using efficientnet++ and multi-organ transfer learning. *Journal of Korea Multimedia Society*, 24(8):1000–1011, 2021. 6, 1, 3, 4, 9, 11
- [13] 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
- [14] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, pages 4548–4557, 2017. 6
- [15] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*, 2018. 6
- [16] 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. 3, 5, 6
- [17] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE transactions on medical imaging*, 39(8):2626–2637, 2020. 3, 5
- [18] Shuanglang Feng, Heming Zhao, Fei Shi, Xuena Cheng, Meng Wang, Yuhui Ma, Dehui Xiang, Weifang Zhu, and Xinjian Chen. Cpfnet: Context pyramid fusion network for medical image segmentation. *IEEE transactions on medical imaging*, 39(10):3008–3018, 2020. 2
- [19] Borko Furht, editor. *Discrete Wavelet Transform (DWT)*, pages 205–207. Springer US, Boston, MA, 2006. 2
- [20] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):652–662, 2019. 3
- [21] A-Rom Gu, Ju-Hyeon Nam, and Sang-Chul Lee. Fbi-net: Frequency-based image forgery localization via multitask learning with self-attention. *IEEE Access*, 10:62751–62762, 2022. 4
- [22] Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging*, 38:2281–2292, 2019. 2, 6, 7, 8, 1, 5, 9, 10, 11
- [23] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016. 5, 6, 1, 3, 4
- [24] Robert M Haralick, Stanley R Sternberg, and Xinhua Zhuang. Image analysis using mathematical morphology. *IEEE transactions on pattern analysis and machine intelligence*, (4):532–550, 1987. 1
- [25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3
- [26] Moein Heidari, Amirhossein Kazerouni, Milad Soltany, Reza Azad, Ehsan Khodapanah Aghdam, Julien Cohen-Adad, and Dorit Merhof. Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6202–6212, 2023. 6, 7, 8, 1, 5, 9, 10, 11
- [27] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 2
- [28] 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 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1055–1059. IEEE, 2020. 3
- [29] W. Jenkins and M. Desai. The discrete frequency fourier transform. *IEEE Transactions on Circuits and Systems*, 33(7):732–734, 1986. 2
- [30] 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 *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pages 451–462. Springer, 2020. 6, 8, 1, 3, 4, 10, 11
- [31] Yuzhu Ji, Haijun Zhang, and QM Jonathan Wu. Salient object detection via multi-scale attention cnn. *Neurocomputing*, 322:130–140, 2018. 2
- [32] Ma Jun, Ge Cheng, Wang Yixin, An Xingle, Gao Jiantao, Yu Ziqi, Zhang Mingqing, Liu Xin, Deng Xueyuan, Cao Shucheng, Wei Hao, Mei Sen, Yang Xiaoyu, Nie Ziwei, Li Chen, Tian Lu, Zhu Yuntao, Zhu Qiongjie, Dong Guoqiang, and He Jian. COVID-19 CT Lung and Infection Segmentation Dataset. 2020. 5, 6, 1, 3, 4, 7
- [33] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988. 1

- [34] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. [2](#)
- [35] et al. Li, Jia. Finding the secret of image saliency in the frequency domain. *IEEE transactions on pattern analysis and machine intelligence*, 2015. [4](#)
- [36] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. [2](#)
- [37] Songtao Liu, Di Huang, et al. Receptive field block net for accurate and fast object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 385–400, 2018. [2](#)
- [38] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. [6](#)
- [39] Stephane G Mallat. Multifrequency channel decompositions of images and wavelet models. *IEEE Transactions on Acoustics, speech, and signal processing*, 37(12):2091–2110, 1989. [3](#)
- [40] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 248–255, 2014. [6](#)
- [41] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. Ph 2-a dermoscopic image database for research and benchmarking. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 5437–5440. IEEE, 2013. [6](#), [1](#), [3](#), [4](#), [5](#)
- [42] Ju-Hyeon Nam, Seo-Hyeong Park, Nur Suriza Syazwany, Yerim Jung, Yu-Han Im, and Sang-Chul Lee. M3fpolypsegnet: Segmentation network with multi-frequency feature fusion for polyp localization in colonoscopy images. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1530–1534. IEEE, 2023. [3](#)
- [43] Shin’ya Nishida, Timothy Ledgeway, and Mark Edwards. Dual multiple-scale processing for motion in the human visual system. *Vision research*, 37(19):2685–2698, 1997. [3](#)
- [44] 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* 2018. *arXiv preprint arXiv:1804.03999*, 1804. [2](#), [6](#), [7](#), [8](#), [1](#), [5](#), [9](#), [10](#), [11](#)
- [45] José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel Van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59:101570, 2020. [6](#), [7](#), [1](#)
- [46] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979. [1](#)
- [47] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018. [2](#)
- [48] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. [6](#)
- [49] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 783–792, 2021. [2](#), [4](#)
- [50] Daniel Riccio, Nadia Brancati, Maria Frucci, and Diego Gragnaniello. A new unsupervised approach for segmenting and counting cells in high-throughput microscopy image sets. *IEEE journal of biomedical and health informatics*, 23(1):437–448, 2018. [1](#)
- [51] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. [2](#), [6](#), [7](#), [8](#), [1](#), [5](#), [9](#), [10](#), [11](#)
- [52] Abhinav Sagar. Dmsanet: Dual multi scale attention network. In *International Conference on Image Analysis and Processing*, pages 633–645. Springer, 2022. [2](#)
- [53] Mufan Sang and John HL Hansen. Multi-frequency information enhanced channel attention module for speaker representation learning. *arXiv preprint arXiv:2207.04540*, 2022. [4](#)
- [54] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9:283–293, 2014. [6](#), [8](#), [1](#), [3](#), [4](#), [10](#), [11](#)
- [55] Jayanthi Sivaswamy, S Krishnadas, Arunava Chakravarty, G Joshi, A Syed Tabish, et al. A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. *JSM Biomedical Imaging Data Papers*, 2(1):1004, 2015. [6](#), [7](#), [1](#)
- [56] Abhishek Srivastava, Debesh Jha, Sukalpa Chanda, Umada Pal, Håvard D Johansen, Dag Johansen, Michael A Riegler, Sharib Ali, and Pål Halvorsen. Msrf-net: a multi-scale residual fusion network for biomedical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26(5):2252–2263, 2021. [2](#), [6](#), [7](#), [8](#), [1](#), [5](#), [9](#), [10](#), [11](#)
- [57] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015. [6](#), [8](#), [1](#), [3](#), [4](#), [10](#), [11](#)

- [58] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020. 2
- [59] Hamid R Tizhoosh. Image thresholding using type ii fuzzy sets. *Pattern recognition*, 38(12):2363–2372, 2005. 1
- [60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2
- [61] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdal, Aaron Courville, et al. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017, 2017. 6, 8, 1, 3, 4, 10, 11
- [62] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017. 2
- [63] Shujun Wang, Lequan Yu, Kang Li, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Boundary and entropy-driven adversarial learning for fundus image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, pages 102–110. Springer, 2019. 7
- [64] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 2
- [65] et al. Xu, Kai. Learning in the frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020. 4
- [66] Qing Xu, Zhicheng Ma, HE Na, and Wenting Duan. Dcsau-net: A deeper and more compact split-attention u-net for medical image segmentation. *Computers in Biology and Medicine*, 154:106626, 2023. 2, 6, 7, 8, 1, 5, 9, 10, 11
- [67] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3684–3692, 2018. 2
- [68] Tongda Yang, Weiming Wang, Gary Cheng, Mingqiang Wei, Haoran Xie, and Fu Lee Wang. Fddl-net: frequency domain decomposition learning for speckle reduction in ultrasound images. *Multimedia Tools and Applications*, 81(29):42769–42781, 2022. 2
- [69] Michael Yeung, Evis Sala, Carola-Bibiane Schönlieb, and Leonardo Rundo. Focus u-net: A novel dual attention-gated cnn for polyp segmentation during colonoscopy. *Computers in biology and medicine*, 137:104815, 2021. 2
- [70] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2736–2746, 2022. 3
- [71] Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Automatic polyp segmentation via multi-scale subtraction network. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 120–130. Springer, 2021. 6
- [72] Xiaoqi Zhao, Hongpeng Jia, Youwei Pang, Long Lv, Feng Tian, Lihe Zhang, Weibing Sun, and Huchuan Lu. M²SNet: Multi-scale in multi-scale subtraction network for medical image segmentation. *arXiv preprint arXiv:2303.10894*, 2023. 5, 6, 7, 8, 1, 9, 10, 11
- [73] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018. 2, 3, 6, 7, 8, 1, 5, 9, 10, 11
- [74] Zheming Zhuang, Nan Li, Alex Noel Joseph Raj, Vijayalakshmi GV Mahesh, and Shunmin Qiu. An rdau-net model for lesion segmentation in breast ultrasound images. *PloS one*, 14(8):e0221535, 2019. 6, 1, 3, 4, 8, 11

Modality-agnostic Domain Generalizable Medical Image Segmentation by Multi-Frequency in Multi-Scale Attention

Supplementary Material

6. Dataset Descriptions

Dataset	Modality	Images	Resolutions	Train	Valid	Test
ISIC2018 [23]	Dermoscopy	2594	Variable	1868	465	261
COVID19-1 [32]	Radiology	1277	512×512	643	251	383
BUSI [3]	Ultrasound	645	Variable	324	160	161
2018 Data Science Bowl [6]	Microscopy	670	Variable	483	120	67
CVC-ClinicDB [5]	Colonoscopy	612	384×288	490	60	62
Kvasir-SEG [30]	Colonoscopy	1000	Variable	800	100	100
REFUGE [45]	Fundus Image	400	2124×2056	280	40	80

Table 5. Details of the medical segmentation *seen* clinical settings used in our experiments.

Dataset	Modality	Images	Resolutions	Test
PH2 [41]	Dermoscopy	200	767×576	200
COVID19-2 [1]	Radiology	2535	512×512	2535
STU [74]	Ultrasound	42	Variable	42
MonuSeg2018 [12]	Microscopy	82	256×256	82
CVC-300 [61]	Colonoscopy	60	574×500	60
CVC-ColonDB [57]	Colonoscopy	380	574×500	380
ETIS [54]	Colonoscopy	196	1255×966	196
Drishti-GS [55]	Fundus Image	50	Variable	50

Table 6. Details of the medical segmentation *unseen* clinical settings used in our experiments.

- **Breast Ultrasound Segmentation:** The BUSI [3] comprises 780 images from 600 female patients, including 133 normal cases, 437 benign cases, and 210 malignant tumors. On the other hand, the STU [74] includes only 42 breast ultrasound images collected by Shantou University. Due to the limited number of images in the STU, it is used only to evaluate the generalizability of each model across different datasets.
- **Skin Lesion Segmentation:** The ISIC 2018 [23] comprises 2,594 images with various sizes. We randomly selected train, validation, and test images with 1,868, 465, and 261, respectively. And, we used PH2 [41] to evaluate the domain generalizability of each model. Note that ISIC2018 [23] and PH2 [41] are *seen*, and *unseen* clinical settings, respectively.
- **COVID19 Lung Infection Segmentation:** COVID19-1 [32] comprises 1,277 high-quality CT images. We randomly selected train, validation, and test images with 643, 251, and 383, respectively. And, we used COVID19-2 [1] to evaluate the domain generalizability of each model. Note that COVID19-2 [1] is used for only testing.

- **Cell Segmentation:** The 2018 Data Science Bowl dataset [6] comprises 670 microscopy images. The dataset consisted of training, validate, and test images with 483, 120, and 67, respectively. We also used MonuSeg2018 [12] for evaluating the domain generalizability of each model. Note that MonuSeg2018 [12] is used for only testing.
- **Polyp Segmentation:** Colorectal cancer is the third most prevalent cancer globally and the second most common cause of death. It typically originates as small, non-cancerous (benign) clusters of cells known as polyps, which develop inside the colon. To evaluate the proposed model, we have used five benchmark datasets, namely CVC-ColonDB [57], ETIS [54], Kvasir [30], CVC-300 [61], and CVC-ClinicDB [5]. The same training set as the latest image polyp segmentation method has been adopted, consisting of 900 samples from Kvasir and 550 samples from CVC-ClinicDB for training. The remaining images and the other three datasets are used for only testing.
- **Fundus Image Segmentation:** To evaluate our method on multi-label segmentation, we utilize the training part of the REFUGE challenge dataset [45] as the training (280) and testing (80) dataset, and the public Drishti-GS [55] dataset as the testing (50) dataset.

7. Efficiency Analysis

Method	Parameters (M)	inference speed (ms)
UNet [51]	34.5	10.1
AttUNet [44]	35.6	13.8
UNet++ [73]	36.6	22.9
CENet [22]	18.9	10.5
TransUNet [7]	53.4	93.4
FRCUNet [4]	40.8	13.4
MSRFNet [56]	22.5	73.8
HiFormer [26]	34.1	24.9
DCSAUNet [66]	25.9	24.3
M2SNet [72]	26.5	32.1
MADGNet	31.4	24.0

Table 7. The number of parameters (M) and inference speed (ms) of different models.

7.1. Parameter Efficiency Proof

Scale Decomposition. For further efficiency, we replace the conventional convolution with kernel size $2s + 1$ by dilated

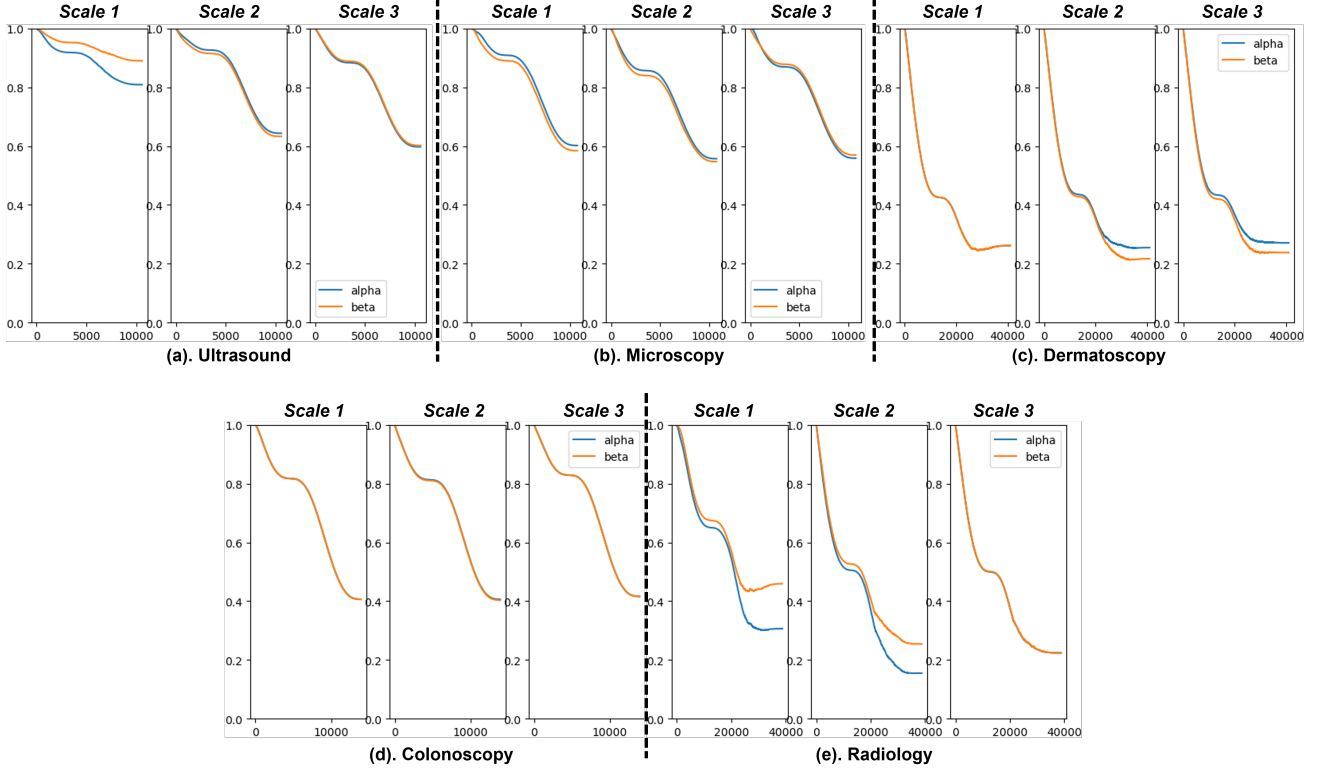


Figure 8. The training results of α_3^s and β_3^s for each modalities ((a) Ultrasound, (b) Microscopy, (c) Dermatology, (d) Colonoscopy, and (e) Radiology) where $s \in \{1, 2, 3\}$.

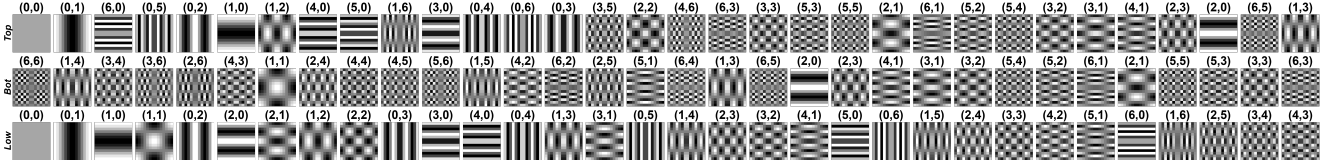


Figure 9. Frequency selection strategies (Top, Bot, Low) [49]. (u_k, v_k) denotes the frequency indices according to frequency selection strategy.

convolution with a kernel size of 3 and dilation size of s . For instance, in Fig. 2. (b), the convolution with kernel size of 5 and 7 in second and third scale branch are replaced into the dilated convolution with a kernel size of 3 and dilation size of 2 and 3, respectively. This replacement can achieve a parameter reduction of $9/(2s+1)^2$ in each scale branch. Suppose that the number of channel at i -th MFMSA block is C . Then, by channel reduction ratio $\gamma \in (0, 1)$, the number of parameters for each scale branch is $9 \times C^2 \gamma^{s-1}$.

MFCA. Note that MFCA contains two fully-connected layer $\mathbf{W}_1 \in \mathbb{R}^{C\gamma^{s-1} \times \frac{C\gamma^{s-1}}{r}}$ and $\mathbf{W}_2 \in \mathbb{R}^{\frac{C\gamma^{s-1}}{r} \times C\gamma^{s-1}}$ with reduction ratio r . Then, the number of parameters in MFCA is $C\gamma^{s-1} \times C\gamma^{s-1} \times \frac{1}{r} \times 2$.

MSSA. Firstly, to extract attention map from frequency-recalibrated feature map $\hat{\mathbf{X}}_i^s$, we apply a 2D convolution

operation with kernel size 1. And then, to aggregate each refined features from different scale branches, we restore the number of channels into C . Hence, the number of parameters in MSSA is $C\gamma^{s-1} + 9 \times C^2 \gamma^{s-1}$.

MFMSA block. Then, we can approximate the number of parameters at s -th scale branch as follows:

$$\begin{aligned} p_s &= 9 \times C^2 \gamma^{s-1} + C\gamma^{s-1} \times C\gamma^{s-1} \times \frac{1}{r} \times 2 \\ &\quad + C\gamma^{s-1} + 9 \times C^2 \gamma^{s-1} \\ &= C\gamma^{s-1} \left(18C + C\gamma^{s-1} \times \frac{2}{r} + 1 \right) \end{aligned} \quad (11)$$

If we do not introduce the channel reduction ratio γ , then the number of parameters at each scale is $p = C(18C +$

Algorithm 2 Ensemble Sub-Decoding Module for Multi-task Learning with Deep Supervision in Multi-label Segmentation

Input: Refined feature map $\mathbf{Y}_i$ from i -th MFMSA block

Output: Core task prediction $\mathbf{T}_i^{c,m}$ and sub-task predictions $\{\mathbf{T}_i^{s_1,m}, \dots, \mathbf{T}_i^{s_L,m}\}$ at i -th decoder for each m -th label.

```

1: for  $m = 1, 2, \dots, M$  do
2:    $\mathbf{P}_i^{c,m} = \text{Conv2D}_1(\mathbf{Y}_i)$ 
3:   for  $l = 1, 2, \dots, L$  do
4:      $\mathbf{P}_i^{s_l,m} = \text{Conv2D}_1(\mathbf{Y}_i \times \sigma(\mathbf{P}_i^{s_{l-1},m}))$ 
5:   end for
6:    $\mathbf{T}_i^{s_L,m} = \text{Up}_{5-i}(\mathbf{P}_i^{s_L,m})$ 
7:   for  $l = L-1, \dots, 0$  do
8:      $\mathbf{T}_i^{s_l,m} = \text{Up}_{5-i}(\mathbf{P}_i^{s_l,m}) + \mathbf{T}_i^{s_{l+1},m}$ 
9:   end for
10:   $\mathbf{O}_i^m = \{\mathbf{T}_i^{c,m}, \mathbf{T}_i^{s_1,m}, \dots, \mathbf{T}_i^{s_L,m}\}$ 
11: end for
12: return  $\{\mathbf{O}_i^1, \dots, \mathbf{O}_i^M\}$ 

```

$2C/r + 1$). Then, we can calculate the parameter reduction ratio $\frac{p_s}{p}$ as follows:

$$\begin{aligned}
\frac{p_s}{p} &= \frac{C\gamma^{s-1}(18C + C\gamma^{s-1} \times \frac{2}{r} + 1)}{C(18C + \frac{2C}{r} + 1)} \\
&= \gamma^{s-1} \left(\frac{18C + C\gamma^{s-1} \times \frac{2}{r} + 1}{18C + \frac{2C}{r} + 1} \right) \\
&\approx \gamma^{s-1} \left(\frac{18C + C\gamma^{s-1} \times \frac{2}{r}}{18C + \frac{2C}{r}} \right) \\
&= \gamma^{s-1} \left(\frac{18r + 2\gamma^{s-1}}{18r + 2} \right) \\
&= (\gamma^{s-1})^2 \left(\frac{\frac{18r}{\gamma^{s-1}} + 2}{18r + 2} \right) \\
&\approx (\gamma^{s-1})^2 \left(\frac{\frac{18r}{\gamma^{s-1}}}{18r} \right) \\
&= \gamma^{s-1}
\end{aligned} \tag{12}$$

8. Ensemble Sub-Decoding Module for Multi-label Segmentation

In this section, we show that E-SDM can be utilized to any segmentation dataset with M multi-label.

Forward Stream. During the forward stream, core and sub-task pseudo predictions $\{\mathbf{P}_i^{c,m}, \mathbf{P}_i^{s_1,m}, \dots, \mathbf{P}_i^{s_L,m}\}$ for each label $m \in \{1, 2, \dots, M\}$ are produced at i -th decoder stage as follows:

$$\begin{cases} \mathbf{P}_i^{c,m} = \text{Conv2D}_1(\mathbf{Y}_i) \\ \mathbf{P}_i^{s_l,m} = \text{Conv2D}_1(\mathbf{Y}_i \times \sigma(\mathbf{P}_i^{s_{l-1},m})) \text{ for } l = 1, \dots, L \end{cases} \tag{13}$$

where $\mathbf{P}_i^{s_0,m} = \mathbf{P}_i^{c,m}$. This stream enables the following sub-task prediction cascadingly focus on the region by spatial attention starting from core pseudo prediction $\mathbf{P}_i^{c,m}$.

Backward Stream. After producing L -th sub-task pseudo prediction $\mathbf{P}_i^{s_L,m}$, to produce final core task prediction $\mathbf{T}_i^{c,m}$ for m -th label, we apply backward stream as follows:

$$\begin{cases} \mathbf{T}_i^{s_L,m} = \text{Up}_{5-i}(\mathbf{P}_i^{s_L,m}) \\ \mathbf{T}_i^{s_l,m} = \text{Up}_{5-i}(\mathbf{P}_i^{s_l,m}) + \mathbf{T}_i^{s_{l+1},m} \text{ for } l = 0, \dots, L-1 \end{cases} \tag{14}$$

where $\mathbf{T}_i^{s_0,m} = \mathbf{T}_i^{c,m}$. To further analyze the Eq 14, we can recursively rewrite from core task $\mathbf{T}_i^{c,m}$ as follows:

$$\begin{aligned}
\mathbf{T}_i^{c,m} &= \mathbf{T}_i^{s_0,m} = \text{Up}_{5-i}(\mathbf{P}_i^{s_0,m}) + \mathbf{T}_i^{s_1,m} \\
&= [\text{Up}_{5-i}(\mathbf{P}_i^{s_0,m}) + \text{Up}_{5-i}(\mathbf{P}_i^{s_1,m})] + \mathbf{T}_i^{s_2,m} = \dots \\
&= \sum_{l=0}^L \text{Up}_{5-i}(\mathbf{P}_i^{s_l,m})
\end{aligned} \tag{15}$$

Consequently, E-SDM can be interpreted as an ensemble of predictions between different tasks for describing the same legion for each m -th label. Algorithm 2 describes the detailed training algorithm for E-SDM in multi-label segmentation.

9. More Detailed Ablation Study on MADGNet

9.1. Ablation on Backbone Model

Backbone	Seen Datasets ([3, 5, 6, 23, 30, 32])					
	DSC $\uparrow$	mIoU $\uparrow$	$F_\beta^w \uparrow$	$S_\alpha \uparrow$	$E_\phi^{max} \uparrow$	MAE $\downarrow$
ResNet50 [25]	86.9	80.6	84.6	84.9	91.9	2.3
Res2Net50 [20]	87.1	80.9	83.1	85.0	92.0	2.4
ViT-B-16 [13]	87.5	81.2	85.0	85.1	92.3	2.4
ResNeSt50 [70] (Ours)	88.5	82.3	85.9	85.7	92.8	2.4
Backbone	Unseen Datasets ([1, 12, 41, 54, 57, 61, 74])					
	DSC $\uparrow$	mIoU $\uparrow$	$F_\beta^w \uparrow$	$S_\alpha \uparrow$	$E_\phi^{max} \uparrow$	MAE $\downarrow$
ResNet50 [25]	69.1	61.0	67.6	73.8	80.7	6.2
Res2Net50 [20]	70.2	61.8	68.7	74.2	81.5	5.9
ViT-B-16 [13]	69.0	61.6	67.8	75.4	81.1	4.9
ResNeSt50 [70] (Ours)	77.1	68.1	75.0	77.2	87.0	6.2

Table 8. Quantitative results for each *Seen* ([3, 5, 6, 23, 30, 32]) and *Unseen* ([1, 12, 41, 54, 57, 61, 74]) datasets according to backbone network. We presents the *mean* performance for each dataset.

In this section, we present the performance of MADGNet according to various backbone network (ResNet50 [25], Res2Net50 [20], ViT-B-16 [13], and **ResNeSt50 [70] (Ours)**) in Tab. 8. We report the *mean* performance of *seen* and *unseen* datasets.

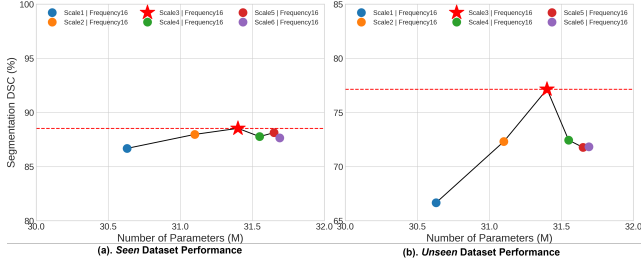


Figure 10. Comparison of parameters (M) vs segmentation performance (DSC) according to number of scale S on average for (a) *seen* and (b) *unseen* datasets.

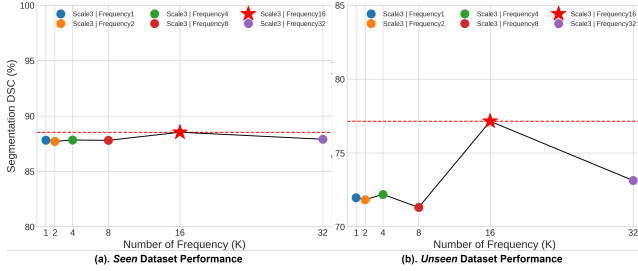


Figure 11. Number of frequency F vs segmentation performance (DSC) on average for (a) *seen* and (b) *unseen* datasets.

9.2. Hyperparameter on MADGNet

Number of Scale branch S .

Scale S	Seen Datasets ([3, 5, 6, 23, 30, 32])					
	DSC $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{max} \uparrow$	MAE $\downarrow$
1	86.7	81.0	84.5	84.8	91.6	2.4
2	88.0	81.7	85.5	85.4	92.6	2.3
3 (Ours)	88.5	82.3	85.9	85.7	92.8	2.4
4	87.8	81.5	85.3	85.3	92.6	2.3
5	88.1	81.9	85.7	85.5	92.7	2.3
6	87.6	81.9	85.3	85.3	92.7	2.4

Scale S	Unseen Datasets ([1, 12, 41, 54, 57, 61, 74])					
	DSC $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{max} \uparrow$	MAE $\downarrow$
1	66.7	59.1	66.0	73.7	79.4	9.6
2	72.3	64.1	70.8	74.9	83.8	5.1
3 (Ours)	77.1	68.1	75.0	77.2	87.0	6.2
4	72.4	63.7	70.6	74.6	82.3	8.3
5	71.8	63.6	70.4	75.8	83.5	5.4
6	71.8	63.6	70.5	76.4	84.0	4.8

Table 9. Quantitative results for each *Seen* ([3, 5, 6, 23, 30, 32]) and *Unseen* ([1, 12, 41, 54, 57, 61, 74]) datasets according to the number of scale S . We presents the *mean* performance for each domain.

In this section, we present the performance of MADGNet according to the number of scales $S \in \{1, 2, 3, 4, 5, 6\}$ with $F = 16$ in Tab. 9 and Fig. 10. We report the *mean* performance of *seen* and *unseen* datasets.

Number of Frequency branch K .

In this section, we present the performance of MADGNet according to the number of frequencies $F \in \{1, 2, 4, 8, 16, 32\}$ with $S = 3$ in Tab. 10 and Fig. 11. We

Frequency K	Seen Datasets ([3, 5, 6, 23, 30, 32])					
	DSC $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{max} \uparrow$	MAE $\downarrow$
1	87.8	81.6	85.4	85.3	92.5	2.3
2	87.7	81.5	85.3	85.2	92.6	2.3
4	87.8	81.5	85.3	85.3	92.5	2.4
8	87.8	81.5	85.3	85.3	92.4	2.3
16 (Ours)	88.5	82.3	85.9	85.7	92.8	2.4
32	87.9	81.7	85.5	85.3	92.7	2.3

Frequency K	Unseen Datasets ([1, 12, 41, 54, 57, 61, 74])					
	DSC $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{max} \uparrow$	MAE $\downarrow$
1	72.0	63.7	70.4	75.9	83.3	5.6
2	71.8	63.7	70.4	76.3	83.8	5.3
4	72.2	64.0	70.9	76.8	84.6	4.6
8	71.3	63.0	69.7	75.3	82.9	6.1
16 (Ours)	77.1	68.1	75.0	77.2	87.0	6.2
32	73.1	64.1	70.9	74.6	82.5	7.8

Table 10. Quantitative results for each *Seen* ([3, 5, 6, 23, 30, 32]) and *Unseen* ([1, 12, 41, 54, 57, 61, 74]) datasets according to the number of frequency K . We presents the *mean* performance for each domain.

report the *mean* performance of *seen* and *unseen* datasets.

Frequency Selection Strategy (Top vs Bot vs Low).

Strategy	Seen Datasets ([3, 5, 6, 23, 30, 32])					
	DSC $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{max} \uparrow$	MAE $\downarrow$
Top (Ours)	88.5	82.3	85.9	85.7	92.8	2.4
Bot	87.7	81.4	85.0	85.2	92.4	2.4
Low	87.1	80.5	84.3	86.2	92.4	2.0

Strategy	Unseen Datasets ([1, 12, 41, 54, 57, 61, 74])					
	DSC $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{max} \uparrow$	MAE $\downarrow$
Top (Ours)	77.1	68.1	75.0	77.2	87.0	6.2
Bot	72.1	63.5	70.3	74.3	82.9	7.3
Low	70.1	62.5	69.2	75.8	82.8	5.0

Table 11. Quantitative results for each *Seen* ([3, 5, 6, 23, 30, 32]) and *Unseen* ([1, 12, 41, 54, 57, 61, 74]) datasets according to frequency selection strategies. We presents the *mean* performance for each domain.

In this section, we present the performance of MADGNet according to frequency selections **Top (Ours)**, Bot, and Low with $S = 3$ and $F = 16$ in Tab. 11. The set of DCT basis images according to each frequency selection strategy can be seen in the Fig. 9. We report the *mean* performance of *seen* and *unseen* datasets.

10. Technical Innovation, Design Principle and Interpretability of MADGNet

Motivated by papers [35, 65], to extract discriminative features in both the frequency and spatial domains, we introduced dual attention modules with multiple statistic information of frequency (MFCA) and two learnable information flow parameters in multi-scale (MSSA). The causal effect of MFMSA block is interpreted as follows: 1) *MFCA emphasizes the salient features while reducing the influence of noisy features by focusing on the frequency of interest*, characterized by high variance of frequency in medical domain (Fig. 1). 2) *MSSA captures more reliable discriminative boundary cues (Fig. 6) for lesions of various*

sizes by combining foreground and background attention with multi-scale attention and information flow parameters. Our approach distinguishes itself by successfully integrating both attentions with dilated convolution and downsampling, a pioneering endeavor in the medical domain.

11. Metrics Descriptions

- *Mean Dice Similarity Coefficient (DSC)* measures the similarity between two samples and is widely used in assessing the performance of segmentation tasks, such as image segmentation or object detection. Higher is better.
- *Mean Intersection over Union (IoU)* measures the ratio of the intersection area to the union area of the predicted and ground truth masks in segmentation tasks. Higher is better.
- *Mean Weighted F-Measure F_{β}^w* is a metric that combines precision and recall into a single value by calculating the harmonic mean. "Weighted" often implies that it might be weighted by class frequency or other factors to provide a balanced measure across different classes. Higher is better.
- *Mean S-Measure S_{α}* is used to evaluate the quality of image segmentation, specifically focusing on the structural similarity between the predicted and ground truth segmentation. Higher is better.
- *Mean E-Measure E_{ϕ}^{max}* assesses the edge accuracy in edge detection or segmentation tasks. It evaluates how well the predicted edges align with the ground truth edges. Higher is better.
- *Mean Mean Absolute Error (MAE)* calculates the average absolute differences between predicted and ground truth values. Lower is better.

12. More Qualitative and Quantitative Results

In this section, we provide the quantitative results with various metrics in Tab. 12, 13, 14, 15, and 16. We report the *mean* performance of three trials for all results. (·) denotes a standard deviation of three trials. **Red** and **Blue** are the first and second best performance results, respectively. We also present more various qualitative results on datasets in Fig. 12, 13, 14, 15, and 16.

Method	ISIC2018 [23] $\Rightarrow$ ISIC2018 [23]					
	DSC $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{max} \uparrow$	MAE $\downarrow$
UNet [51]	87.3 (0.8)	80.2 (0.7)	87.9 (0.0)	80.4 (0.1)	91.3 (0.0)	4.7 (0.0)
AttUNet [44]	87.8 (0.1)	80.5 (0.1)	86.5 (0.2)	80.5 (0.1)	92.0 (0.1)	4.5 (0.0)
UNet++ [73]	87.3 (0.2)	80.2 (0.1)	86.3 (0.2)	80.1 (0.1)	91.6 (0.2)	4.7 (0.0)
CENet [22]	89.1 (0.2)	82.1 (0.1)	88.1 (0.2)	81.3 (0.1)	93.0 (0.2)	4.3 (0.1)
TransUNet [7]	87.3 (0.2)	81.2 (0.2)	88.6 (0.2)	80.8 (0.2)	91.9 (0.2)	4.2 (0.1)
FRCUNet [4]	88.9 (0.1)	83.1 (0.2)	89.3 (0.0)	82.0 (0.1)	93.9 (0.2)	3.7 (0.1)
MSRFNet [56]	88.2 (0.2)	81.3 (0.2)	86.9 (0.2)	80.7 (0.1)	92.0 (0.2)	4.7 (0.1)
HiFormer [26]	88.7 (0.5)	81.9 (0.5)	87.6 (0.6)	80.8 (0.5)	92.6 (0.5)	4.4 (0.3)
DCSAUNet [66]	89.0 (0.3)	82.0 (0.3)	87.8 (0.3)	81.4 (0.1)	92.9 (0.3)	4.4 (0.1)
M2SNet [72]	89.2 (0.2)	83.4 (0.2)	88.9 (0.1)	81.8 (0.1)	93.8 (0.1)	3.7 (0.0)
MADGNet	90.2 (0.1)	83.7 (0.2)	89.2 (0.2)	82.0 (0.1)	94.1 (0.3)	3.6 (0.2)
Method	ISIC2018 [23] $\Rightarrow$ PH2 [41]					
	DSC $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{max} \uparrow$	MAE $\downarrow$
UNet [51]	90.3 (0.1)	83.5 (0.1)	88.4 (0.1)	74.8 (0.1)	90.8 (0.1)	6.9 (0.0)
AttUNet [44]	89.9 (0.2)	82.6 (0.3)	87.3 (0.3)	74.8 (0.2)	90.8 (0.2)	6.7 (0.2)
UNet++ [73]	88.0 (0.3)	80.1 (0.3)	85.7 (0.2)	73.2 (0.1)	89.2 (0.2)	7.9 (0.1)
CENet [22]	90.5 (0.1)	83.3 (0.1)	87.3 (0.1)	75.1 (0.0)	91.5 (0.1)	6.0 (0.1)
TransUNet [7]	89.5 (0.3)	82.1 (0.4)	86.9 (0.4)	74.3 (0.2)	90.3 (0.2)	6.7 (0.2)
FRCUNet [4]	90.6 (0.1)	83.4 (0.2)	87.4 (0.2)	75.4 (0.2)	91.7 (0.1)	5.9 (0.1)
MSRFNet [56]	90.5 (0.3)	83.5 (0.3)	87.5 (0.3)	75.0 (0.0)	91.4 (0.2)	6.0 (0.3)
HiFormer [26]	86.9 (1.6)	79.1 (1.8)	83.2 (1.9)	72.9 (1.1)	88.6 (1.4)	8.0 (0.9)
DCSAUNet [66]	89.0 (0.4)	81.5 (0.3)	85.7 (0.2)	74.0 (0.3)	90.2 (0.3)	6.9 (0.4)
M2SNet [72]	90.7 (0.3)	83.5 (0.5)	87.6 (0.4)	75.5 (0.3)	92.0 (0.2)	5.9 (0.2)
MADGNet	91.3 (0.1)	84.6 (0.1)	88.4 (0.1)	76.2 (0.1)	92.8 (0.1)	5.1 (0.1)

Table 12. Segmentation results on **Skin Lesion Segmentation (Dermatoscopy)** [23, 41]. We train each model on ISIC2018 [23] train dataset and evaluate on ISIC2018 [23] and PH2 [41] test datasets.

Method	COVID19-1 [32] $\Rightarrow$ COVID19-1 [32]					
	DSC $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{max} \uparrow$	MAE $\downarrow$
UNet [51]	47.7 (0.6)	38.6 (0.6)	36.1 (0.2)	69.6 (0.1)	62.7 (0.7)	2.1 (0.0)
AttUNet [44]	57.5 (0.2)	48.4 (0.2)	45.3 (1.8)	74.5 (1.2)	66.0 (2.3)	1.7 (0.0)
UNet++ [73]	65.6 (0.7)	57.1 (0.8)	54.4 (8.9)	78.8 (3.3)	73.2 (5.2)	1.3 (0.3)
CENet [22]	76.3 (0.4)	69.2 (0.5)	64.4 (0.2)	83.2 (0.2)	76.6 (0.3)	0.6 (0.0)
TransUNet [7]	75.6 (0.4)	68.8 (0.2)	63.4 (0.2)	82.7 (0.3)	75.5 (0.1)	0.7 (0.0)
FRCUNet [4]	77.3 (0.3)	70.4 (0.2)	66.0 (0.4)	84.0 (0.2)	78.4 (0.6)	0.7 (0.0)
MSRFNet [56]	75.2 (0.4)	68.0 (0.4)	63.4 (0.4)	82.7 (0.2)	76.3 (0.6)	0.8 (0.0)
HiFormer [26]	72.9 (1.4)	63.3 (1.5)	60.2 (1.0)	80.8 (0.8)	76.0 (1.0)	0.8 (0.1)
DCSAUNet [66]	75.3 (0.4)	68.2 (0.4)	63.1 (0.6)	83.0 (0.3)	77.3 (0.5)	0.7 (0.0)
M2SNet [72]	81.7 (0.4)	74.7 (0.5)	68.3 (0.7)	85.7 (0.2)	80.1 (0.4)	0.6 (0.0)
MADGNet	83.7 (0.2)	76.8 (0.2)	70.2 (0.2)	86.3 (0.2)	81.5 (0.1)	0.5 (0.0)
Method	COVID19-1 [32] $\Rightarrow$ COVID19-2 [1]					
	DSC $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{max} \uparrow$	MAE $\downarrow$
UNet [51]	47.1 (0.7)	37.7 (0.6)	46.7 (0.8)	68.7 (0.2)	68.6 (1.0)	1.0 (0.0)
AttUNet [44]	43.7 (0.8)	35.2 (0.8)	44.5 (0.7)	67.9 (0.5)	64.0 (0.6)	1.0 (0.0)
UNet++ [73]	50.5 (3.8)	40.9 (3.7)	50.6 (4.6)	69.8 (1.3)	75.7 (2.6)	1.0 (0.2)
CENet [22]	60.1 (0.3)	49.9 (0.3)	61.1 (0.3)	73.4 (0.1)	80.1 (0.3)	1.1 (0.0)
TransUNet [7]	56.9 (1.0)	48.0 (0.7)	58.0 (0.1)	72.5 (0.2)	80.0 (1.3)	0.8 (0.0)
FRCUNet [4]	62.9 (1.1)	52.7 (0.9)	63.7 (1.2)	74.5 (0.3)	82.1 (0.7)	1.4 (0.2)
MSRFNet [56]	58.3 (0.8)	48.4 (0.6)	59.1 (0.9)	72.7 (0.2)	79.8 (0.8)	1.0 (0.1)
HiFormer [26]	54.1 (1.0)	44.5 (0.8)	55.2 (1.0)	70.9 (0.5)	78.0 (0.7)	1.0 (0.1)
DCSAUNet [66]	52.4 (1.2)	44.0 (0.7)	52.0 (1.2)	71.3 (0.1)	76.3 (3.1)	1.0 (0.1)
M2SNet [72]	68.6 (0.1)	58.9 (0.2)	68.5 (0.2)	76.9 (0.1)	86.1 (0.4)	1.1 (0.1)
MADGNet	72.2 (0.3)	62.6 (0.3)	72.3 (0.5)	78.2 (0.1)	88.1 (0.2)	1.0 (0.1)

Table 13. Segmentation results on **COVID19 Infection Segmentation (Radiology)** [1, 32]. We train each model on COVID19-1 [32] train dataset and evaluate on COVID19-1 [32] and COVID19-2 [1] test datasets.

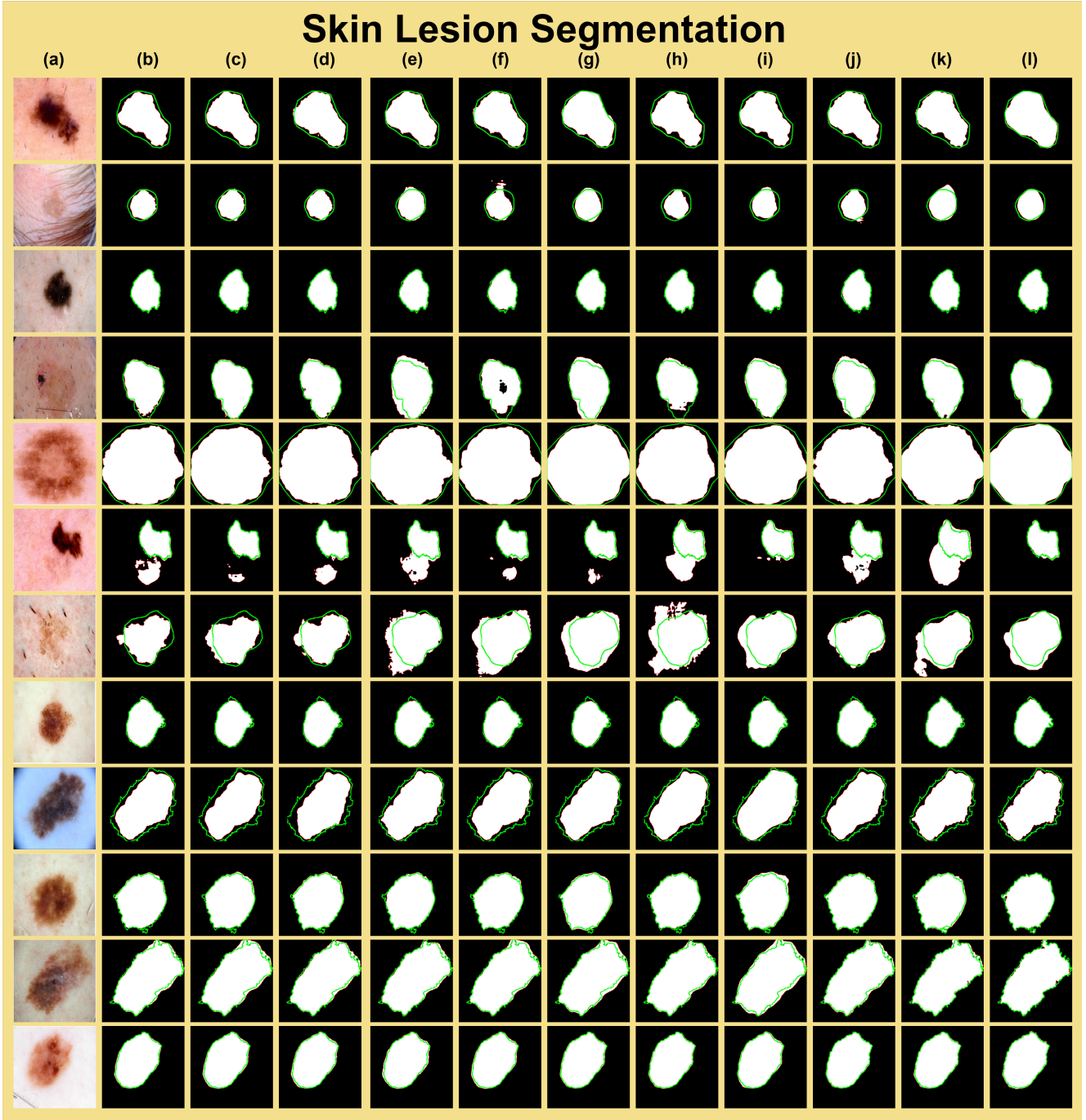


Figure 12. Qualitative comparison of other methods and MADGNet on **Skin Lesion Segmentation (Dermoscopy)** [23, 41]. (a) Input images, (b) UNet [51], (c) AttUNet [44], (d) UNet++ [73], (e) CENet [22], (f) TransUNet [7], (g) FRCUNet [4], (h) MSRFNet [56], (i) HiFormer [26], (j) DCSAUNet [66], (k) M2SNet [72], (l) **MADGNet (Ours)**. **Green** and **Red** lines denote the boundaries of the ground truth and prediction, respectively.

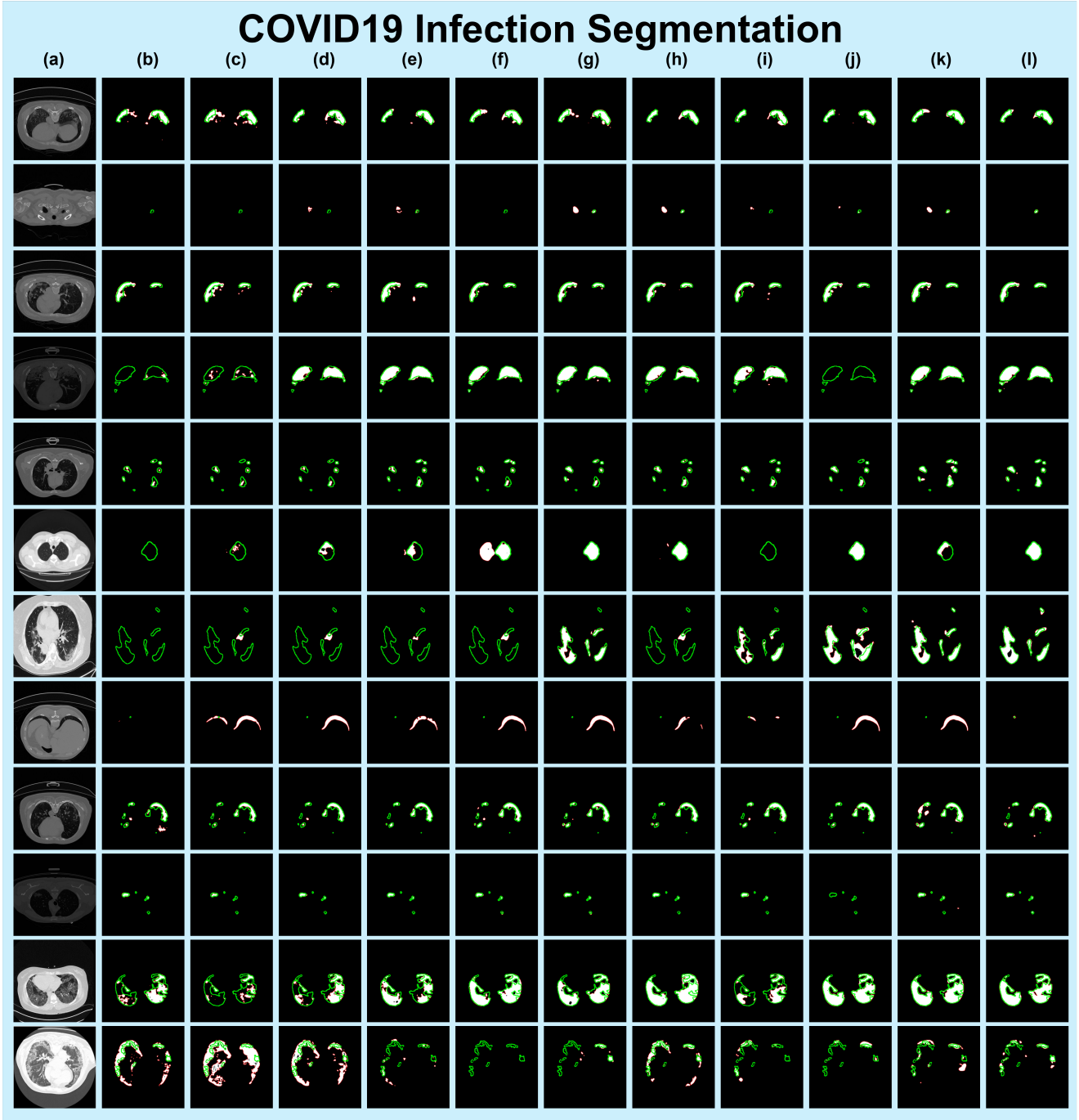


Figure 13. Qualitative comparison of other methods and MADGNet on **COVID19 Infection Segmentation (Radiology)** [1, 32]. (a) Input images, (b) UNet [51], (c) AttUNet [44], (d) UNet++ [73], (e) CENet [22], (f) TransUNet [7], (g) FRCUNet [4], (h) MSRFNet [56], (i) HiFormer [26], (j) DCSAUNet [66], (k) M2SNet [72], (l) **MADGNet (Ours)**. **Green** and **Red** lines denote the boundaries of the ground truth and prediction, respectively.

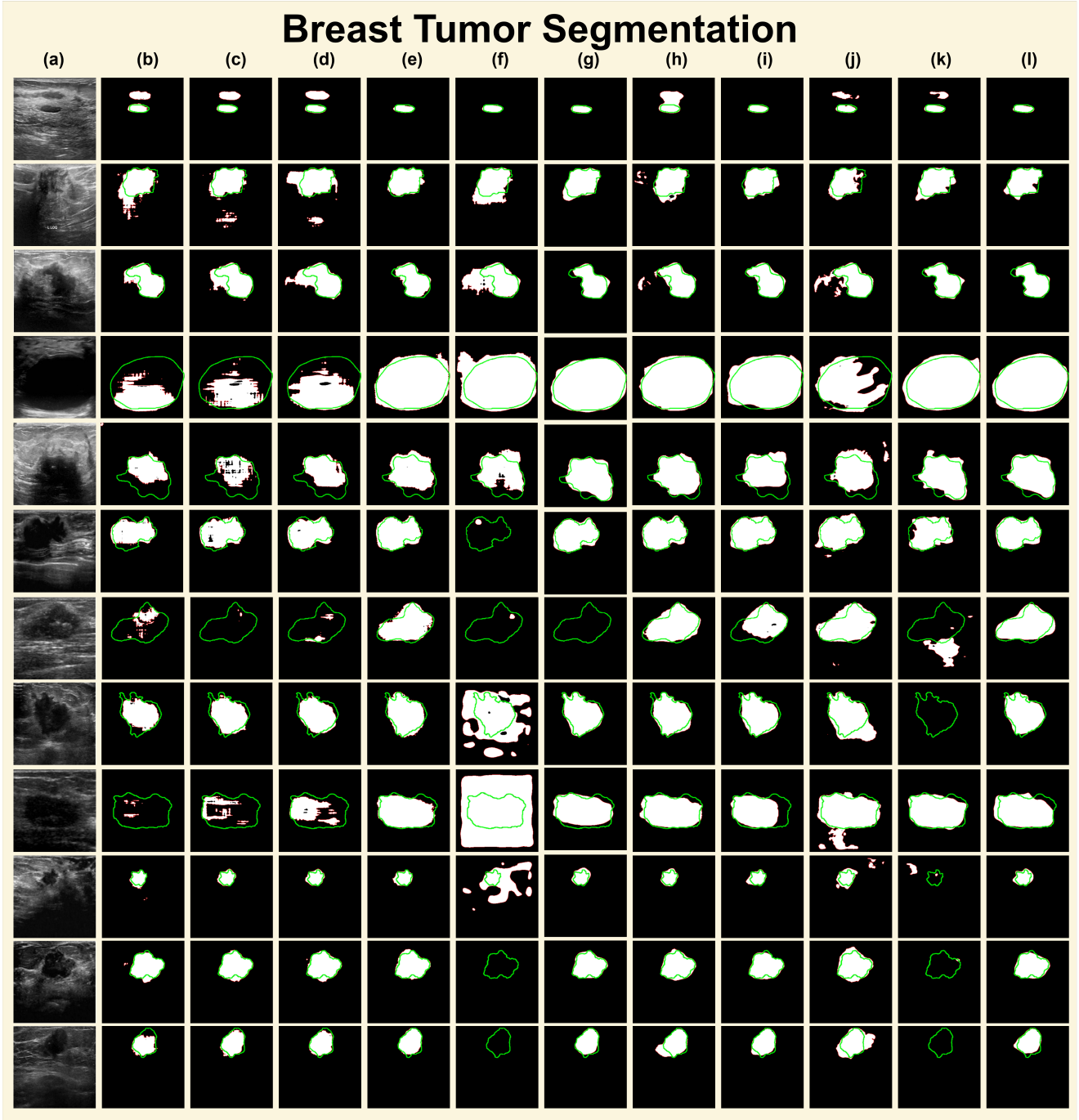


Figure 14. Qualitative comparison of other methods and MADGNet on **Breast Tumor Segmentation (Ultrasound)** [3, 74]. (a) Input images, (b) UNet [51], (c) AttUNet [44], (d) UNet++ [73], (e) CENet [22], (f) TransUNet [7], (g) FRCUNet [4], (h) MSRFNet [56], (i) HiFormer [26], (j) DCSAUNet [66], (k) M2SNet [72], (l) **MADGNet (Ours)**. **Green** and **Red** lines denote the boundaries of the ground truth and prediction, respectively.

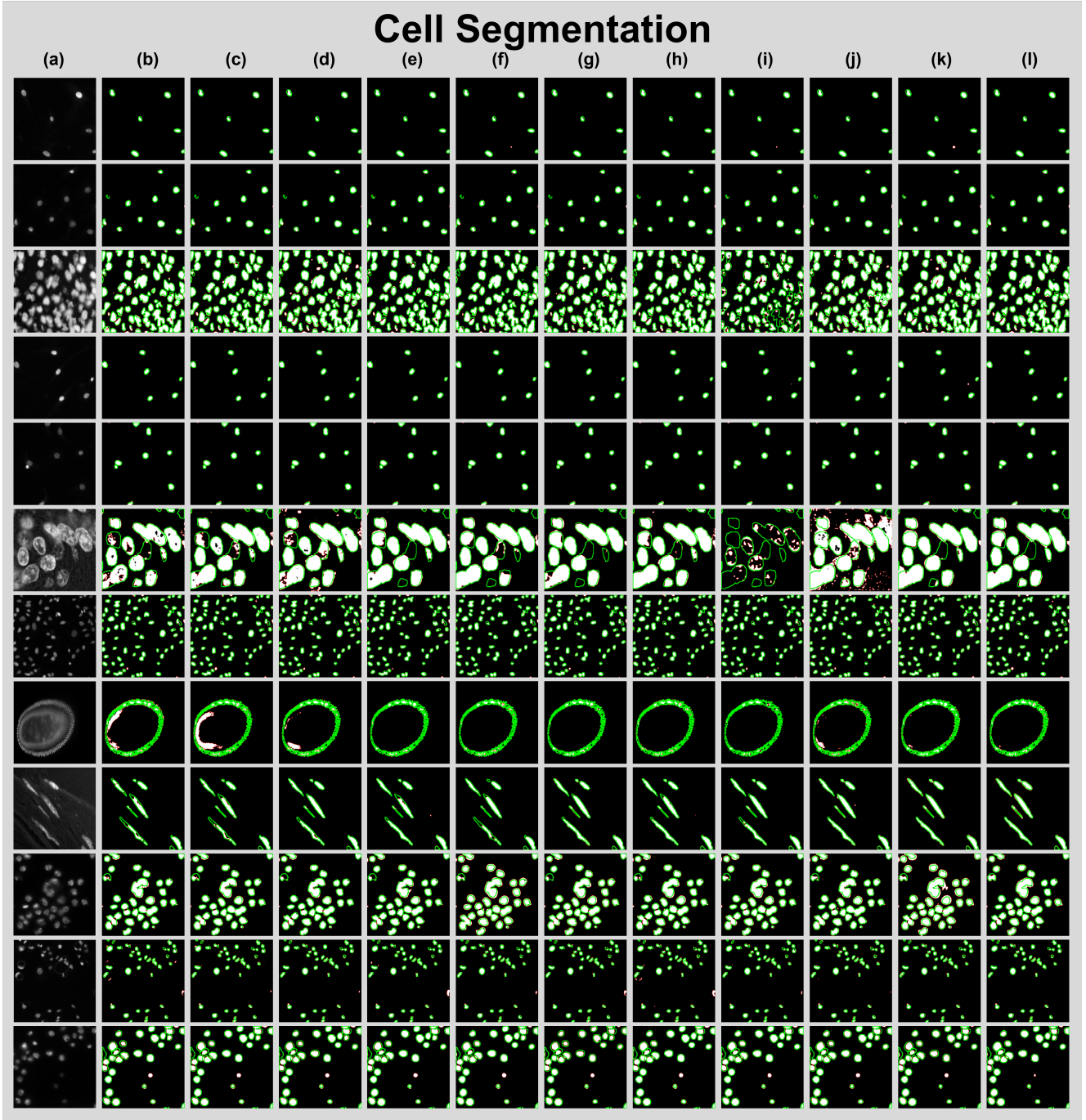


Figure 15. Qualitative comparison of other methods and MADGNet on **Cell Segmentation (Microscopy)** [6, 12]. (a) Input images, (b) UNet [51], (c) AttUNet [44], (d) UNet++ [73], (e) CENet [22], (f) TransUNet [7], (g) FRCUNet [4], (h) MSRFNet [56], (i) HiFormer [26], (j) DCSAUNet [66], (k) M2SNet [72], (l) **MADGNet (Ours)**. **Green** and **Red** lines denote the boundaries of the ground truth and prediction, respectively.

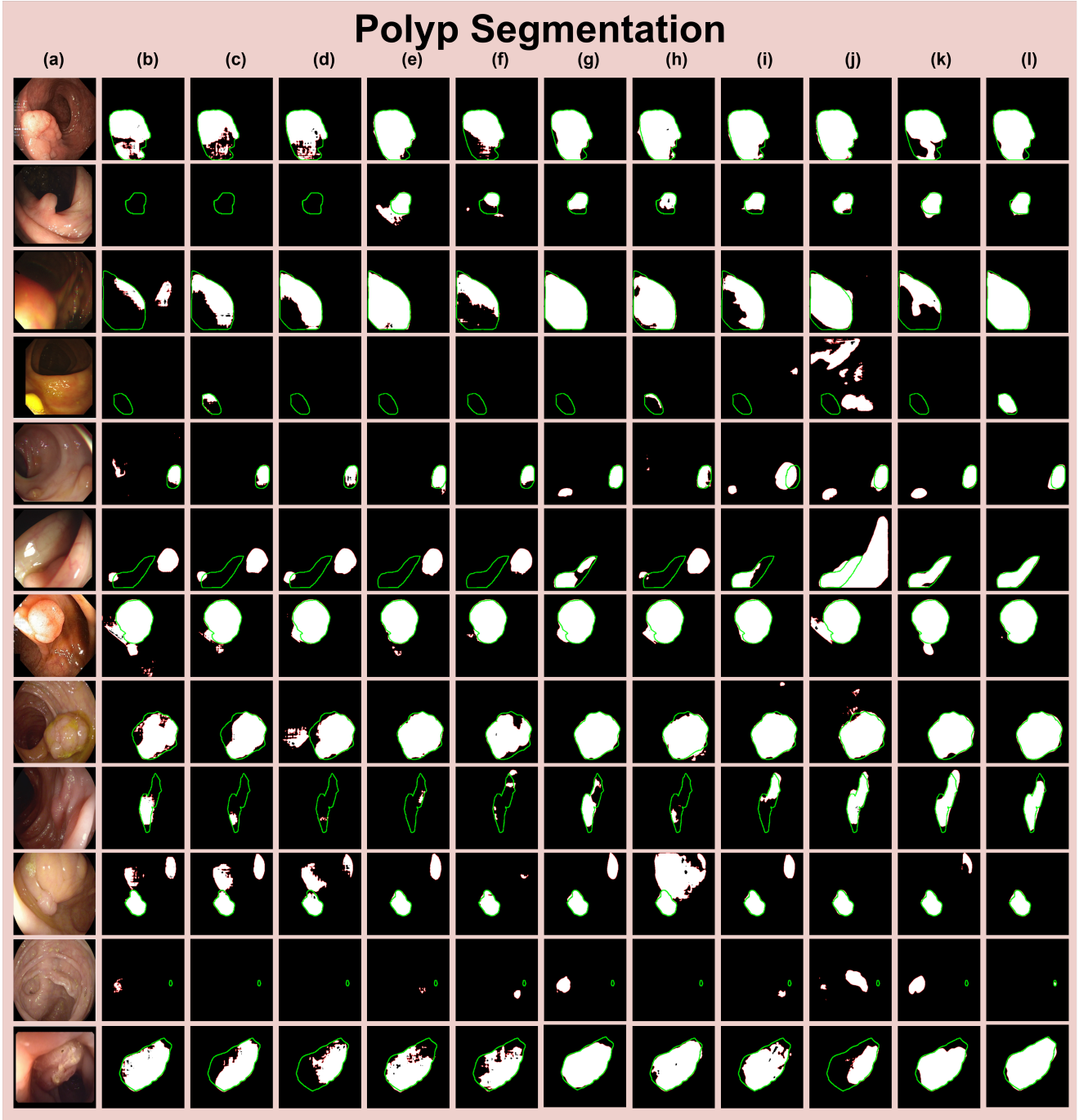


Figure 16. Qualitative comparison of other methods and MADGNet on **Polyp Segmentation (Colonoscopy)** [5, 30, 54, 57, 61]. (a) Input images, (b) UNet [51], (c) AttUNet [44], (d) UNet++ [73], (e) CENet [22], (f) TransUNet [7], (g) FRCUNet [4], (h) MSRFNet [56], (i) HiFormer [26], (j) DCSAUNet [66], (k) M2SNet [72], (l) **MADGNet (Ours)**. **Green** and **Red** lines denote the boundaries of the ground truth and prediction, respectively.

Method	BUSI [3] $\Rightarrow$ BUSI [3]					
	DSC $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{max} \uparrow$	MAE $\downarrow$
UNet [51]	69.5 (0.3)	60.2 (0.2)	67.2 (0.3)	76.9 (0.1)	83.2 (0.2)	4.8 (0.0)
AttUNet [44]	71.3 (0.4)	62.3 (0.6)	68.9 (0.5)	78.1 (0.3)	84.4 (0.1)	4.8 (0.0)
UNet++ [73]	72.4 (0.1)	62.5 (0.2)	68.7 (0.3)	78.4 (0.2)	85.0 (0.2)	5.0 (0.1)
CENet [22]	79.7 (0.6)	71.5 (0.5)	78.1 (0.6)	82.8 (0.3)	91.1 (0.2)	3.9 (0.0)
TransUNet [7]	75.5 (0.5)	68.4 (0.1)	73.8 (0.2)	79.8 (0.1)	88.6 (0.6)	4.2 (0.2)
FRCUNet [4]	81.2 (0.2)	73.3 (0.3)	79.9 (0.3)	83.5 (0.2)	91.9 (0.1)	3.7 (0.1)
MSRFNet [56]	76.6 (0.7)	68.1 (0.7)	75.1 (0.9)	80.9 (0.3)	88.5 (0.4)	4.2 (0.1)
HiFormer [26]	79.3 (0.2)	70.8 (0.1)	77.7 (0.0)	82.3 (0.1)	90.8 (0.3)	4.1 (0.1)
DCSAUNet [66]	73.7 (0.5)	65.0 (0.5)	71.5 (0.4)	79.6 (0.3)	86.0 (0.3)	4.6 (0.1)
M2SNet [72]	80.4 (0.8)	72.5 (0.7)	78.7 (0.6)	83.0 (0.5)	91.2 (0.4)	4.1 (0.2)
MADGNet	81.3 (0.4)	73.4 (0.4)	79.5 (0.4)	83.8 (0.2)	91.7 (0.3)	3.6 (0.1)

Method	BUSI [3] $\Rightarrow$ STU [74]					
	DSC $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{max} \uparrow$	MAE $\downarrow$
UNet [51]	71.6 (1.0)	61.6 (0.7)	71.6 (0.8)	76.1 (0.4)	82.4 (0.9)	5.2 (0.2)
AttUNet [44]	77.0 (1.6)	68.0 (1.7)	76.4 (1.2)	79.8 (1.0)	86.7 (1.4)	4.4 (0.3)
UNet++ [73]	77.3 (0.4)	67.8 (0.3)	76.1 (0.3)	79.4 (0.5)	87.6 (0.3)	4.4 (0.1)
CENet [22]	86.0 (0.7)	77.2 (0.9)	84.2 (0.6)	84.6 (0.4)	93.7 (0.4)	2.8 (0.2)
TransUNet [7]	41.4 (9.5)	32.1 (4.2)	40.8 (8.7)	60.2 (4.3)	58.1 (8.4)	9.7 (0.7)
FRCUNet [4]	86.5 (2.3)	77.2 (2.7)	84.9 (2.1)	85.2 (1.6)	94.1 (2.0)	2.8 (0.5)
MSRFNet [56]	84.0 (5.5)	75.2 (8.2)	82.5 (5.0)	83.5 (5.7)	92.2 (3.3)	3.1 (0.2)
HiFormer [26]	80.7 (2.9)	71.3 (3.2)	78.9 (3.0)	81.2 (1.6)	90.1 (2.2)	3.7 (0.5)
DCSAUNet [66]	86.1 (0.5)	76.5 (0.8)	82.7 (0.8)	84.9 (0.4)	94.7 (0.5)	3.2 (0.1)
M2SNet [72]	79.4 (0.7)	69.3 (0.6)	76.4 (0.8)	81.3 (0.4)	90.7 (0.9)	4.3 (0.2)
MADGNet	88.4 (1.0)	79.9 (1.5)	86.4 (1.5)	86.2 (0.9)	95.9 (0.5)	2.6 (0.4)

Table 14. Segmentation results on **Breast Tumor Segmentation (Ultrasound)** [3, 74]. We train each model on BUSI [3] train dataset and evaluate on BUSI [3] and STU [74] test datasets.

Method	DSB2018 [6] $\Rightarrow$ DSB2018 [6]					
	DSC $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{max} \uparrow$	MAE $\downarrow$
UNet [51]	91.1 (0.2)	84.3 (0.3)	92.1 (0.1)	83.3 (0.0)	96.8 (0.0)	2.5 (0.0)
AttUNet [44]	91.6 (0.1)	85.0 (0.1)	92.5 (0.0)	83.7 (0.0)	97.2 (0.0)	2.4 (0.0)
UNet++ [73]	91.6 (0.1)	85.0 (0.1)	92.8 (0.1)	83.6 (0.0)	97.1 (0.0)	2.4 (0.0)
CENet [22]	91.3 (0.1)	84.6 (0.1)	92.5 (0.1)	83.6 (0.1)	97.2 (0.1)	2.3 (0.0)
TransUNet [7]	91.8 (0.3)	85.2 (0.2)	92.8 (0.1)	83.8 (0.2)	97.3 (0.2)	2.3 (0.1)
FRCUNet [4]	90.8 (0.3)	83.8 (0.4)	92.1 (0.2)	83.2 (0.2)	97.0 (0.2)	2.5 (0.1)
MSRFNet [56]	91.9 (0.1)	85.3 (0.1)	92.7 (0.1)	83.7 (0.1)	97.5 (0.0)	2.3 (0.0)
HiFormer [26]	90.7 (0.2)	83.8 (0.4)	91.1 (1.1)	82.5 (0.8)	96.0 (1.1)	2.7 (0.4)
DCSAUNet [66]	91.1 (0.2)	84.4 (0.2)	91.8 (0.2)	82.9 (0.3)	96.6 (0.2)	2.8 (0.2)
M2SNet [72]	91.6 (0.2)	85.1 (0.3)	92.0 (0.2)	83.5 (0.1)	97.5 (0.1)	2.2 (0.1)
MADGNet	92.0 (0.0)	85.5 (0.1)	92.3 (0.4)	83.8 (0.2)	97.4 (0.3)	2.3 (0.1)

Method	DSB2018 [6] $\Rightarrow$ MonuSeg2018 [12]					
	DSC $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{max} \uparrow$	MAE $\downarrow$
UNet [51]	29.2 (5.1)	18.9 (3.5)	28.0 (5.2)	38.3 (1.8)	49.9 (6.0)	32.5 (2.6)
AttUNet [44]	39.0 (3.1)	26.5 (2.4)	40.9 (2.2)	39.3 (2.9)	61.4 (5.0)	24.5 (4.6)
UNet++ [73]	25.4 (0.8)	15.3 (0.5)	21.1 (0.4)	5.5 (1.7)	21.1 (3.1)	66.6 (4.5)
CENet [22]	27.7 (1.5)	16.9 (1.0)	26.7 (1.1)	37.7 (1.9)	59.6 (4.9)	27.5 (5.1)
TransUNet [7]	15.9 (8.5)	9.6 (5.5)	17.0 (6.2)	32.6 (4.1)	39.0 (5.5)	23.4 (7.6)
FRCUNet [4]	26.1 (5.6)	16.8 (4.3)	26.9 (7.7)	40.6 (6.7)	49.3 (9.7)	22.9 (10.5)
MSRFNet [56]	9.1 (1.0)	5.3 (0.7)	12.0 (0.9)	35.6 (1.6)	46.4 (0.4)	18.7 (0.4)
HiFormer [26]	21.9 (8.9)	13.2 (5.7)	22.5 (6.7)	39.9 (1.4)	48.2 (0.3)	23.9 (9.1)
DCSAUNet [66]	4.3 (0.3)	2.4 (0.9)	5.2 (3.2)	18.0 (7.3)	37.0 (1.2)	25.9 (5.9)
M2SNet [72]	36.3 (0.9)	23.1 (0.8)	29.1 (1.6)	20.4 (1.5)	34.5 (6.8)	45.1 (5.3)
MADGNet	46.7 (4.3)	32.0 (2.9)	43.0 (4.6)	40.8 (2.6)	60.7 (6.3)	29.2 (5.3)

Table 15. Segmentation results on **Cell Segmentation (Microscopy)** [6, 12]. We train each model on DSB2018 [6] train dataset and evaluate on DSB2018 [6] and MonuSeg2018 [12] test datasets.

Method	CVC-ClinicDB [5] + Kvasir-SEG [30] $\rightarrow$ CVC-ClinicDB [5]					
	DSC $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{max} \uparrow$	MAE $\downarrow$
UNet [51]	76.5 (0.8)	69.1 (0.9)	75.1 (0.8)	83.0 (0.4)	86.4 (0.6)	2.7 (0.0)
AttUNet [44]	80.1 (0.6)	74.2 (0.5)	79.8 (0.7)	85.1 (0.4)	88.5 (0.5)	2.1 (0.1)
UNet++ [73]	79.7 (0.2)	73.6 (0.4)	79.4 (0.1)	85.1 (0.2)	88.3 (0.5)	2.2 (0.0)
CENet [22]	89.3 (0.3)	84.0 (0.2)	89.1 (0.2)	89.8 (0.2)	96.0 (0.6)	1.1 (0.0)
TransUNet [7]	87.4 (0.2)	82.9 (0.1)	87.2 (0.1)	88.5 (0.2)	95.2 (0.1)	1.3 (0.0)
FRCUNet [4]	91.8 (0.2)	87.0 (0.2)	91.3 (0.3)	91.1 (0.1)	97.1 (0.3)	0.7 (0.0)
MSRFNet [56]	83.2 (0.9)	76.5 (1.1)	81.9 (1.2)	86.4 (0.5)	91.3 (1.0)	1.7 (0.0)
HiFormer [26]	89.1 (0.6)	83.7 (0.6)	88.8 (0.5)	89.5 (0.2)	96.1 (0.8)	1.1 (0.2)
DCSAUNet [66]	80.5 (1.2)	73.7 (1.1)	79.6 (1.1)	84.9 (0.6)	89.9 (1.0)	2.4 (0.2)
M2SNet [72]	92.8 (0.8)	88.2 (0.8)	92.3 (0.7)	91.4 (0.4)	97.7 (0.5)	0.7 (0.1)
MADGNet	93.9 (0.6)	89.5 (0.5)	93.6 (0.6)	92.2 (0.2)	98.5 (0.7)	0.7 (0.0)

Method	CVC-ClinicDB [5] + Kvasir-SEG [30] $\rightarrow$ Kvasir-SEG [30]					
	DSC $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{max} \uparrow$	MAE $\downarrow$
UNet [51]	80.5 (0.3)	72.6 (0.4)	78.2 (0.4)	79.9 (0.2)	88.2 (0.2)	5.2 (0.2)
AttUNet [44]	83.9 (0.1)	77.1 (0.1)	83.1 (0.0)	81.9 (0.0)	90.0 (0.1)	4.4 (0.1)
UNet++ [73]	84.3 (0.3)	77.4 (0.2)	83.1 (0.3)	82.1 (0.1)	90.5 (0.2)	4.6 (0.1)
CENet [22]	89.5 (0.7)	83.9 (0.7)	88.9 (0.7)	85.3 (0.3)	94.1 (0.4)	3.0 (0.2)
TransUNet [7]	86.4 (0.4)	81.3 (0.4)	85.4 (0.4)	83.0 (0.4)	92.1 (0.5)	4.0 (0.3)
FRCUNet [4]	88.8 (0.4)	83.5 (0.6)	88.4 (0.6)	85.1 (0.2)	93.6 (0.4)	3.3 (0.1)
MSRFNet [56]	86.1 (0.5)	79.3 (0.4)	84.9 (0.7)	82.8 (0.1)	92.0 (0.4)	4.0 (0.1)
HiFormer [26]	88.1 (1.0)	82.3 (1.2)	87.3 (1.1)	84.6 (0.5)	93.9 (0.6)	3.1 (0.3)
DCSAUNet [66]	82.6 (0.5)	75.2 (0.5)	80.7 (0.3)	81.3 (0.7)	90.1 (0.1)	4.9 (0.2)
M2SNet [72]	90.2 (0.5)	85.1 (0.6)	89.4 (0.8)	85.6 (0.5)	94.6 (0.7)	2.8 (0.1)
MADGNet	90.7 (0.8)	85.3 (0.8)	89.9 (0.8)	85.6 (0.5)	94.7 (1.0)	3.1 (0.2)

Method	CVC-ClinicDB [5] + Kvasir-SEG [30] $\rightarrow$ CVC-300 [61]					
	DSC $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{max} \uparrow$	MAE $\downarrow$
UNet [51]	66.1 (2.3)	58.5 (2.1)	65.0 (2.2)	79.7 (1.0)	80.0 (2.4)	1.7 (0.1)
AttUNet [44]	63.0 (0.3)	57.2 (0.4)	62.4 (0.4)	79.1 (0.3)	76.6 (0.9)	1.8 (0.0)
UNet++ [73]	64.3 (2.2)	58.4 (2.0)	63.7 (2.3)	79.5 (1.1)	77.4 (1.5)	1.8 (0.1)
CENet [22]	85.4 (1.6)	78.2 (1.4)	84.2 (1.8)	90.2 (0.5)	94.0 (1.5)	0.8 (0.1)
TransUNet [7]	85.0 (0.6)	77.3 (0.3)	83.1 (0.7)	89.4 (0.3)	95.2 (0.7)	1.1 (0.1)
FRCUNet [4]	86.7 (0.7)	79.4 (0.3)	85.1 (0.2)	90.5 (0.3)	95.0 (0.2)	0.8 (0.1)
MSRFNet [56]	72.3 (2.2)	65.4 (2.2)	71.2 (2.0)	83.5 (1.6)	84.6 (1.7)	1.4 (0.1)
HiFormer [26]	84.7 (1.1)	77.5 (1.1)	83.2 (0.7)	89.7 (0.6)	94.0 (1.3)	0.8 (0.3)
DCSAUNet [66]	68.9 (4.0)	59.8 (3.9)	66.3 (3.8)	81.1 (2.1)	83.8 (2.9)	2.0 (0.3)
M2SNet [72]	89.9 (0.2)	83.2 (0.3)	88.3 (0.2)	93.0 (0.2)	96.9 (0.2)	0.6 (0.0)
MADGNet	87.4 (0.4)	79.9 (0.4)	84.5 (0.5)	92.0 (0.2)	94.7 (0.5)	0.9 (0.1)

Method	CVC-ClinicDB [5] + Kvasir-SEG [30] $\rightarrow$ CVC-ColonDB [57]					
	DSC $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{max} \uparrow$	MAE $\downarrow$
UNet [51]	56.8 (1.3)	49.0 (1.2)	55.9 (1.2)	72.6 (0.6)	73.9 (1.6)	5.1 (0.1)
AttUNet [44]	56.8 (1.6)	50.0 (1.5)	56.2 (1.7)	73.0 (0.7)	72.3 (1.3)	4.9 (0.1)
UNet++ [73]	57.5 (0.4)	50.2 (0.4)	56.6 (0.3)	73.3 (0.3)	73.9 (0.5)	5.0 (0.1)
CENet [22]	65.9 (1.6)	59.2 (0.1)	65.8 (0.1)	77.7 (0.1)	79.5 (0.4)	4.0 (0.2)
TransUNet [7]	63.7 (0.1)	58.4 (0.4)	62.8 (0.9)	75.8 (0.4)	79.3 (1.6)	4.8 (0.0)
FRCUNet [4]	69.1 (1.0)	62.6 (0.9)	68.5 (1.0)	79.3 (0.6)	81.4 (0.6)	4.0 (0.1)
MSRFNet [56]	61.5 (1.0)	54.8 (0.8)	60.8 (0.8)	75.4 (0.5)	76.1 (0.9)	4.5 (0.1)
HiFormer [26]	67.6 (1.4)	60.5 (1.3)	66.9 (1.4)	78.6 (0.7)	81.2 (1.4)	4.2 (0.0)
DCSAUNet [66]	57.8 (0.4)	49.3 (0.4)	54.9 (0.6)	73.3 (0.3)	76.0 (1.3)	5.8 (0.3)
M2SNet [72]	75.8 (0.7)	68.5 (0.5)	73.7 (0.7)	84.2 (0.3)	86.9 (0.1)	3.8 (0.1)
MADGNet	77.5 (1.1)	69.7 (1.2)	76.2 (1.2)	83.3 (0.8)	88.0 (1.0)	3.2 (0.2)

Method	CVC-ClinicDB [5] + Kvasir-SEG [30] $\rightarrow$ ETIS [54]					
	DSC $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{max} \uparrow$	MAE $\downarrow$
UNet [51]	41.6 (1.1)	35.4 (1.0)	39.5 (1.0)	67.2 (0.6)	61.7 (0.2)	2.7 (0.2)
AttUNet [44]	38.4 (0.3)	33.5 (0.1)	37.6 (0.4)	65.4 (0.2)	59.7 (1.2)	2.6 (0.1)
UNet++ [73]	39.1 (2.4)	34.0 (2.1)	38.3 (2.4)	65.8 (1.0)	59.3 (1.9)	2.7 (0.1)
CENet [22]	57.0 (3.4)	51.4 (0.5)	56.0 (0.4)	74.9 (0.1)	73.7 (0.4)	2.2 (0.2)
TransUNet [7]	50.1 (0.5)	44.0 (2.3)	48.8 (1.8)	70.7 (1.3)	68.7 (2.0)	2.6 (0.1)
FRCUNet [4]	65.1 (1.0)	58.4 (0.5)	62.9 (0.6)	78.7 (0.4)	81.0 (1.3)	2.2 (0.3)
MSRFNet [56]	38.3 (0.6)	33.7 (0.7)	36.9 (0.6)	66.0 (0.2)	58.4 (0.9)	3.6 (0.5)
HiFormer [26]	56.7 (3.2)	50.1 (3.3)	55.2 (3.0)	74.1 (1.7)	74.7 (2.5)	1.8 (0.1)
DCSAUNet [66]	42.9 (3.0)	36.1 (2.9)	40.5 (3.5)	67.9 (1.4)	69.3 (2.0)	4.1 (0.9)
M2SNet [72]	74.9 (1.3)	67.8 (1.4)	71.2 (1.6)	84.6 (0.1)	87.2 (0.7)	1.7 (0.3)
MADGNet	77.0 (0.3)	69.7 (0.5)	75.3 (0.2)	84.6 (0.5)	88.4 (0.6)	1.6 (0.4)

Table 16. Segmentation results on **Polyp Segmentation (Colonoscopy)** [5, 30, 54, 57, 61]. We train each model on CVC-ClinicDB [5] + Kvasir-SEG [30] train dataset and evaluate on CVC-ClinicDB [5], Kvasir-SEG [30], CVC-300 [61], CVC-ColonDB [57], and ETIS [54] test datasets.