I-MedSAM: Implicit Medical Image Segmentation with Segment Anything

Xiaobao Wei[1,2,3,*], Jiajun Cao[1,4,*], Yizhu Jin[1,5],
Ming Lu[1], Guangyu Wang[6], and Shanghang Zhang[1,†]

[1]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University [2]University of Chinese Academy of Sciences
[3]Institute of Software, Chinese Academy of Sciences [4]Xi'an Jiaotong University
[5]Beihang University [6]Beijing University of Posts and Telecommunications
weixiaobao0210@gmail.com

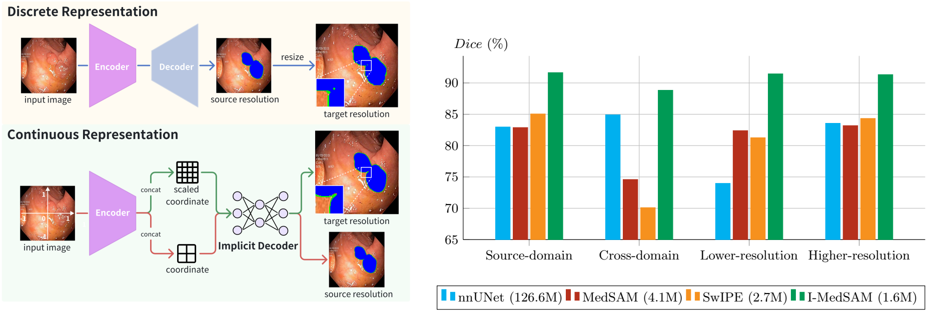
Abstract. With the development of Deep Neural Networks (DNNs), many efforts have been made to handle medical image segmentation. Traditional methods such as nnUNet train specific segmentation models on the individual datasets. Plenty of recent methods have been proposed to adapt the foundational Segment Anything Model (SAM) to medical image segmentation. However, they still focus on discrete representations to generate pixel-wise predictions, which are spatially inflexible and scale poorly to higher resolution. In contrast, implicit methods learn continuous representations for segmentation, which is crucial for medical image segmentation. In this paper, we propose I-MedSAM, which leverages the benefits of both continuous representations and SAM, to obtain better cross-domain ability and accurate boundary delineation. Since medical image segmentation needs to predict detailed segmentation boundaries, we designed a novel adapter to enhance the SAM features with high-frequency information during Parameter-Efficient Fine-Tuning (PEFT). To convert the SAM features and coordinates into continuous segmentation output, we utilize Implicit Neural Representation (INR) to learn an implicit segmentation decoder. We also propose an uncertainty-guided sampling strategy for efficient learning of INR. Extensive evaluations on 2D medical image segmentation tasks have shown that our proposed method with only 1.6M trainable parameters outperforms existing methods including discrete and implicit methods. The code will be available at: https://github.com/ucwxb/I-MedSAM.

Keywords: Medical Image Segmentation · Implicit Neural Representation · Segment Anything

1 Introduction

Medical image segmentation, as a pivotal component of auxiliary disease diagnosis, holds a crucial role in medical image applications. The advent of deep

[*] Equal Contribution.
[†] Corresponding Author.



(a) Illustrations of discrete and continuous representation for medical segmentation.

(b) Segmentation quality comparison between our I-MedSAM and baselines.

Fig. 1: (a) Continuous representation with implicit decoders exhibits superior scale flexibility. (b) I-MedSAM with the fewest trainable params (1.6M) surpasses the state-of-the-art discrete and implicit approaches and exhibits a solid generalization ability when facing data shifts. Please refer to Sec. 4 for more experiment details.

learning has spurred the widespread adoption of neural networks customized for medical images. For example, nnUNet [37] leverages the downsampling and upsampling modules to aggregate multi-scale contextual features. Transformers [41] uses the self-attention mechanism to significantly augment the representation capacity of deep neural networks, improving the accuracy in medical image segmentation [5]. Recent advancements have witnessed the integration of foundation models as backbones in various works. The Segment Anything Model (SAM) [24] demonstrates unprecedented zero-shot segmentation ability. Therefore, diverse adapters based on parameter-efficient fine-tuning (PEFT) are crafted to fine-tune SAM for medical images [29, 42, 43, 46].

Despite their notable effectiveness, these methods primarily focus on pixel-wise or voxel-wise predictions [10, 13, 19, 29, 37, 46]. While they achieve promising results, the discrete representations present challenges in spatial flexibility and introduce discretization artifacts when scaling to arbitrary input sizes. Additionally, the discrete representations give rise to ambiguity when extracting the nuanced details crucial for precise boundary delineation [31], which is important in medical image analysis. The delineation of boundaries can signify the transitions between different human tissues or anatomical structures, thus providing essential information for accurately separating these instances. Usually, the intricacy and subtlety of this delineation process need additional refinement [1, 34].

Compared with discrete representations, continuous representations learn Implicit Neural Representations (INRs) [32] to transform discrete representations into continuous space. As shown in Fig. 1a, numerous approaches learn a mapping from encoded image features and grid coordinates to the segmentation output, enabling adaptability to various output resolutions [22, 33, 36, 47]. However, current approaches show unsatisfying domain transfer ability due to the limited representation capabilities of their pre-trained image encoders. Additionally, the boundary information of images demonstrates a strong correlation with features in the frequency domain [11, 27], which is also ignored by most previous methodologies. Lastly, existing methods adopt random sampling across coordinates, underestimating the influence of sampling strategies when learning INRs.

To address the aforementioned limitations, we propose I-MedSAM, a model that leverages the benefits of both continuous representations and SAM, aiming to enhance cross-domain capabilities and achieve precise boundary delineation. Given the medical images, I-MedSAM extracts the features from SAM with the proposed frequency adapter, which aggregates high-frequency information from the frequency domain. These features along with the grid coordinates are decoded into segmentation outputs by the learned INRs. We employ a two-stage implicit segmentation decoder, consisting of two INRs in a coarse-to-fine manner. The first INR produces coarse segmentation results and features. Subsequently, a novel uncertainty-guided sampling strategy is applied to sample *Top-K variance* feature points along with their corresponding grid coordinates. Finally, these selected samples are fed into the second INR to obtain refined segmentation results. Notably, I-MedSAM is trained end-to-end with a minimal number of trainable parameters, yet it achieves state-of-the-art performance compared with all baseline methods. Our main contributions are summarized as follows:

- We propose I-MedSAM, a novel method that leverages the advantages of SAM and continuous representations.
- We propose a novel frequency adapter that utilizes high-frequency information to enhance features, thereby accurately segmenting boundaries.
- We propose a novel coarse-to-fine INR decoder with an uncertainty-guided sampling (UGS) strategy, to learn a mapping from features and coordinates to segmentation output.
- We perform detailed evaluations of I-MedSAM on 2D medical image segmentation. As shown in Fig. 1b, I-MedSAM outperforms state-of-the-art continuous and discrete methods. Experiments also demonstrate that I-MedSAM is robust to scale and domain shifts.

2 Related Work

Implicit Neural Representation. The concept of signal representation is fundamental across various domains, especially in the field of computer vision [31, 32]. Traditional methods for encoding signals discretize the input space into pixel or voxel grids [4,6,7,9,10,15,19,37,38,46]. Different from these discrete methods, Implicit Neural Representation (INR) learns generator functions that map input coordinates into the signal values [8]. Numerous studies employ INR for diverse tasks, including medical data reconstruction, rendering, compression, registration, super-resolution and segmentation [2, 25, 30, 32, 44]. For segmentation, conventional methods typically consist of a trained feature encoder and a decoder. The encoder encodes medical data into features, and the decoder subsequently decodes features along with their coordinates into segmentation output [17, 22, 23, 36, 39, 40, 45, 47]. However, current methods exhibit an imbalance in emphasizing either global or local features and demonstrate relatively low out-of-distribution ability. In contrast, our proposed approach leverages the segmentation foundation model SAM to enrich feature extraction. Moreover, we introduce an innovative uncertainty-guided sampling (UGS) strategy into the INR, enabling the adaptive selection of samples to train the implicit segmentation decoder. Please refer to the appendix for more related work.

3 Methodology

In this section, we initially provide a concise overview of the implicit image segmentation problem. Then, we proceed to elaborate on the pipeline of I-MedSAM. Finally, we elucidate the novel designs introduced in I-MedSAM.

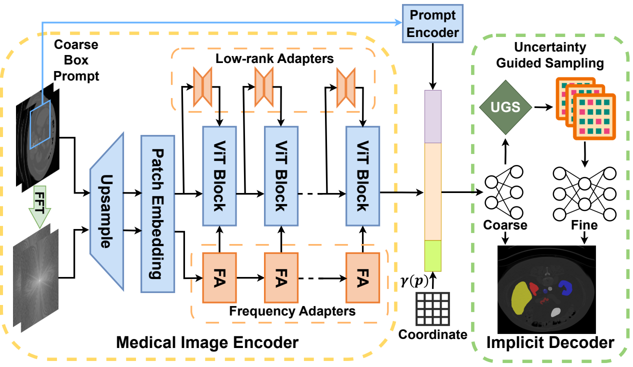


Fig. 2: The overall pipeline of I-MedSAM. First, given the medical images and a coarse bounding box as a prompt, I-MedSAM utilizes the medical image encoder and the prompt encoder to generate discrete features. For the medical image encoder, we design low-rank adapters and frequency adapters to extract information from the spatial domain and frequency domain. Then I-MedSAM interpolates all features to align with the encoded coordinates and decodes them in coarse to fine neural fields. We propose an Uncertainty Guided Sampling (UGS) strategy to adaptively choose the highest variance points and refine predictions. I-MedSAM merges the predictions from coarse and fine neural fields as the final prediction maps.

3.1 Preliminaries

In traditional discrete segmentation with C classes, neural networks aim to learn a direct mapping from input medical images $X \in \mathbb{R}^{H\times W\times 3}$ to class probabilities $O \in \mathbb{R}^{H\times W\times C}$ at the same resolution, in which H and W means height and width of input images, respectively.

On the other hand, implicit image segmentation seeks to map each pixel of medical images X with its coordinate $p_i = (x, y)$, where $x, y \in [-1, 1]$, to class probabilities $\hat{o}_i \in \mathbb{R}^C$, denoted as $\mathcal{N}_\theta : (p_i, X_i) \rightarrow \hat{o}_i$. Here, $\mathcal{N}_\theta$ represents a neural network parameterized by weights θ. This formulation incorporates coordinates directly on pixels, adjusting the spatial granularity of input coordinates for predictions at arbitrary resolution, from source resolution $H \times W$ to target resolution $H' \times W'$, which can be represented as $X \in \mathbb{R}^{H \times W \times 3} \rightarrow O \in \mathbb{R}^{H' \times W' \times C}$. Moreover, it allows the direct application of pixel-wise loss functions like Cross Entropy or Dice. Additionally, the zero-isosurface in $\mathcal{N}_\theta$'s implicit space represents object boundaries, providing an additional advantage for boundary modeling.

3.2 Overall Pipeline

As depicted in Fig. 2, I-MedSAM comprises two main parts. The first part integrates an image encoder with its adapters, forming Enc_I, and a prompt encoder Enc_P, following SAM's design. Specifically, recognizing the significant role of the frequency domain in segmentation boundary representation, a frequency adapter is devised for extracting frequency features. Taking a medical image and a prompt bounding box as inputs, multi-scale features are extracted from both spatial and frequency domains. In scenarios involving cross-resolution, the extracted features need to be interpolated from the source resolution to achieve segmentation output at the target resolution.

The second part is the implicit segmentation decoder Dec, comprising two stacked INRs: one "coarse" Dec_c with shallow layers and one "fine" Dec_f with deeper layers. Typically, Dec_c generates a coarse segmentation map, and Dec_f refines it on sampled points. The selection of these points is determined by the pixel-wise uncertainty of segmentation predictions, assessed through MC-Dropout and Top-K algorithms. Detailed explanations of these two parts will be provided in the following sections.

3.3 Medical Image Encoder

In this section, we introduce the frequency adapter and low-rank adapter integrated into SAM, to extract features from both frequency and spatial domains.

Frequency Adapter. Discrete Fourier Transform (DFT) is a common and effective method for transforming an image into the frequency domain. In practice, the Fast Fourier Transform (FFT) is employed for efficient computation of DFT, the spectrum representation of $f_{h,w}$ can be formulated as:

$$\mathcal{F}_{u,v} = \sum_{h=1}^{H} \sum_{w=1}^{W} f_{h,w} \cdot \mathrm{e}^{-j2\pi\left(\frac{h}{H}u + \frac{w}{W}v\right)}. \tag{1}$$

Subsequently, the amplitude and phase spectrum of $\mathcal{F}_{u,v}$ can be obtained as $|\mathcal{F}_{u,v}|$ and $arg(\mathcal{F}_{u,v})$, respectively. Experiment results in Tab. 6 indicate that the amplitude spectrum exhibits superior representation ability compared to the phase spectrum. Therefore, we default to using the amplitude spectrum for our proposed frequency adapter (FA).

As illustrated in Fig. 3, the individual FA comprises a linear down-projection layer, a GELU activation layer, and a linear up-projection layer. In total, we utilize n instances of FA as a sequence, corresponding to the number of Vision Transformer (ViT) Blocks of Enc_I.

Low-Rank Adapter. In contrast to fine-tuning all parameters in the image encoder Enc_I, we leverage the Low-Rank Adapter (LoRA) [16] to update a small fraction of parameters, adapting SAM to medical images, as illustrated in Fig. 3. Given the encoded token sequence $F \in \mathbb{R}^{B\times N\times C_{in}}$, the resulting token sequence $\hat{F} \in \mathbb{R}^{B\times N\times C_{out}}$ is generated using a projection layer $W_p \in \mathbb{R}^{C_{out}\times C_{in}}$, denoted as $\hat{F} = W_pF$. LoRA proposes that the adjustment to W_p should be gradual and consistent. It recommends utilizing a low-rank approximation $A \in \mathbb{R}^{r\times C_{in}}$ and $B \in \mathbb{R}^{C_{out}\times r}$ to represent this gradual update, which can be formulated as:

$$\hat{W}_p = W_p + \Delta W_p = W_p + BA. \tag{2}$$

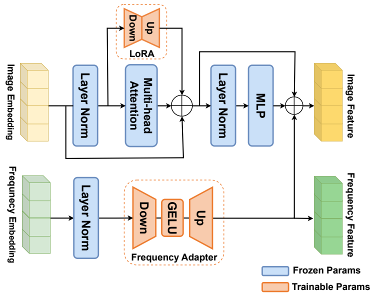


Fig. 3: Illustration of the proposed frequency adapter and LoRA in the image encoder. The image/frequency embedding from patch embedding undergoes two separate branches in the encoder.

As the multi-head attention mechanism determines the regions to focus on, it is reasonable to apply LoRA to the frozen projection layers of query Q, key K, or value V to influence the attention scores. We notice that I-MedSAM performs better when LoRA is applied to the query Q and value V projection layers, which can be expressed as:

$$\begin{cases} Q = \hat{W}_qF = W_qF + B_qA_qF \\ K = W_kF \\ V = \hat{W}_vF = W_vF + B_vA_vF \end{cases} \tag{3}$$

where W_q, W_k and W_v are frozen projection layers from SAM's image encoder, and A_q, B_q, A_v, B_v are trainable LoRA parameters.

3.4 Implicit Segmentation Decoder

In this section, we introduce a coarse-to-fine implicit neural representation with an uncertainty-guided sampling (UGS) strategy to decode features from encoders into the segmentation maps at target resolutions.

Coarse to Fine Implicit Neural Representation. Given features from the image encoder Enc_I and the prompt encoder Enc_P, we interpolate them from source to target resolutions and concatenate them with coordinates p. These coordinates p are generated at the target resolutions and normalized to $[-1, 1]$. To address potential biased learning resulting from the direct use of input coordinates [35], we encode the coordinates into a higher-dimensional space using a high-frequency positional encoding function, which is defined as:

$$\gamma(p) = (\sin(2^0\pi p), \cos(2^0\pi p), \cdots, \sin(2^{L-1}\pi p), \cos(2^{L-1}\pi p)\,) \tag{4}$$

where the hyperparameter L is set to 10 in our experiments following the previous work. The encoded coordinates, the encoded features from both the image and prompt encoders are concatenated to feed into the decoder:

$$Z^p = Concat(\gamma(p), Interp(Enc_I(X)), Enc_I(P)). \tag{5}$$

Here, X and P represent the input medical image and the corresponding coarse bounding box prompt, respectively. The function $Interp$ refers to the interpolation function based on bilinear algorithms, which is used to interpolate the encoded features from source to target resolution, in alignment with the encoded coordinates.

Inspired by NeRF [31], we depart from the one-stage INR approaches to introduce a two-stage decoding process. This involves optimizing two INRs simultaneously: one "coarse" Dec_c, with shallow layers, and one "fine" Dec_f, with deeper layers. Dec_c produces a coarse segmentation map, $\hat{o}_i^c$, serving as reference for Dec_f to refine. Additionally, Dec_c generates coarse features, z_i^c, employed by Dec_f in its refinement process.

We employ MC-dropout to calculate the uncertainty of features $\hat{o}_i^c$ for each pixel. Subsequently, a Top-K percentage of feature points is sampled based on this uncertainty, denoted as z_i^s (with $s \in \mathcal{S}$). Finally, the predictions from the "coarse" and "fine" INRs are combined to produce the output of I-MedSAM. The decoding process is formulated as follows:

$$\begin{aligned}
\hat{o}_i^c,\, z_i^c &= Dec_c(z_i^p) \\
z_i^s &= UGS(z_i^c),\, s \in \mathcal{S} \\
\hat{o}_i^f &= Dec_f(z_i^s) \\
\hat{O} &= \hat{O}^c(\mathcal{S} \setminus s) \cup \hat{O}^f(s)\,, \hat{o}_i \in \hat{O}.
\end{aligned} \tag{6}$$

Here, UGS represents Uncertainty Guided Sampling, which will be further illustrated in the following section.

Uncertainty Guided Sampling. In the sampling process, we select feature points that require refinement from the "coarse" INR Dec_c and feed them into the "fine" INR Dec_f, based on uncertainty estimation. Drawing inspiration from MC-Dropout methods [12,14], we apply dropout T times to obtain T prediction results of coarse segmentation probabilities, $\{o_i^c\}_{t=1}^T$, given the input features z_i^p, denoted as $\{p_t(o_i^c|z_i^p)\}_{t=1}^T$. The uncertainty is calculated as the variance of predictions for each feature point, expressed as:

$$\begin{cases} \mu_i = \dfrac{1}{T}\displaystyle\sum_{t=1}^{T}(p_t(o_i^c|z_i^p)) \\ u_i = \dfrac{1}{T}\displaystyle\sum_{t=1}^{T}(p_t(o_i^c|z_i^p) - \mu_i)^2. \end{cases} \tag{7}$$

Subsequently, we sample the feature points with the highest Top-K percentage uncertainty to form z_i^s for Dec_f to refine. This estimation of uncertainty reflects the variation in prediction difficulty among different samples. It adaptively selects pixels with higher difficulty for refinement by Dec_f, achieving more accurate segmentation results.

3.5 Training I-MedSAM

To optimize the trainable parameters of I-MedSAM, we freeze the pre-trained image encoder Enc_I, while only unfreezing the proposed adapters, prompt encoder Enc_P and INRs. We utilize SAM's image encoder with LoRA and our proposed frequency adapter to extract features for input medical images X, while extracting prompt features in terms of coarse bounding box P for targeted segmentation objects. The coarse bounding box P is randomly adjusted the height and width following the previous work [46]. Then we concatenate features along with the mapped coordinate values and decode them with the proposed two-stage INR decoder. With the coarse to fine INR and the uncertainty guided sampling strategy, I-MedSAM obtains the coarse and refined point-wise segmentation probabilities $\{\hat{o}_i^c, \hat{o}_i^f\}_{i \in X}$, combined as $\hat{o}_i$. For training optimization, we adopt pixel-wise segmentation loss, which can be formulated as:

$$L_{seg}(o_i, \hat{o}_i) = 0.5 \cdot L_{ce}(o_i, \hat{o}_i) + 0.5 \cdot L_{dc}(o_i, \hat{o}_i) \tag{8}$$

where L_{ce} and L_{dc} stand for Cross Entropy loss and Dice loss respectively. We apply the loss to supervise both coarse and refined segmentation maps progressively. Within the training process, we decrease weights for coarse supervision and increase weights for refined supervision until I-MedSAM converges.

4 Experiments

In this section, we present extensive experiments to evaluate the effectiveness of I-MedSAM for medical image segmentation. We first introduce the experimental settings including datasets and training details. Then we compare our method with the SOTA implicit and discrete approaches on the binary polyp segmentation [20] and multi-class organ segmentation [26] qualitatively and quantitatively. We further evaluate the performance and robustness of I-MedSAM when facing data shifts. Finally, we conduct a comprehensive ablation study to evaluate the contribution of each component. Due to space limitations, we provide more details and visualization results in the supplementary material.

Table 1: Overall segmentation results versus the state-of-the-art discrete approaches and implicit approaches. The Trainable Params columns report unfrozen parameters in training and the Dice columns report averaged scores with standard deviation.

Binary Polyp Segmentation			Multi-class Organ Segmentation		
Method	Dice (%)↑	Trainable Params (M)↓	Method	Dice (%)↑	Trainable Params (M)↓
Discrete Approaches					
U-Net [37]	63.89±1.30	7.9	U-Net [37]	74.47±1.57	16.3
PraNet [10]	82.56±1.08	30.5	UNETR [15]	81.14±0.85	92.6
Res2UNet [13]	81.62±0.97	25.4	Res2UNet [13]	79.23±0.66	38.3
nnUNet [19]	82.97±0.89	126.6	nnUNet [19]	85.15±0.67	126.6
MedSAM [29]	82.88±0.55	4.1	MedSAM [29]	85.85±0.81	52.7
Implicit Approaches					
OSSNet [36]	76.11±1.14	5.2	OSSNet [36]	73.38±1.65	7.6
IOSNet [22]	78.37±0.76	4.1	IOSNet [22]	76.75±1.37	6.2
SwIPE [47]	85.05±0.82	2.7	SwIPE [47]	81.21±0.94	4.4
I-MedSAM (ours)	**91.49**±0.52	**1.6**	I-MedSAM (ours)	**89.91**±0.68	**3.5**

4.1 Experimental Settings

Datasets. We assess the performance of our model on two distinct tasks: binary polyp segmentation and multi-class abdominal organ segmentation. For binary polyp segmentation, we conduct experiments using the challenging **Kvasir-Sessile** dataset [20], consisting of 196 RGB images of small sessile polyps. Additionally, we evaluate the generalization capability of our model by testing the pre-trained I-MedSAM directly to the **CVC-ClinicDB** dataset [3], comprising 612 images from 31 colonoscopy sequences.

For multi-organ segmentation, training is conducted on the **BCV** dataset [26], comprising 30 CT scans with annotations for 13 organs. Model robustness is also evaluated using the pre-trained I-MedSAM on diverse CT images in **AMOS** [21] (200 training CTs, maintaining the same setup as in [48]). Since our work is dedicated to showing the effectiveness of 2D medical image segmentation, we just slice-wise segment CT data. Following the data prepossess in SwIPE [47], all datasets are divided with a train:validation:test ratio of 60:20:20, and the reported dice scores are all based on the test set.

Implementation Details. The training process of I-MedSAM involves fine-tuning the encoder of SAM [24], utilizing ViT-B as the backbone. We set LoRA ranks to 4 and incorporate amplitude information in the frequency adapters. For implicit segmentation decoders, we set latent MLP dimensions as $[1024, 512]$ for Dec_c and $[512, 256, 256, 128]$ for Dec_f. We sample the highest uncertainty points with a proportion of 12.5% and set the dropout probabilities to 0.5.

For multi-organ segmentation, we slightly modify the number of the last layer in Dec_c and Dec_f to match target segmentation classes. I-MedSAM are optimized by AdamW [28] with α=0.5, β=0.1, and λ_{ada}=5×10^{-5} for adapters in the encoder and λ_{dec}=1×10^{-3}. For fair comparisons, all methods are trained for 1000 epochs on the same experiment settings. The reported test dice score and Hausdorff Distance [18] correspond to the best validation epoch. Image input sizes are 384×384 for Sessile and 512×512 for each slice of BCV.

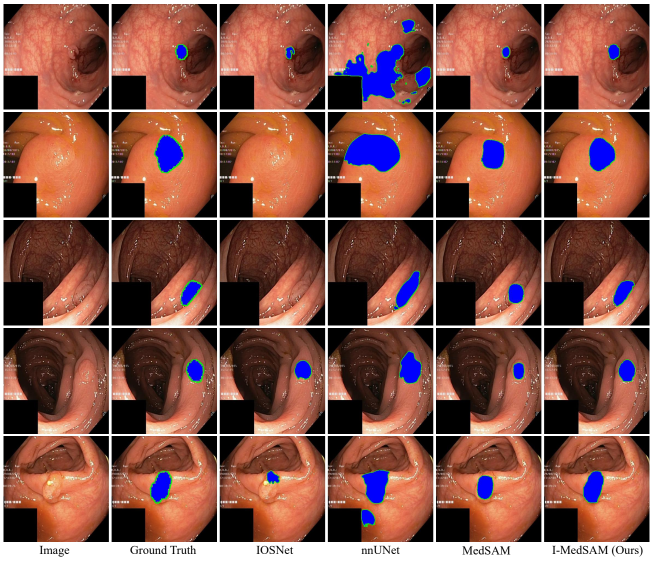


Fig. 4: Qualitative comparison on Kvasir-Sessile dataset for binary polyp segmentation.

Baselines. We divide baselines into two sets: discrete approaches and implicit (continuous) approaches. Discrete approaches include U-Net [37], PraNet [10], Res2UNet [13], nnUNet [19], UNETR [15] and MedSAM [29]. Specifically, MedSAM [29] is also a SAM-based approach, where SAM's original decoder is directly finetined. Implicit approaches include OSSNet [36], IOSNet [22] and SwIPE [47]. Given the absence of code availability from the prior state-of-the-art SwIPE [47] for extended comparison and the demonstrated superiority of IOSNet [22] over OSSNet [36] in implicit approaches, we primarily compare our method with IOSNet [22] in several experiment settings.

4.2 Quantitative Comparison

In this section, we first report the dice score compared with baselines. Then we conduct experiments across different resolutions and domains to evaluate the robustness and generalization ability under data shifts. Finally, we implement Hausdorff Distance (HD distance) [18] to compare the quality of the segmentation boundary with baselines across different experiment settings.

Segmentation Comparison. We make a comparison with both the discrete methods and implicit methods in terms of trainable parameters and Dice score with standard deviation on Polyp Sessile for binary-class segmentation and CT BCV for multi-class segmentation, as demonstrated in Tab. 1. On the smaller polyp dataset, we observe notable improvements over the best-known implicit methods and discrete methods with much fewer trainable parameters (1.26% of nnNet [19] and 59.26% of SwIPE [47]). For multi-organ segmentation on BCV, performance gains are also significant compared with the best-known

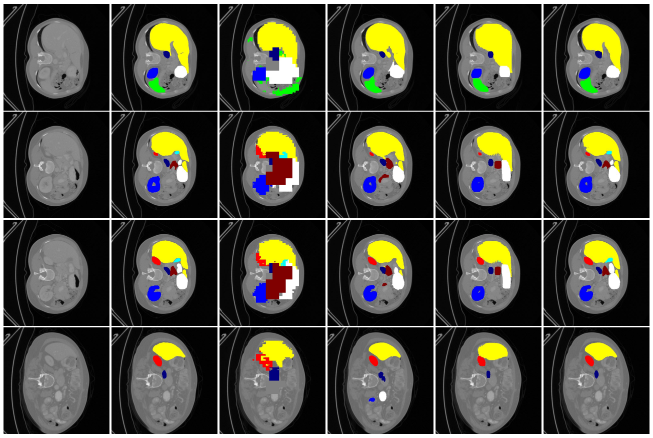


Fig. 5: Qualitative comparison on BCV dataset for multi-organ segmentation.

implicit methods and discrete methods. On the one hand, SAM, with the assistance of the proposed frequency adapters, generates abundant features that enhance the quality of segmentation boundaries. In contrast, SwIPE employs Res2Net-50 [13] as its backbone, which offers less sufficient features compared to SAM's, leading to inferior segmentation quality. On the other hand, with the proposed uncertainty-guided sampling in INR decoders, I-MedSAM adaptively selects and refines uncertain pixels with the highest variance, leading to more accurate segmentation maps.

Robustness under Data Shifts. We compare the robustness across resolutions and domains with the best discrete and implicit methods on binary-class polyp segmentation.

Table 2: Cross-resolution from 384×384 to 128×128, from 384×384 to 896×896 on Kvasir-Sessile.

Method		Dice (%)↑	
		384 → 128	384 → 896
Discrete Approaches			
PraNet	[10]	72.64	74.95
PraNet*	[10]	68.79	43.92
nnUNet	[19]	73.97	83.56
nnUNet*	[19]	65.34	76.36
MedSAM	[29]	82.39	83.19
MedSAM*	[29]	82.37	83.32
Implicit Approaches			
IOSNet	[22]	76.18	78.01
SwIPE	[47]	81.26	84.33
I-MedSAM (ours)		**91.45**	**91.33**

Firstly, to adapt the pre-trained I-MedSAM model, originally trained on standard 384×384 images, to different target resolutions such as 128×128 for lower resolutions and 896×896 for higher resolutions, the input coordinates (384×384) are scaled accordingly to match the target resolutions. Subsequently, the dice score is computed at these corresponding target resolutions. For discrete methods, the resolution of output images remains the same as that of the input images. We input the original images at their source resolution (384×384) and resize the output segmentation maps to the target resolution for evaluation. Additionally, we denote discrete baselines with the suffix $*$, where the original medical images are resized to the target resolution and then provided as input to these methods to directly generate segmentation at the target resolution.

As shown in Tab. 2, implicit methods exhibit spatial flexibility and consistently outperform discrete methods. Among implicit approaches, our I-MedSAM achieves the highest performance across various output resolutions. This superior performance can be attributed to the efficacy of the proposed frequency adapters and uncertainty-guided sampling, enhancing I-MedSAM's capability to provide accurate predictions at arbitrary resolutions.

Secondly, we investigate the robustness of model performance across different datasets for the same task. In the binary-class polyp segmentation task, all methods are pre-trained on the Kvasir-Sessile dataset and evaluated directly on the CVC dataset. Similarly, in the multi-class abdominal organ segmentation task, all methods are pre-trained on the BCV dataset and evaluated on the AMOS dataset, focusing solely on the liver class. As shown in Tab. 3, leveraging the generalization ability of SAM, I-MedSAM outperforms the top discrete method, achieving dice scores of 88.83% and 86.28% respectively. For additional visualization comparisons regarding data shifts, please refer to the supplementary materials.

Table 3: Cross-domain on binary polyp segmentation and multi-class abdominal organ segmentation.

Method	Dice (%)
Kvasir-Sessile → CVC	
PraNet [10]	68.37
nnUNet [19]	84.91
MedSAM [29]	74.59
IOSNet [22]	59.42
SwIPE [47]	70.10
I-MedSAM (ours)	**88.83**
BCV → AMOS	
UNETR [10]	81.75
nnUNet [19]	79.63
MedSAM [29]	71.98
IOSNet [22]	79.48
SwIPE [47]	82.81
I-MedSAM (ours)	**86.28**

Boundary Comparison. We further utilize the Hausdorff Distance (HD distance) [18] to assess segmentation boundary quality. I-MedSAM achieves a lower HD distance, indicating superior boundary quality. For more detailed boundary visualizations, please refer to Fig.4 and the supplementary materials.

Table 4: Hausdorff Distance comparison on various experiment settings.

HD distance (↓)	Kvasir-Sessile	Kvasir-Sessile → CVC	384 → 128	384 → 896	BCV	BCV → AMOS
nnUNet [19]	31.30	82.31	13.69	72.31	6.50	80.39
MedSAM [29]	21.53	30.15	8.04	51.82	10.62	52.14
IOSNet [22]	51.72	81.60	35.33	87.86	21.46	61.19
I-MedSAM(Ours)	**11.59**	**19.76**	**7.91**	**32.77**	**5.95**	**37.53**

4.3 Qualitative Comparison

As shown in Fig. 4 and Fig. 5, we conduct qualitative comparisons on Kvasir-Sessile and BCV datasets. Due to the unavailability of code for inference, SwIPE has been omitted from the visual comparison. We also provide input medical images along with corresponding ground truth segmentation masks. The segmentation boundaries are delineated by green lines in Fig. 4. From the figures, it can be witnessed that I-MedSAM obtains better segmentation boundaries. Thanks to the proposed frequency adapters and uncertainty guided sampling techniques, I-MedSAM can efficiently aggregate high-frequency information from the input, which is beneficial to the accuracy of final segmentation maps. Due to the space limitation, please refer to supplementary materials for more qualitative results.

4.4 Ablation Study

We conduct ablation studies focusing on three aspects: component-wise ablations, the incorporation of frequency adapter, and point numbers for sampling. In each of our ablation experiments, other hyper-parameters remain consistent with the implementation details.

Table 5: Effectiveness of each component of the pipeline. We evaluate the Dice metric for both cross-domain and cross-resolution tasks.

LoRA	FA	INR	Kvasir-Sessile	Cross-domain	Cross-resolution	
				Kvasir-Sessile → CVC	384 → 128	384 → 896
✓			83.61	82.57	72.73	76.46
✓	✓		88.74	82.61	75.69	78.59
✓		✓	88.83	83.40	88.16	88.43
✓	✓	✓	**91.49**	**88.83**	**91.45**	**91.33**

Component-wise ablations. To demonstrate the effectiveness of each component, we conduct a component-wise ablation on Kvasir-Sessile [20], cross-domain and cross-resolution tasks, as can be seen in Tab. 5. We employ LoRA alone as a baseline for binary segmentation, obtaining a competitive performance, which is consistent with the baseline SAMed [46]. Following the incorporation of frequency adapters and INR decoders separately, the model's performance exhibits improvements. It can be observed that the INR decoder exhibits a more pronounced advantage in cross-domain and cross-resolution tasks. Furthermore, simultaneously employing frequency adapters (FA) and INR decoders can achieve a synergistic effect where 1+1>2.

Incorporating the frequency adapter. Tab. 6 indicates the effectiveness of the frequency adapter, and it can be observed that amplitude information is more helpful for spectrum representation compared to phase information. The results also demonstrate the segmentation boundary benefits from the frequency adapter.

Points Number for Sampling. Tab. 7 represents ablation on points number for Uncertainty Guided Sampling (UGS). This experiment reveals that I-MedSAM generates high-quality segmentation masks with the help of the proposed uncertainty guided sampling method. Excessive sampling points do not necessarily im-

Table 6: Ablation study on Frequency Adapter (FA). FA_{pha} stands for utilizing of the phase spectrum of DFT, while FA_{amp} stands for the amplitude spectrum of DFT.

Setting	w/o FA	FA_{pha}	FA_{amp}
Dice (%)	88.83	90.60	**91.49**
HD	15.44	12.67	**11.59**

prove the final segmentation results and may lead to increased memory consumption. Conversely, insufficient sampling points may limit the areas that require refinement. Therefore, a proportion of 12.5% for UGS is deemed appropriate for I-MedSAM. The parameter can be adjusted according to the specific tasks.

Table 7: Ablation study on the number of sampled feature points for Uncertainty Guided Sampling (UGS). "Top-K" denotes the selection of a specified proportion, K%, from all feature points.

Setting	w/o UGS	Top-50%	Top-25%	Top-12.5%	Top-6.25%	Top-3.125%
Dice (%)	87.77	90.27	89.59	**91.49**	91.01	90.48
HD Distance	16.15	13.88	14.12	**11.59**	12.99	14.53

4.5 Effect of different training annotations

To illustrate the generalization ability between I-MedSAM and baselines, we further conduct experiments on different proportions of training annotation amount. Following the experiment settings in SwIPE [47], based on the divided training set, we train I-MedSAM on 10%, 25%, 50% and 100% of the training set. As shown in Sec. 4.5, our I-MedSAM outperforms all baselines at various training annotation amounts. Thanks to the great generalization ability of SAM's encoder in I-MedSAM, I-MedSAM maintains higher segmentation performance even with relatively limited training annotations.

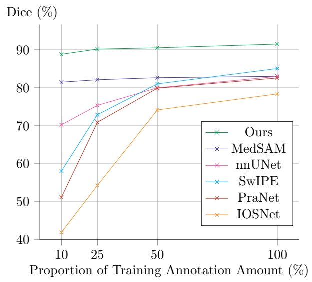


Fig. 6: Effect on different proportions of training annotation amount.

5 Conclusion

In this paper, we introduce I-MedSAM to enhance cross-domain ability and adaptability to diverse output resolutions in medical image segmentation. By integrating SAM's generalized representations into the INR space, I-MedSAM achieves state-of-the-art performance across various experimental scenarios. Specifically addressing the challenge of precise boundary delineation in 2D medical images, we incorporate a frequency adapter for parameter-efficient fine-tuning to SAM, showcasing the potential benefits of complementing spatial domain information with frequency domain insights for foundation models. Additionally, the employment of the uncertainty-guided sampling strategy in coarse-to-fine INRs proves effective in the selection and refinement of challenging samples in continuous space. These findings suggest avenues for future research to establish stronger connections between different representation spaces.

Acknowledgement

Shanghang Zhang is supported by the National Science and Technology Major Project of China (No. 2022ZD0117801).

References

1. Alahmadi, M.D.: Boundary aware u-net for medical image segmentation. Arabian Journal for Science and Engineering **48**(8), 9929–9940 (2023)
2. Amiranashvili, T., Lüdke, D., Li, H.B., Menze, B., Zachow, S.: Learning shape reconstruction from sparse measurements with neural implicit functions. In: International Conference on Medical Imaging with Deep Learning. pp. 22–34. PMLR (2022)
3. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics **43**, 99–111 (2015)
4. Bui, N.T., Hoang, D.H., Tran, M.T., Le, N.: Sam3d: Segment anything model in volumetric medical images. arXiv preprint arXiv:2309.03493 (2023)
5. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
6. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)
7. Cheng, Z., Wei, Q., Zhu, H., Wang, Y., Qu, L., Shao, W., Zhou, Y.: Unleashing the potential of sam for medical adaptation via hierarchical decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3511–3522 (2024)
8. Chibane, J., Alldieck, T., Pons-Moll, G.: Implicit functions in feature space for 3d shape reconstruction and completion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6970–6981 (2020)
9. Deng, G., Zou, K., Ren, K., Wang, M., Yuan, X., Ying, S., Fu, H.: Sam-u: Multi-box prompts triggered uncertainty estimation for reliable sam in medical image. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 368–377. Springer (2023)
10. Fan, D., Ji, G., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: PraNet: Parallel reverse attention network for polyp segmentation. In: MICCAI. pp. 263–273. Springer (2020)
11. Feng, Z., Wen, L., Wang, P., Yan, B., Wu, X., Zhou, J., Wang, Y.: Diffdp: Radiotherapy dose prediction via a diffusion model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 191–201. Springer (2023)
12. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. international conference on machine learning (2015)
13. Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2Net: A new multi-scale backbone architecture. IEEE TPAMI **43**(2), 652–662 (2019)
14. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. international conference on machine learning (2017)

15. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H., Xu, D.: UNETR: Transformers for 3D medical image segmentation. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 574–584 (2022)
16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
17. Hu, H., Chen, Y., Xu, J., Borse, S., Cai, H., Porikli, F., Wang, X.: Learning implicit feature alignment function for semantic segmentation. In: ECCV. pp. 487–505. Springer (2022)
18. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the hausdorff distance. IEEE Transactions on pattern analysis and machine intelligence **15**(9), 850–863 (1993)
19. Isensee, F., Jaeger, P., Kohl, S., Petersen, J., Maier-Hein, K.H.: nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
20. Jha, D., Smedsrud, P.H., Johansen, D., de Lange, T., Johansen, H.D., Halvorsen, P., Riegler, M.A.: A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation. IEEE Journal of Biomedical and Health Informatics **25**(6), 2029–2040 (2021)
21. Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. ArXiv:2206.08023 (2022)
22. Khan, M., Fang, Y.: Implicit neural representations for medical imaging segmentation. In: MICCAI (2022)
23. Khan, M.O., Fang, Y.: Implicit neural representations for medical imaging segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 433–443. Springer (2022)
24. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
25. Kuang, K., Zhang, L., Li, J., Li, H., Chen, J., Du, B., Yang, J.: What makes for automatic reconstruction of pulmonary segments. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 495–505. Springer (2022)
26. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. vol. 5, p. 12 (2015)
27. Li, Y., Shao, H.C., Liang, X., Chen, L., Li, R., Jiang, S., Wang, J., Zhang, Y.: Zero-shot medical image translation via frequency-guided diffusion models. arXiv preprint arXiv:2304.02742 (2023)
28. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2017)
29. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 1–9 (2024)
30. McGinnis, J., Shit, S., Li, H.B., Sideri-Lampretsa, V., Graf, R., Dannecker, M., Pan, J., Stolt-Ansó, N., Mühlau, M., Kirschke, J.S., et al.: Single-subject multi-contrast mri super-resolution via implicit neural representations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 173–183. Springer (2023)

31. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
32. Molaei, A., Aminimehr, A., Tavakoli, A., Kazerouni, A., Azad, B., Azad, R., Merhof, D.: Implicit neural representation in medical imaging: A comparative survey. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2381–2391 (2023)
33. Naval Marimont, S., Tarroni, G.: Implicit field learning for unsupervised anomaly detection in medical images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24. pp. 189–198. Springer (2021)
34. Pasupathy, A.: The neural basis of image segmentation in the primate brain. Neuroscience **296**, 101–109 (2015)
35. Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., Courville, A.: On the spectral bias of neural networks. In: International Conference on Machine Learning. pp. 5301–5310. PMLR (2019)
36. Reich, C., Prangemeier, T., Cetin, O., Koeppl, H.: OSS-Net: Memory efficient high resolution semantic segmentation of 3D medical data. In: British Machine Vision Conference (2021)
37. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
38. Shi, H., Han, S., Huang, S., Liao, Y., Li, G., Kong, X., Zhu, H., Wang, X., Liu, S.: Mask-enhanced segment anything model for tumor lesion semantic segmentation. arXiv preprint arXiv:2403.05912 (2024)
39. Sørensen, K., Camara, O., Backer, O., Kofoed, K., Paulsen, R.: NUDF: Neural unsigned distance fields for high resolution 3D medical image segmentation. ISBI pp. 1–5 (2022)
40. Stolt-Ansó, N., McGinnis, J., Pan, J., Hammernik, K., Rueckert, D.: Nisf: Neural implicit segmentation functions. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 734–744. Springer (2023)
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
42. Wei, X., Zhang, R., Wu, J., Liu, J., Lu, M., Guo, Y., Zhang, S.: Noc: High-quality neural object cloning with 3d lifting of segment anything. arXiv preprint arXiv:2309.12790 (2023)
43. Wu, J., Fu, R., Fang, H., Liu, Y., Wang, Z., Xu, Y., Jin, Y., Arbel, T.: Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
44. Yang, J., Wickramasinghe, U., Ni, B., Fua, P.: ImplicitAtlas: Learning deformable shape templates in medical imaging. In: CVPR. pp. 15861–15871 (2022)
45. You, C., Dai, W., Min, Y., Staib, L., Duncan, J.S.: Implicit anatomical rendering for medical image segmentation with stochastic experts. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 561–571. Springer (2023)
46. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023)
47. Zhang, Y., Gu, P., Sapkota, N., Chen, D.Z.: Swipe: Efficient and robust medical image segmentation with implicit patch embeddings. In: International Conference

on Medical Image Computing and Computer-Assisted Intervention. pp. 315–326. Springer (2023)

48. Zhang, Y., Sapkota, N., Gu, P., Peng, Y., Zheng, H., Chen, D.Z.: Keep your friends close & enemies farther: Debiasing contrastive learning with spatial priors in 3D radiology images. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1824–1829. IEEE (2022)