COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

INHA UNIVERSITY

CVPR
JUNE 17-21, 2024
SEATTLE, WA

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


Ju-Hyeon Nam[1]


Nur Suriza Syazwany[1]

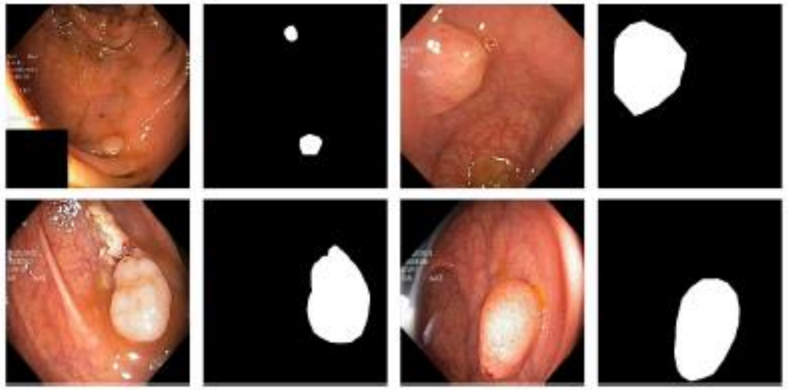

Su Jung Kim[1]


Sang-Chul Lee[1,2]

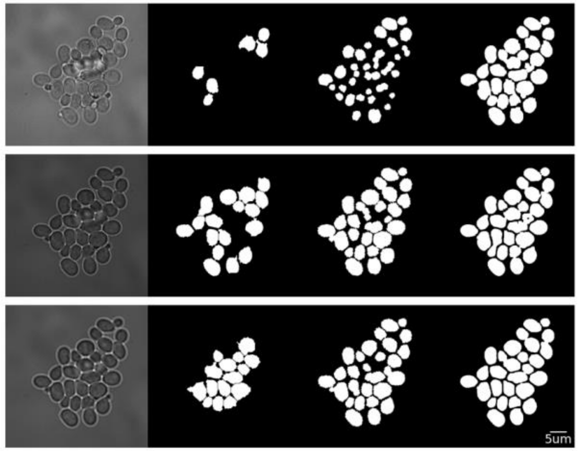
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING OF INHA UNIVERSITY, REPUBLIC OF KOREA[1]

DEEPCARDIO[2]

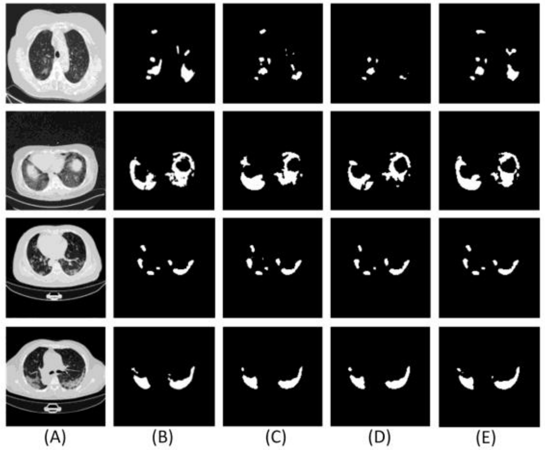
Medical Image Segmentation



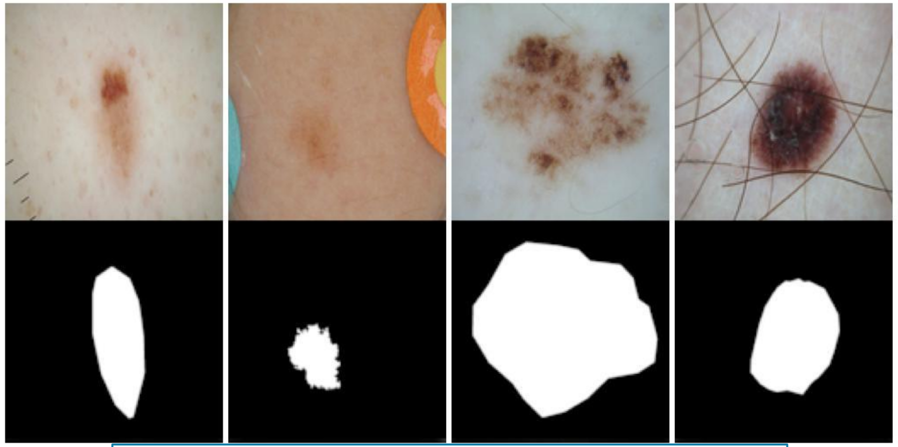
Polyp Segmentation



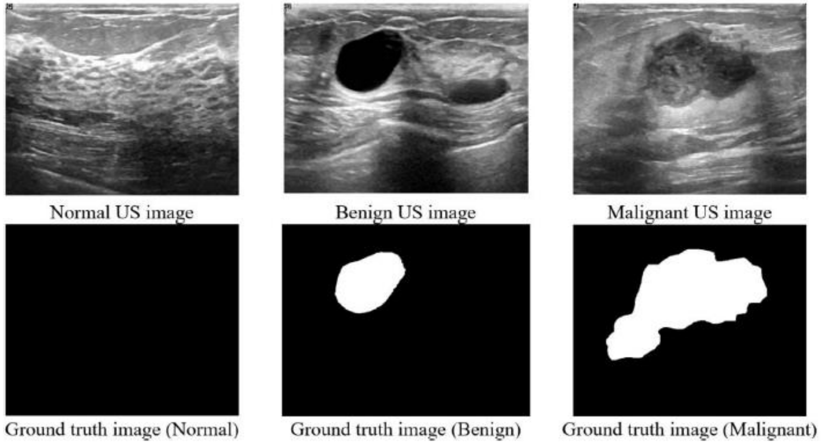
Cell Segmentation



lung Infection Segmentation



Skin Cancer Segmentation



Breast Tumor Segmentation

Related Works: Multi-Scale based Method

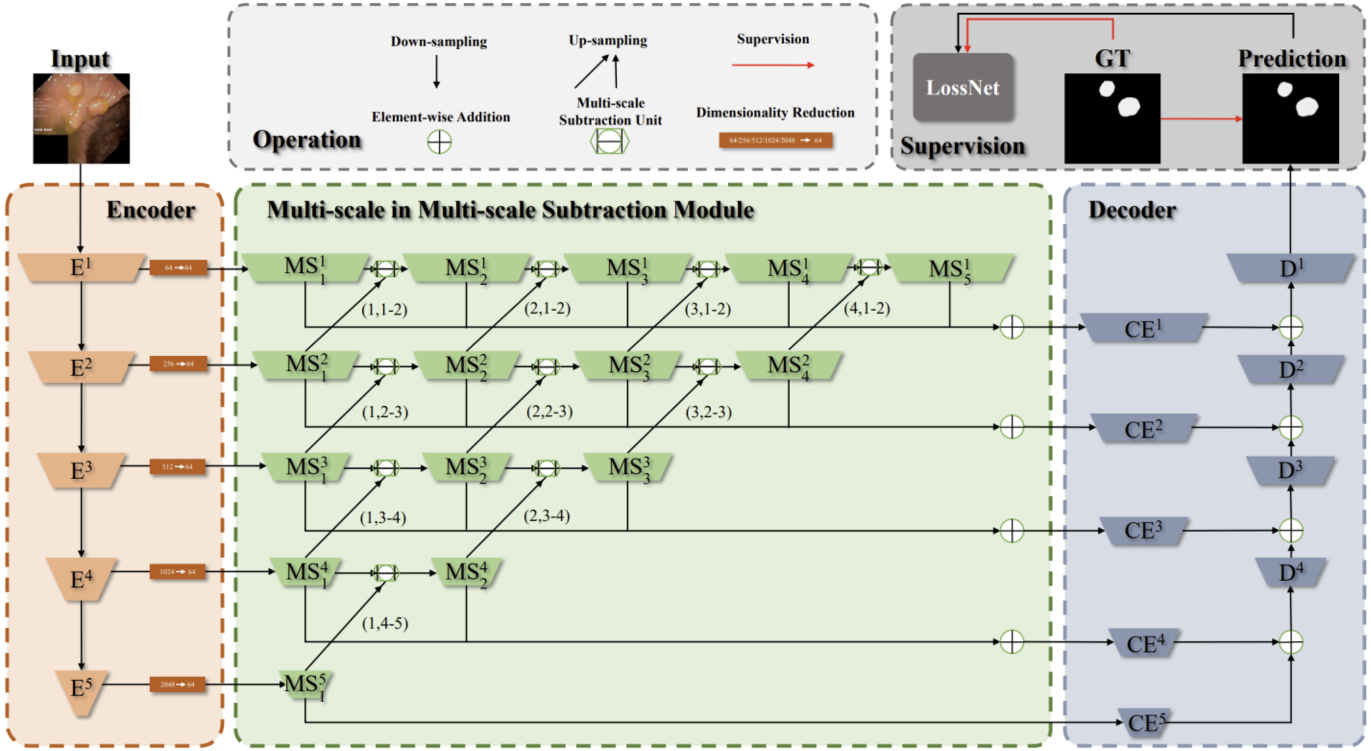


Fig. 2: Overview of the proposed multi-scale subtraction network.

- ✓ Multi-Scale Subtraction Module
- ✓ Context Enhancement Module
- ✓ Feature Map Loss
- ✓ Vulnerable to severe noise image

Zhao, Xiaoqi, et al. "M $^{2} $ SNet: Multi-scale in Multi-scale Subtraction Network for Medical Image Segmentation." *arXiv preprint arXiv:2303.10894* (2023).

Related Works: Multi-Frequency based Method

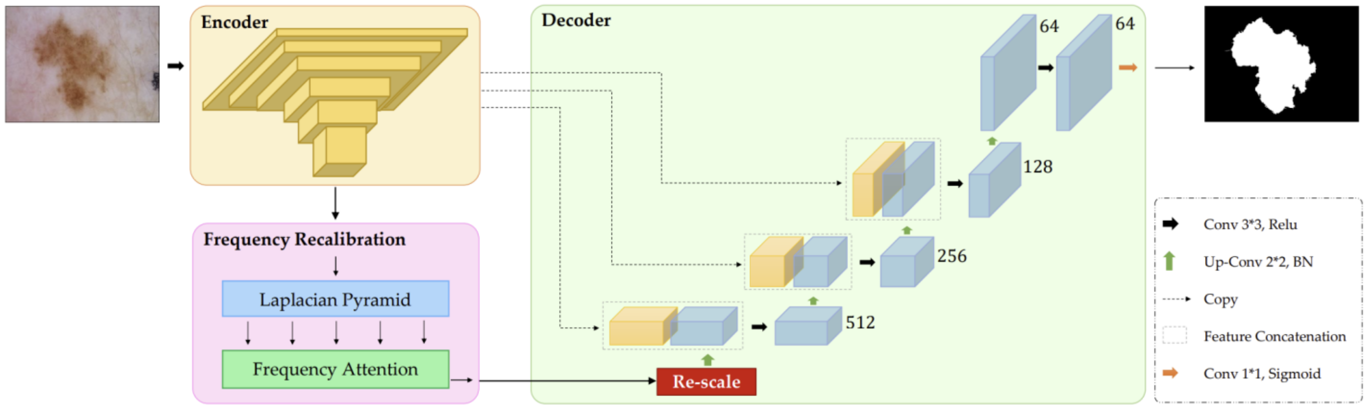
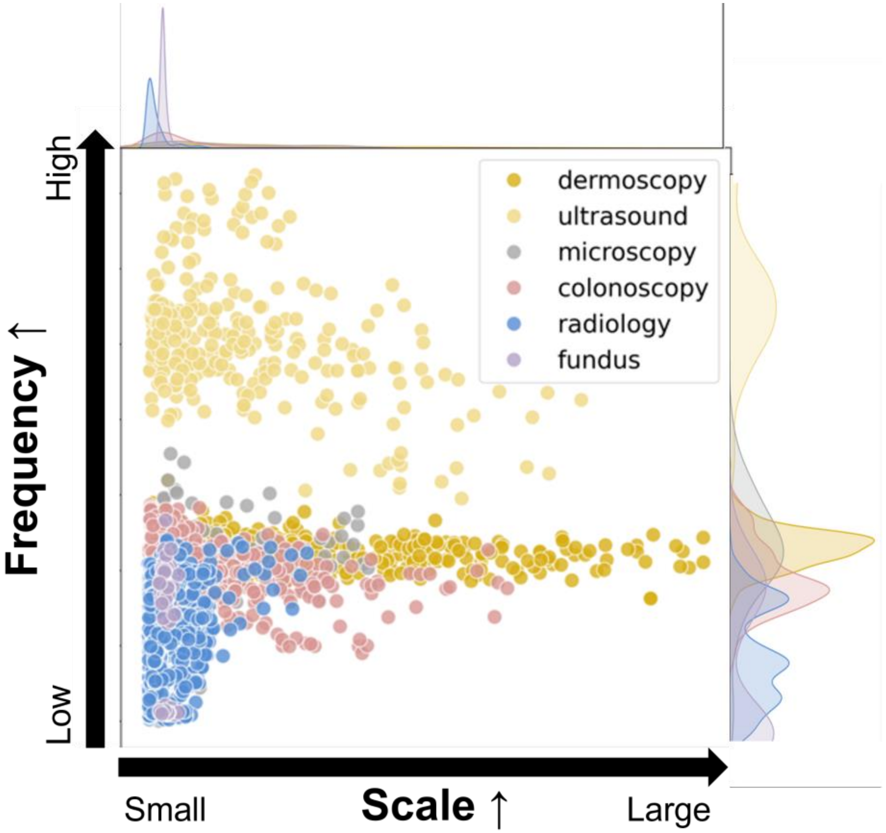


Figure 2. FRCU-Net with 1) Laplacian pyramid to take convolutional features to frequency domain and 2) frequency attention mechanism for a non-linearly weighted combination of all levels of the pramid.

- ✓ Multi-Frequency Recalibration Module
- ✓ Laplacian Pyramid-based Method
- ✓ Hard to capture various lesion

Azad, Reza, et al. "Deep frequency re-calibration u-net for medical image segmentation." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

Observations



Scale vs **Frequency** distribution per modality

Observations

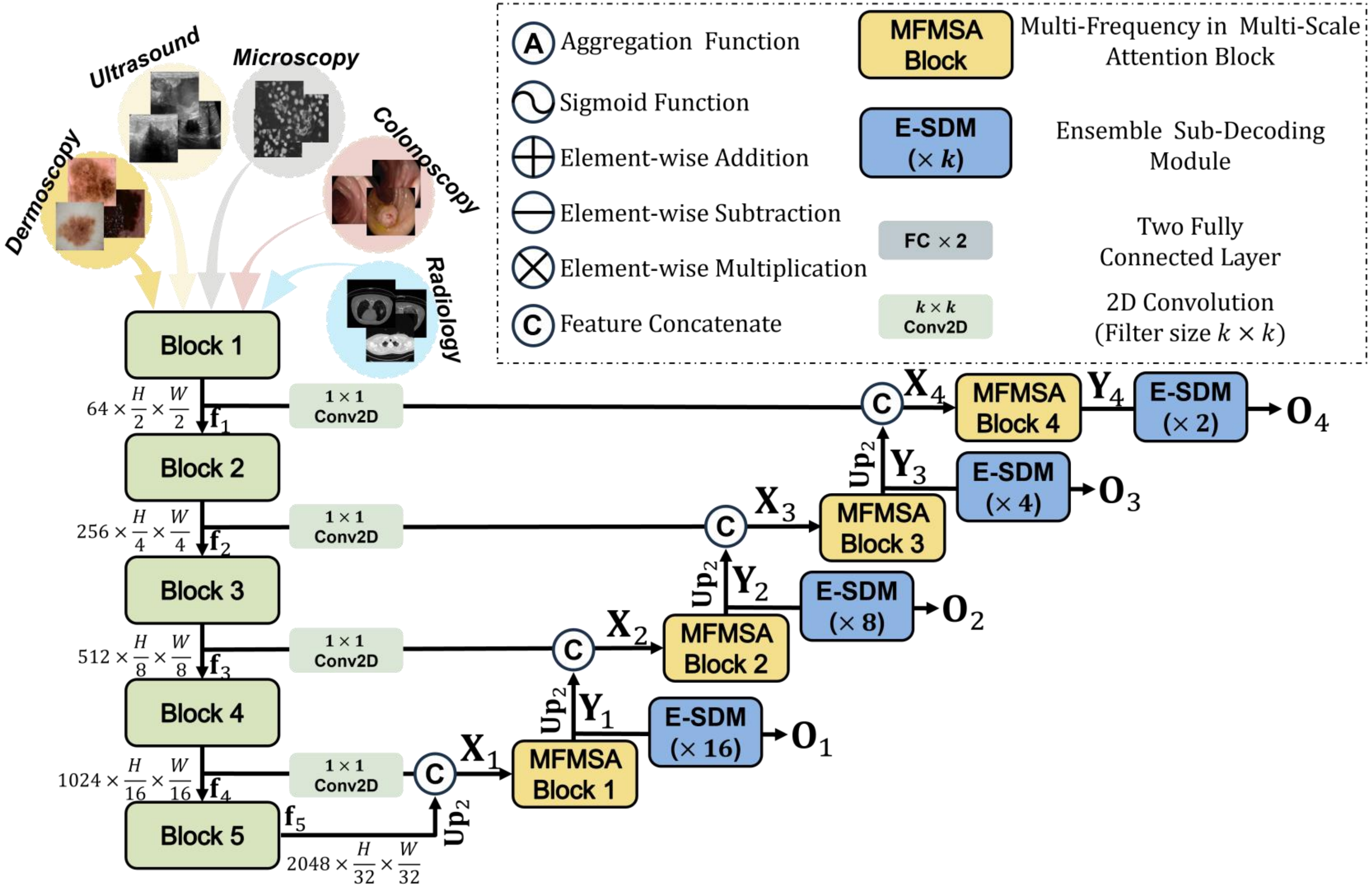
➢Frequency variance is higher than scale variance, which previous papers mainly focused on.

* Frequency = ratio of the high-frequency and full-frequency
* Scale = The size of lesions

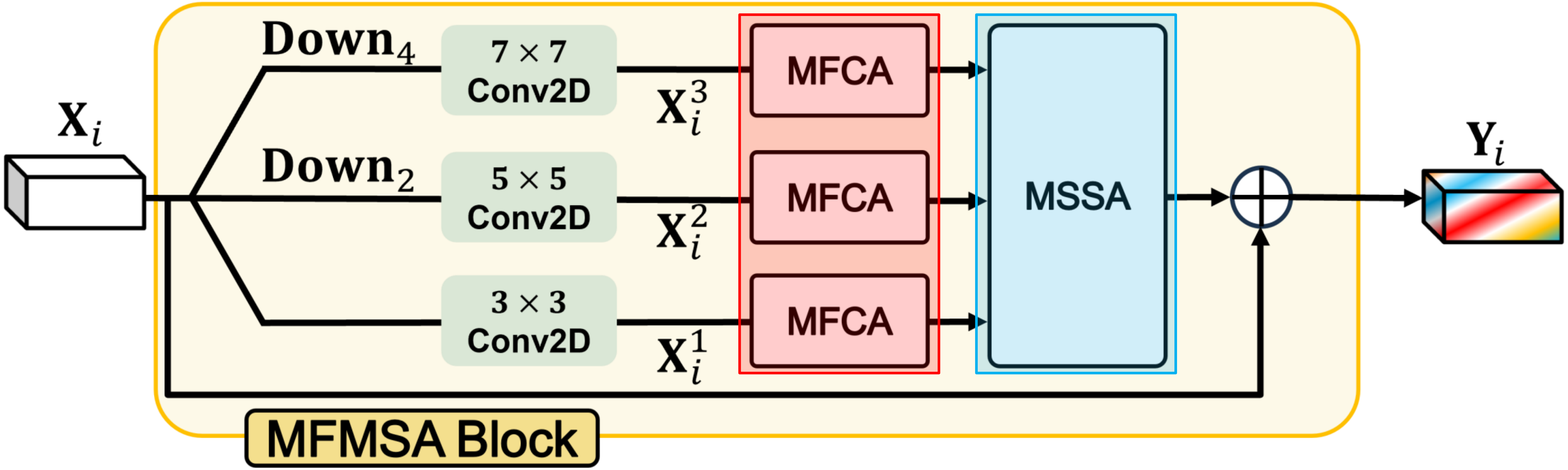
Motivations

- Human vision seamlessly combines scales and frequencies for interpreting the environment.
- Since medical images contains various lesion sizes, it requires multi-scale features for precise segmentation
- As medical images show higher frequency variance than scale, incorporating multi-frequency information is crucial for effective segmentation models.
- Upsampling low-resolution feature maps for loss calculation compromises model representation, leading to information loss in predicting details.

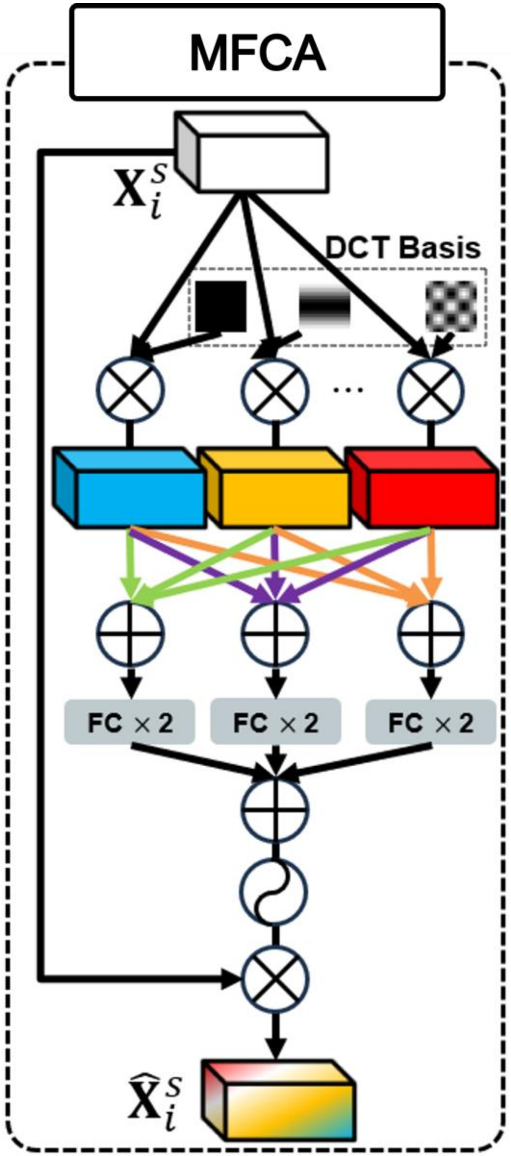
Modality-Agnostic Domain Generalizable Network



Multi-Frequency in Multi-Scale Attention Block



Multi-Frequency Channel Attention



- ✓ DCT-based Channel Attention Module

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

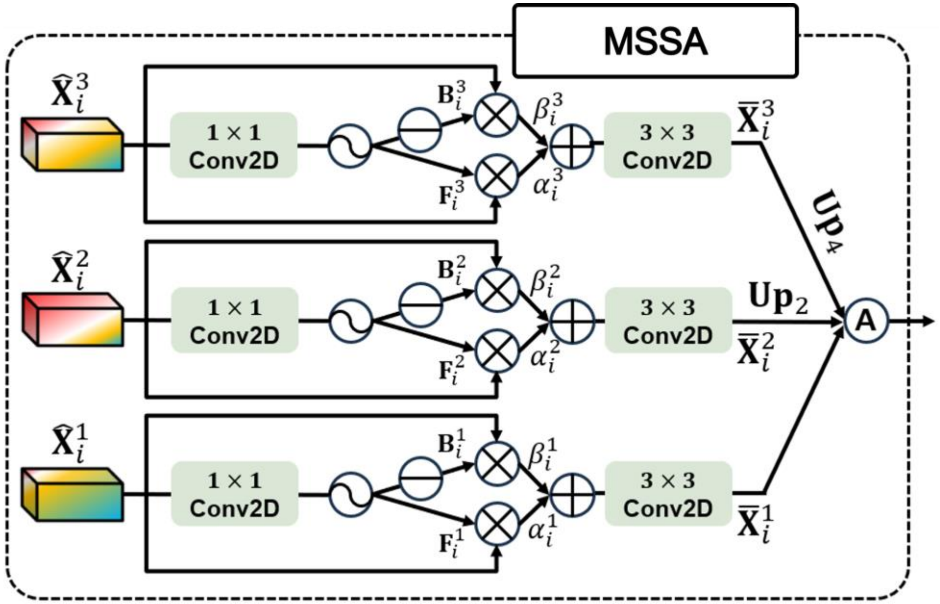
- ✓ Extract various statistic feature for suppressing noise effect

$$\mathbf{M}_i^s = \sigma\Big(\sum_{d\in\{\text{avg},\text{max},\text{min}\}} \mathbf{W}_2(\delta(\mathbf{W}_1 \mathbf{Z}_d))\Big)$$

- ✓ Recalibrate the feature map at s-th scale

$$\hat{\mathbf{X}}_i^s = \mathbf{X}_i^s \times \mathbf{M}_i^s$$

Multi-Scale Spatial Attention



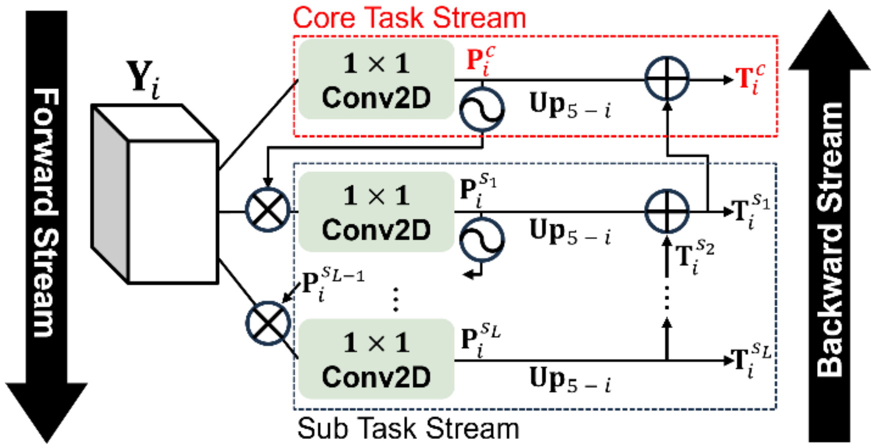
✓ Introduce learnable parameters to control information flow

$$\bar{\mathbf{X}}_i^s = \mathbf{Conv2D}_3(\alpha_i^s(\hat{\mathbf{X}}_i^s \times \mathbf{F}_i^s) + \beta_i^s(\hat{\mathbf{X}}_i^s \times \mathbf{B}_i^s))$$

✓ Aggregate each refined feature from different scale branch

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

Ensemble Sub-Decoding Module



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 $\{\mathbf{T}_i^{s_1}, \dots, \mathbf{T}_i^{s_L}\}$ at i-th decoder

```
P_i^c = Conv2D_1(Y_i)
for l = 1, 2, ..., L do
    P_i^{s_l} = Conv2D_1(Y_i × σ(P_i^{s_{l-1}})).
end for
T_i^{s_L} = Up_{5-i}(P_i^{s_L})
for l = L-1, ..., 0 do
    T_i^{s_l} = Up_{5-i}(P_i^{s_l}) + T_i^{s_{l+1}}
end for
return O_i = {T_i^c, T_i^{s_1}, ..., T_i^{s_L}}
```

Why Ensemble?

$$\mathbf{T}_i^{c} = \mathbf{T}_i^{s_0} = \mathbf{Up}_{5-i}\left(\mathbf{P}_i^{s_0}\right) + \mathbf{T}_i^{s_1}$$
$$= \left[\mathbf{Up}_{5-i}\left(\mathbf{P}_i^{s_0}\right) + \mathbf{Up}_{5-i}\left(\mathbf{P}_i^{s_1}\right)\right] + \mathbf{T}_i^{s_2}$$
$$= \cdots$$
$$= \sum_{l=0}^{L} \mathbf{Up}_{5-i}\left(\mathbf{P}_i^{s_l}\right)$$

- Our decoder has **an ensemble effect** as it **aggregates predictions of different tasks** for the same legion.

Loss Function

Structure Loss Functions with 4 Stage Deep Supervision

$$\mathcal{L}_{total} = \sum_{i=1}^{4} \sum_{t \in \{c, s_1, s_2, \dots, s_L\}} \lambda_t \mathcal{L}_t(\mathbf{G}^t, \mathbf{Up}_{5-i}(\mathbf{T}_i^t))$$

- $\mathcal{L}_t$: Loss function for the task t
  1. Region Prediction (Core Task) Loss: $\mathcal{L}_R = \mathcal{L}_{IoU}^{w} + \mathcal{L}_{bce}^{w}$
  2. Boundary Prediction (Sub Task 1) Loss: $\mathcal{L}_B = \mathcal{L}_{bce}$
  3. Distance Map Prediction (Sub Task 2) Loss: $\mathcal{L}_D = \mathcal{L}_{mse}$

Experiment

✓ Quantitative Results for *Seen* Clinical Settings

Method	Dermoscopy		Radiology		Ultrasound		Microscopy		Colonoscopy				*P*-value
	ISIC2018 [23]		COVID19-1 [33]		BUSI [3]		DSB2018 [6]		CVC-ClinicDB [5]		Kvasir [31]		
	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	
UNet [52]	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 [45]	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++ [74]	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 [57]	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 [67]	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 [73]	***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. $(\cdot)$ denotes the standard deviations of multiple experiment results.

Experiment

CVPR
JUNE 17-21, 2024
SEATTLE, WA

✓ Quantitative Results for *Unseen* Clinical Settings

Method	Dermoscopy		Radiology		Ultrasound		Microscopy		Colonoscopy						P-value
	PH2 [42]		COVID19-2 [1]		STU [75]		MonuSeg2018 [12]		CVC-300 [62]		CVC-ColonDB [58]		ETIS [55]		
	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	
UNet [52]	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 [45]	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++ [74]	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 [57]	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 [67]	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 [73]	*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. $(\cdot)$ denotes the standard deviations of multiple experiment results.

Experiment

✓ Qualitative Results

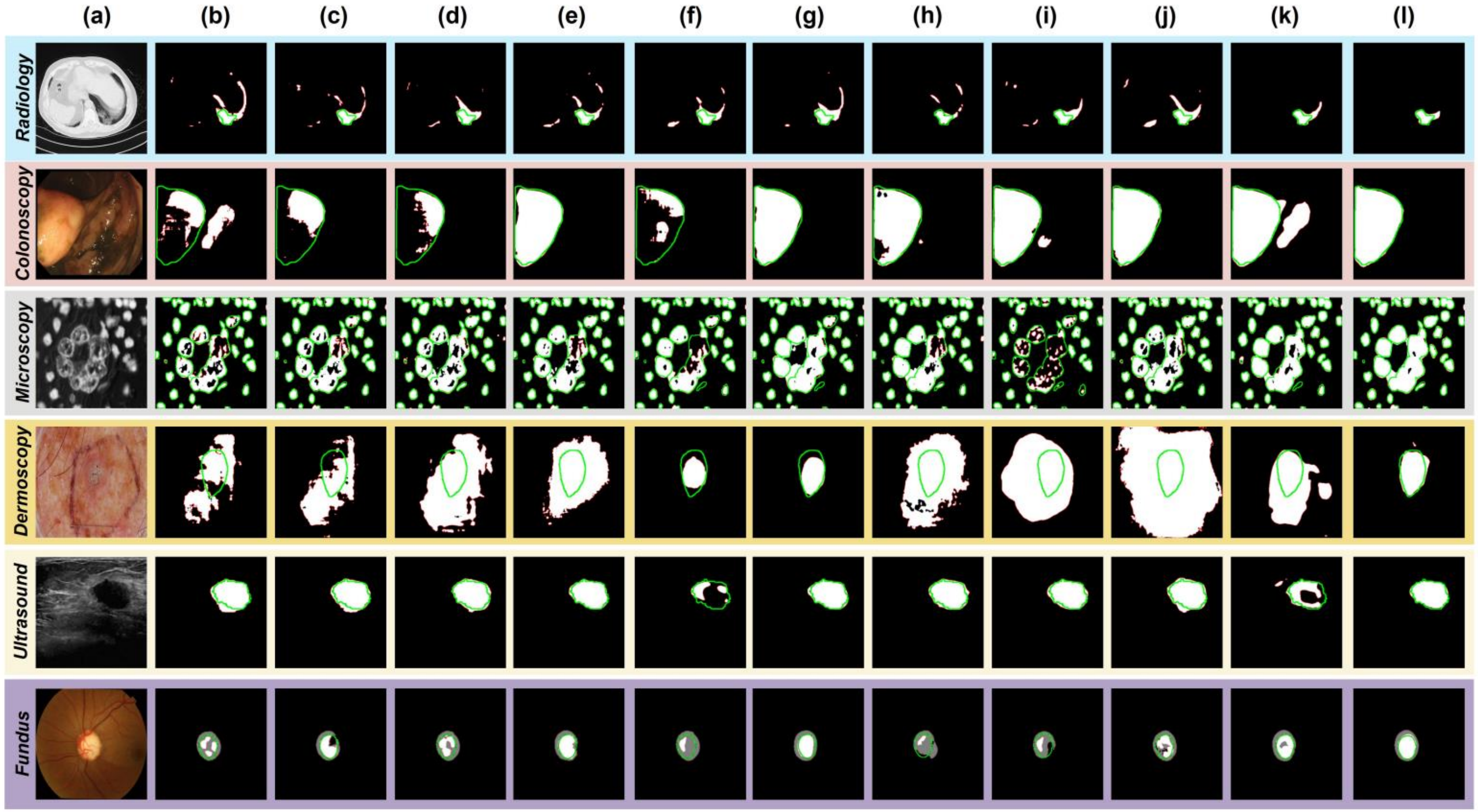


Figure 5. Qualitative comparison of other methods and MADGNet. (a) Input images. (b) UNet [52]. (c) AttUNet [45]. (d) UNet++ [74]. (e) CENet [22]. (f) TransUNet [7]. (g) FRCUNet [4], (h) MSRFNet [57]. (i) HiFormer [26]. (j) DCSAUNet [67]. (k) M2SNet [73]. (l) **MADGNet (Ours)**. **Green** and **Red** lines denote the boundaries of the ground truth and prediction, respectively.

Experiment

✓ Ablation Study: Effectiveness of Multi-Scale & Multi-Frequency Attention

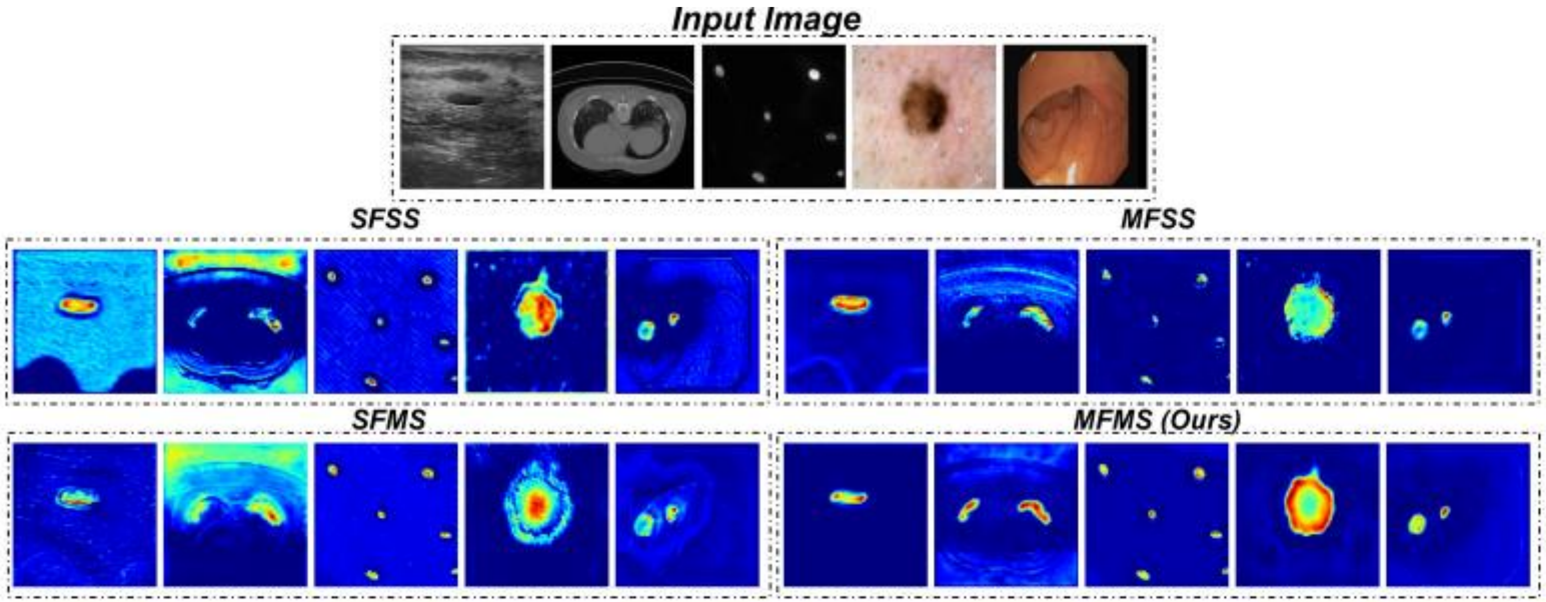


Figure 6. Feature visualization of SFSS, MFSS, SFMS, **MFMS**.

Experiment

✓ Ablation Study: Effectiveness of Ensemble Sub-Decoding Module

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, 31]) and *unseen* ([55, 58, 62]) 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.

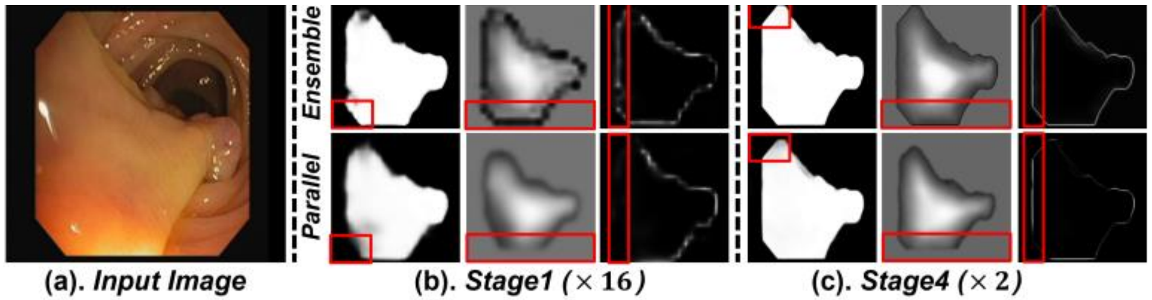


Figure 7. Qualitative results between ensemble and parallel manners. (a) Input Image, (b) and (c) Predictions from Stage1 ($\times$16 **Up**) and Stage4 ($\times$2 **Up**). First and second rows in (b) and (c) are predictions with **ensemble (Ours)** and parallel manners.

Conclusion

- We propose MADGNet, leveraging the benefits of multi-scale and multi-frequency features, which are crucial for effective medical image segmentation.
- MFMSA enhances boundary cues extraction, improving segmentation accuracy.
- E-SDM mitigates information loss during multi-task learning, enhancing segmentation performance.

Thank you