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

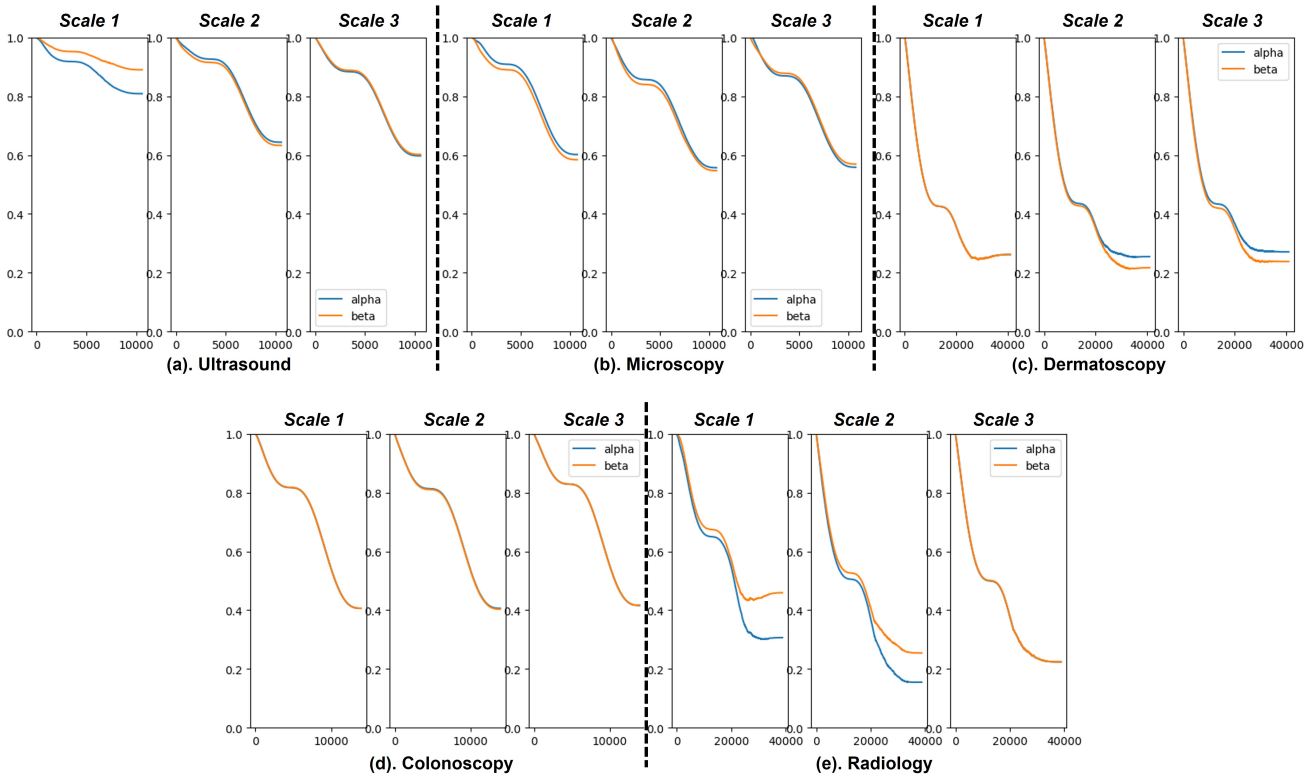

Figure 8. The training results of α_3^s and β_3^s for each modalities ((a) Ultrasound, (b) Microscopy, (c) Dermoscopy, (d) Colonoscopy, and (e) Radiology) where $s \in \{1, 2, 3\}$.

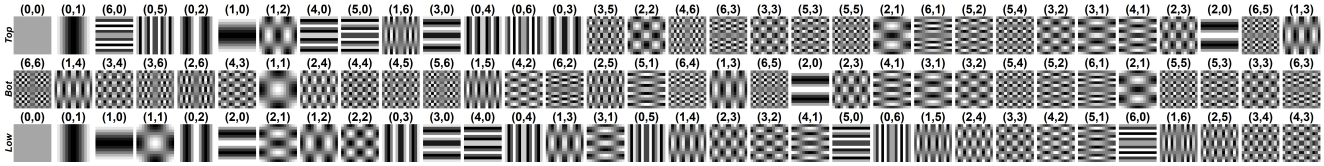

Figure 9. Frequency selection strategies (Top, Bot, Low) [49]. (u_k, v_k) denotes the frequency indices according to frequency selection strategy.

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:

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

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 $\{\mathbf{T}_i^{s_1,m}, \ldots, \mathbf{T}_i^{s_L,m}\}$ at i-th decoder for each m-th label.

for $m = 1, 2, \ldots, M$ **do**
  $\mathbf{P}_i^{c,m} = \textbf{Conv2D}_1(\mathbf{Y}_i)$
  for $l = 1, 2, \ldots, L$ **do**
    $\mathbf{P}_i^{s_l,m} = \textbf{Conv2D}_1(\mathbf{Y}_i \times \sigma\left(\mathbf{P}_i^{s_{l-1},m}\right))$.
  end for
  $\mathbf{T}_i^{s_L,m} = \textbf{Up}_{5-i}\left(\mathbf{P}_i^{s_L,m}\right)$
  for $l = L-1, \ldots, 0$ **do**
    $\mathbf{T}_i^{s_l,m} = \textbf{Up}_{5-i}\left(\mathbf{P}_i^{s_l,m}\right) + \mathbf{T}_i^{s_{l+1},m}$
  end for
  $\mathbf{O}_i^m = \{\mathbf{T}_i^{c,m}, \mathbf{T}_i^{s_1,m}, \ldots, \mathbf{T}_i^{s_L,m}\}$
end for
return $\{\mathbf{O}_i^1, \ldots, \mathbf{O}_i^M\}$

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

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

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 $\{\mathbf{P}_i^{c,m}, \mathbf{P}_i^{s_1,m}, \ldots, \mathbf{P}_i^{s_L,m}\}$ for each label $m \in \{1, 2, \ldots, M\}$ are produced at i-th decoder stage as follows:

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

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:

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

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:

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

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 ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
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 ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
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.

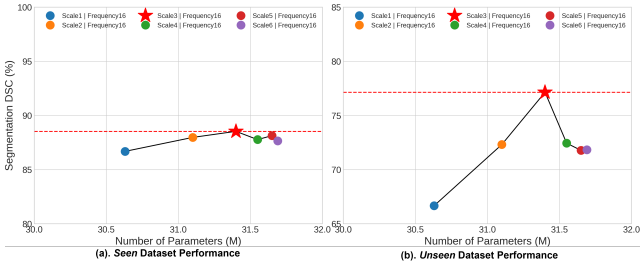

Figure 10. Comparison of parameters (M) vs segmentation performance (DSC) according to number of scale S on average for (a) *seen* and (b) *unseen* datasets.

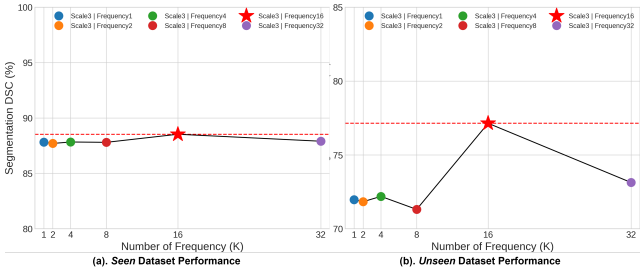

Figure 11. Number of frequency F vs segmentation performance (DSC) on average for (a) *seen* and (b) *unseen* datasets.

9.2. Hyperparameter on MADGNet

Number of Scale branch S.

Scale S	*Seen* Datasets ([3, 5, 6, 23, 30, 32])					
	DSC ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
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 ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
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 report the *mean* performance of *seen* and *unseen* datasets.

Frequency K	*Seen* Datasets ([3, 5, 6, 23, 30, 32])					
	DSC ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
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 ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
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.

Frequency Selection Strategy (Top vs Bot vs Low).

Strategy	*Seen* Datasets ([3, 5, 6, 23, 30, 32])					
	DSC ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
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 ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
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 Qualtative 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. $(\cdot)$ 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] ⇒ ISIC2018 [23]					
	DSC ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
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] ⇒ PH2 [41]					
	DSC ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
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] ⇒ COVID19-1 [32]					
	DSC ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
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] ⇒ COVID19-2 [1]					
	DSC ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
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.

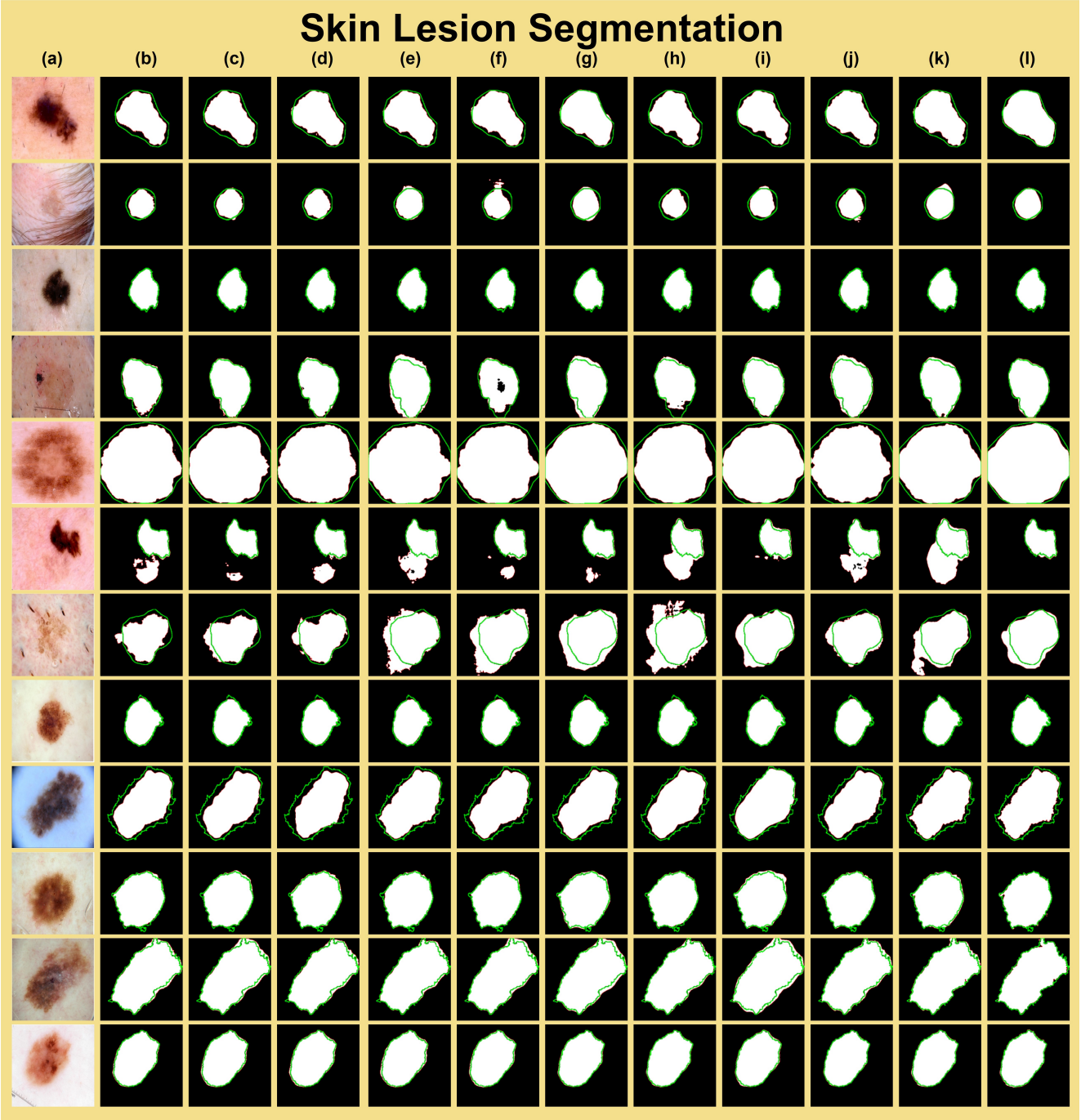

Figure 12. Qualitative comparison of other methods and MADGNet on **Skin Lesion Segmentation (Dermoscopy)** [23, 41]. (a) Input images, (b) UNet [51]. (c) AttUNet [44], (d) UNet++ [73], (e) CENet [22], (f) TransUNet [7], (g) FRCUNet [4], (h) MSRFNet [56], (i) HiFormer [26], (j) DCSAUNet [66], (k) M2SNet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

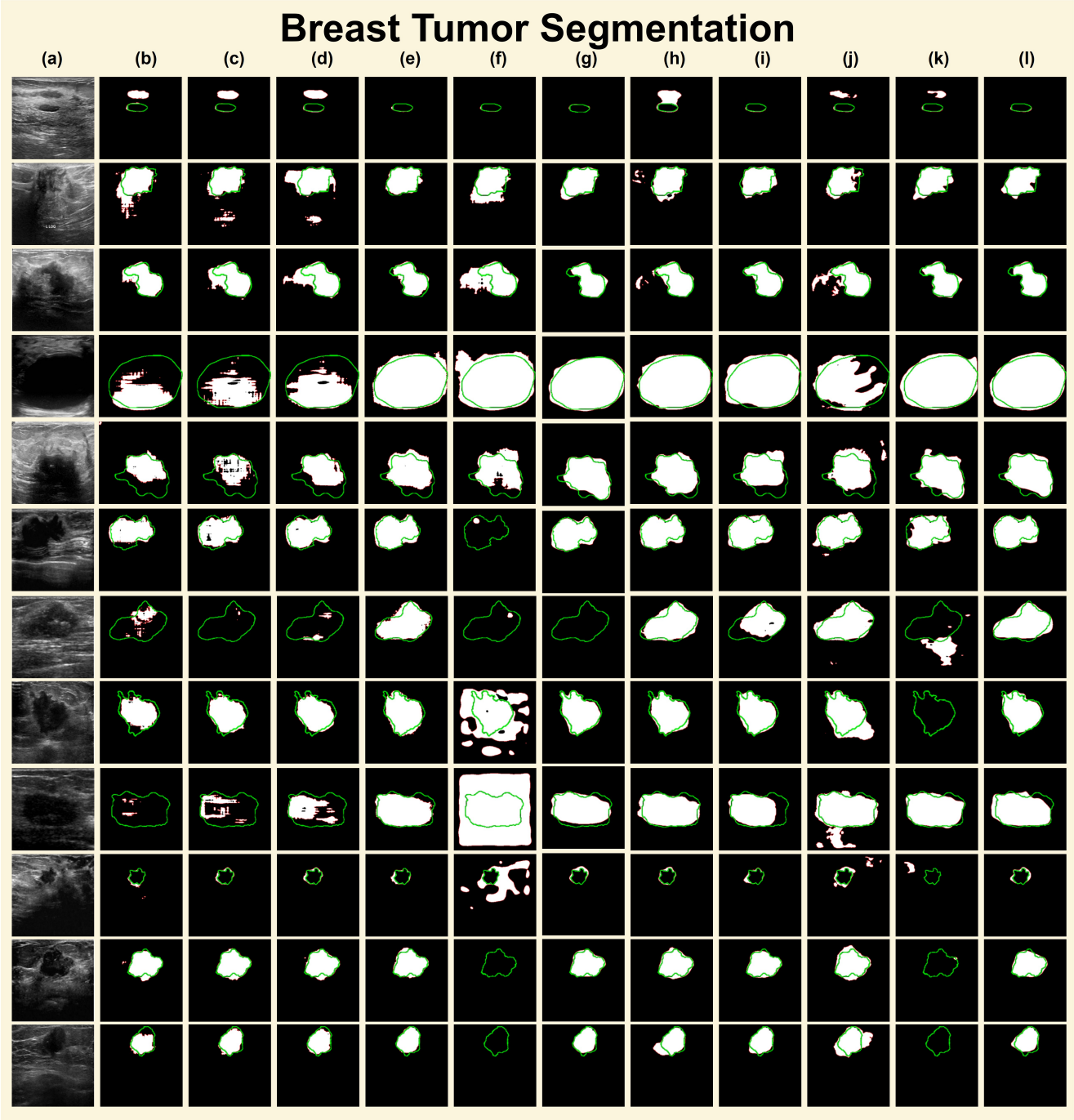

Figure 14. Qualitative comparison of other methods and MADGNet on **Breast Tumor Segmentation (Ultrasound)** [3, 74]. (a) Input images, (b) UNet [51]. (c) AttUNet [44], (d) UNet++ [73], (e) CENet [22], (f) TransUNet [7], (g) FRCUNet [4], (h) MSRFNet [56], (i) HiFormer [26], (j) DCSAUNet [66], (k) M2SNet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

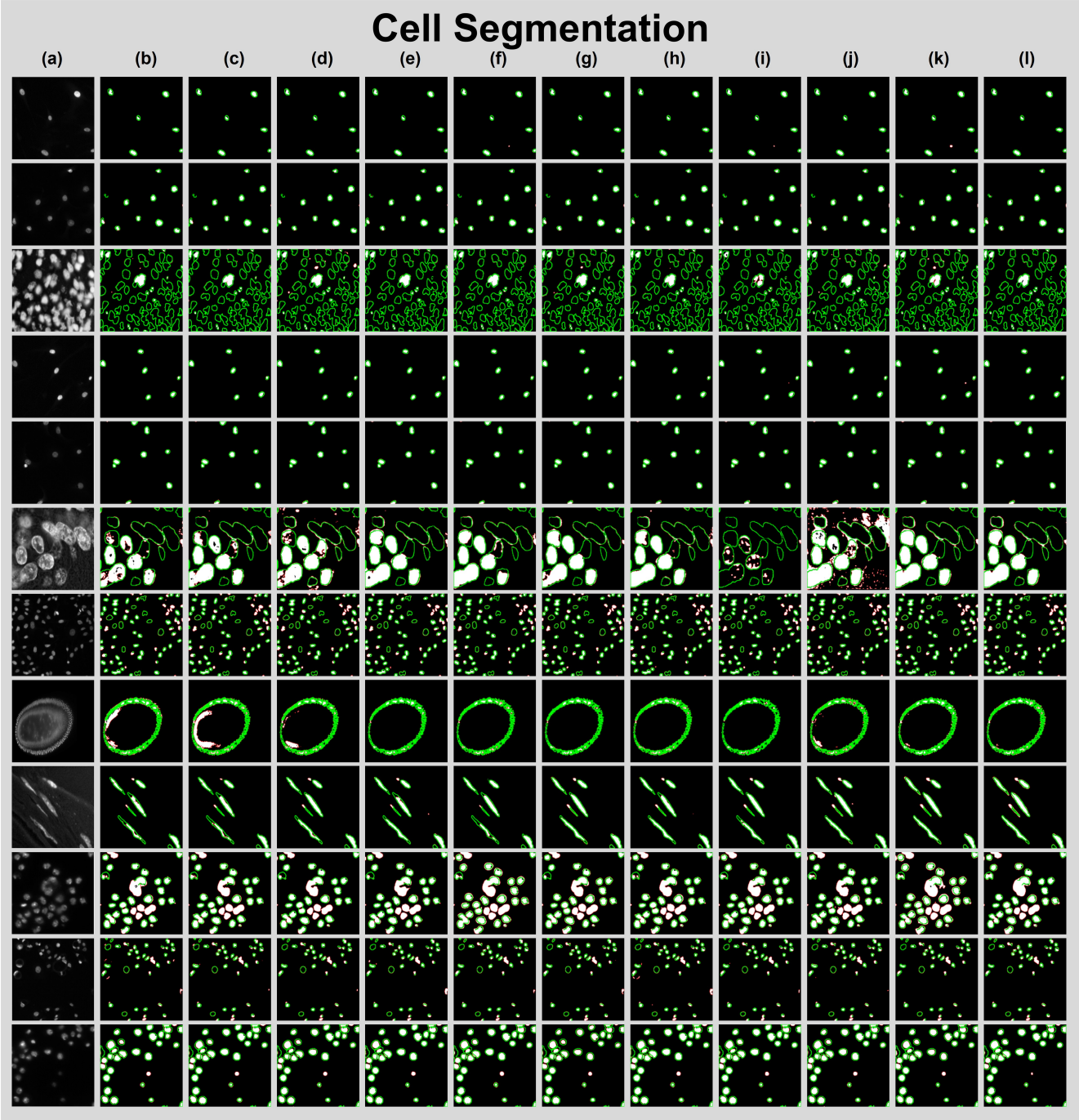

Figure 15. Qualitative comparison of other methods and MADGNet on **Cell Segmentation (Microscopy)** [6, 12]. (a) Input images, (b) UNet [51]. (c) AttUNet [44], (d) UNet++ [73], (e) CENet [22], (f) TransUNet [7], (g) FRCUNet [4], (h) MSRFNet [56], (i) HiFormer [26], (j) DCSAUNet [66], (k) M2SNet [72],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

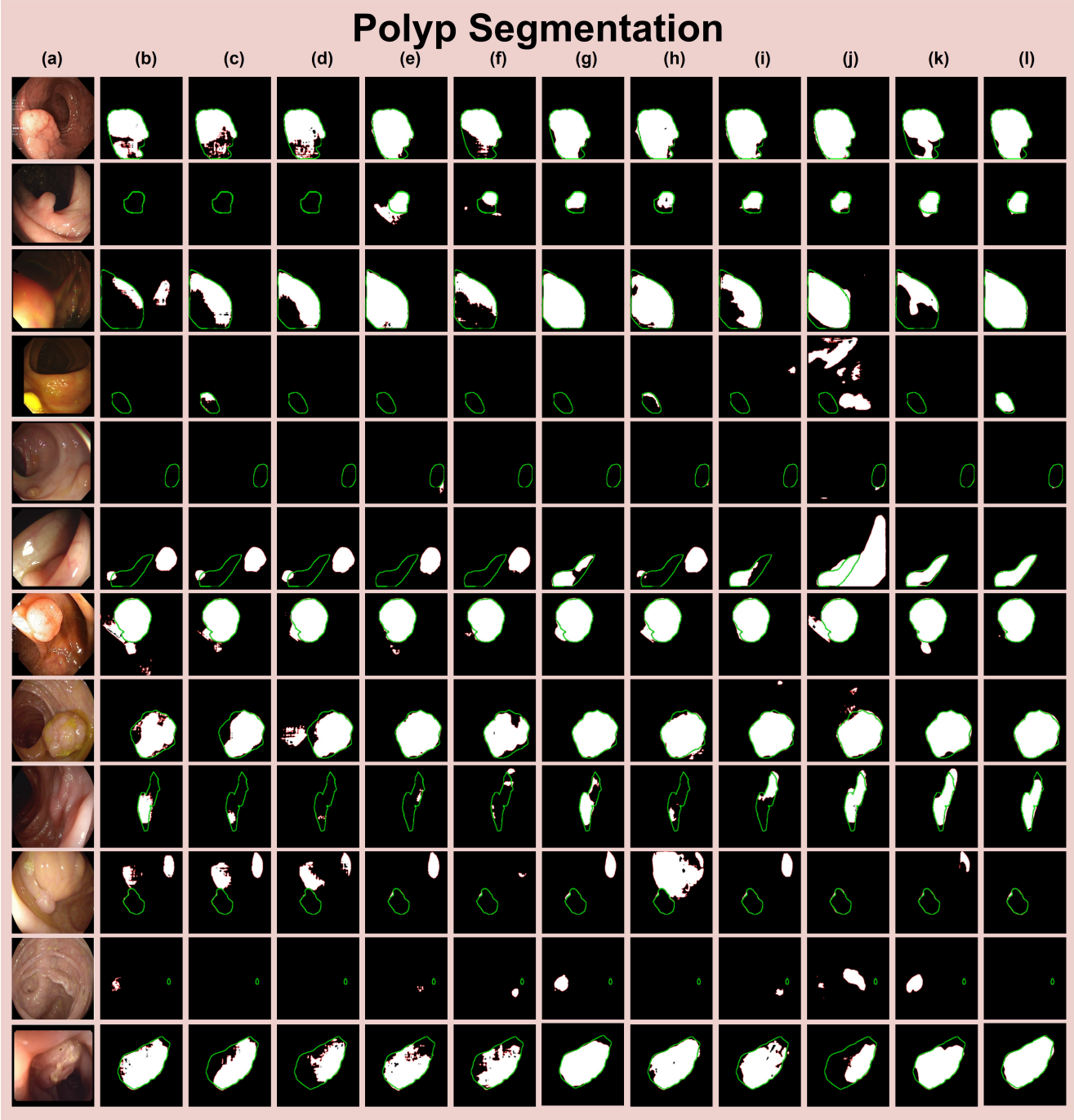

Figure 16. Qualitative comparison of other methods and MADGNet on **Polyp Segmentation (Colonoscopy)** [5, 30, 54, 57, 61]. (a) Input images, (b) UNet [51]. (c) AttUNet [44], (d) UNet++ [73], (e) CENet [22], (f) TransUNet [7], (g) FRCUNet [4], (h) MSRFNet [56], (i) HiFormer [26], (j) DCSAUNet [66], (k)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

Method	BUSI [3] ⇒ BUSI [3]					
	DSC ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
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] ⇒ STU [74]					
	DSC ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
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] ⇒ DSB2018 [6]					
	DSC ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
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] ⇒ MonuSeg2018 [12]					
	DSC ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
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] → CVC-ClinicDB [5]					
	DSC ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
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] → Kvasir-SEG [30]					
	DSC ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
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] → CVC-300 [61]					
	DSC ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
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] → CVC-ColonDB [57]					
	DSC ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
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] → ETIS [54]					
	DSC ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
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.