

# M<sup>2</sup>SNet: Multi-scale in Multi-scale Subtraction Network for Medical Image Segmentation

Xiaoqi Zhao<sup>1,3</sup>, Hongpeng Jia<sup>1</sup>, Youwei Pang<sup>1,4</sup>, Long Lv<sup>2</sup>, Feng Tian<sup>2</sup>, Lihe Zhang<sup>1</sup>, Weibing Sun<sup>2</sup> and Huchuan Lu<sup>1</sup>

<sup>1</sup>IIAU-Lab, Dalian University of Technology, Dalian, China.

<sup>2</sup>Zhongshan Hospital of Dalian University, Dalian, China.

<sup>3</sup>Yale University, USA.

<sup>4</sup>Nanyang Technological University, Singapore.

## Abstract

Accurate medical image segmentation is critical for early medical diagnosis. Most existing methods are based on U-shape structure and use element-wise addition or concatenation to fuse different level features progressively in decoder. However, both the two operations easily generate plenty of redundant information, which will weaken the complementarity between different level features, resulting in inaccurate localization and blurred edges of lesions. To address this challenge, we propose a general multi-scale in multi-scale subtraction network (M<sup>2</sup>SNet) to finish diverse segmentation from medical image. Specifically, we first design a basic subtraction unit (SU) to produce the difference features between adjacent levels in encoder. Next, we expand the single-scale SU to the intra-layer multi-scale SU, which can provide the decoder with both pixel-level and structure-level difference information. Then, we pyramidally equip the multi-scale SUs at different levels with varying receptive fields, thereby achieving the inter-layer multi-scale feature aggregation and obtaining rich multi-scale difference information. In addition, we build a training-free network “LossNet” to comprehensively supervise the task-aware features from bottom layer to top layer, which drives our multi-scale subtraction network to capture the detailed and structural cues simultaneously. Without bells and whistles, our method performs favorably against most state-of-the-art methods under different evaluation metrics on eleven datasets of four different medical image segmentation tasks of diverse image modalities, including color colonoscopy imaging, ultrasound imaging, computed tomography (CT), and optical coherence tomography (OCT). The source code can be available at <https://github.com/Xiaoqi-Zhao-DLUT/MSNet>.

**Keywords:** Medical Image Segmentation, Subtraction Unit, Multi-scale in Multi-scale, Difference Information, LossNet

## 1 Introduction

As the important role in computer-aided diagnosis system, accurate medical image segmentation technique can provide the doctors with great guidance for making clinical decisions. There are three general challenges in accurate

segmentation: **Firstly**, U-shape structures [1, 2] have received considerable attention due to their abilities of utilizing multi-level information to reconstruct high-resolution feature maps. In UNet [2], the up-sampled feature maps are concatenated with feature maps skipped from the encoder and convolutions and non-linearities

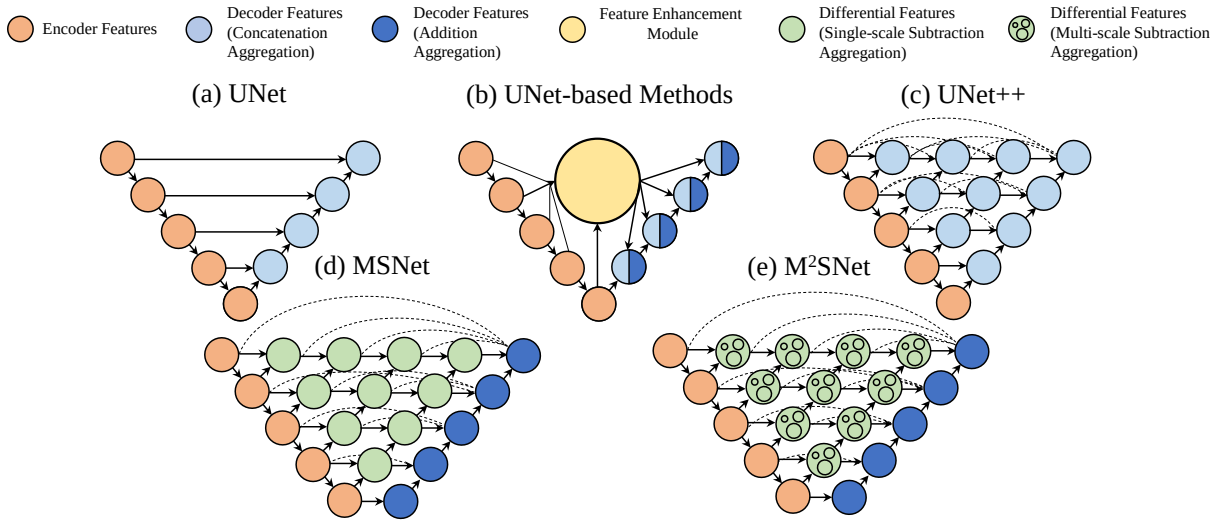


**Fig. 1** Illustration of different medical image segmentation architectures.

are added between up-sampling steps, as shown in Fig. 1 (a). Subsequent UNet-based methods design diverse feature enhancement modules via attention mechanism [3–5], gate mechanism [6–8], transformer technique [9–11], as shown in Fig. 1 (b). UNet++ [12] uses nested and dense skip connections to reduce the semantic gap between the feature maps of encoder and decoder, as shown in Fig. 1 (c). Generally speaking, different level features in encoder have different characteristics. High-level ones have more semantic information which helps localize the objects, while low-level ones have more detailed information which can capture the subtle boundaries of objects. The decoder leverages the level-specific and cross-level characteristics to generate the final high-resolution prediction. Nevertheless, the aforementioned methods directly use an element-wise addition or concatenation to fuse any two level features from the encoder and transmit them to the decoder. These simple operations do not pay more attention to differential information between different levels. This drawback not only generates redundant information to dilute the really useful features but also weakens the characteristics of level-specific features, which results in that the network can not balance accurate localization and subtle boundary refinement. **Secondly**, due to the limited receptive field, a single-scale convolutional kernel is difficult to capture context information of size-varying objects. Some methods [1, 2, 12–15] rely on the inter-layer multi-scale features and

progressively integrate the semantic context and texture details from diverse scale representations. Others [5, 8, 16] focus on extracting the intra-layer multi-scale information based on the atrous spatial pyramid pooling module [17] (ASPP) or DenseASPP [18] in their networks. However, the ASPP-like multi-scale convolution modules will produce many extra parameters and computations. Many methods [6, 19–22] usually equip several ASPP modules into the encoder/decoder blocks of different levels, while some ones [23–26] install it on the highest-level encoder block. **Thirdly**, the form of the loss function directly provides the direction for the gradient optimization of the network. In segmentation field, there are many loss functions are proposed to supervise the prediction at the different levels, such as the L1 loss, cross-entropy loss and weighted cross-entropy loss [27] in the pixel level, the SSIM [28] loss and uncertainty-aware loss [29] in the region level, the IoU loss, Dice loss and consistency-enhanced loss [14] in the global level. Although these basic loss functions and their variants have different optimization characteristics, the designs of complex manual math forms are really time-consuming for many researches. In order to obtain comprehensive performance, models usually integrate a variety of loss functions, which places great demands on the training skills of the researchers. Therefore, we think that it is necessary to introduce an intelligent loss function without complex

manual designs to comprehensively supervise the segmentation prediction.

In this paper, we propose a novel multi-scale in multi-scale subtraction network (M<sup>2</sup>SNet) for general medical image segmentation. Firstly, we design a subtraction unit (SU) and apply it to each pair of adjacent level features. The SU highlights the useful difference information between the features and eliminates the interference from the redundant parts. Secondly, we collect the extreme multi-scale information with the help of the proposed multi-scale in multi-scale subtraction module. For the inter-layer multi-scale information, we pyramidally concatenate multiple subtraction units to capture the large-span cross-level information. Then, we aggregate level-specific features and multi-path cross-level differential features and then generate the final prediction in decoder. For the intra-layer multi-scale information, we improve the single-scale subtraction unit to the multi-scale subtraction unit through a group of full one filters with different kernel sizes, which can achieve naturally multi-scale subtraction aggregation without introducing extra parameters. As shown in Fig. 1, MSNet equips the inter-layer multi-scale subtraction module and M<sup>2</sup>SNet has both the inter-layer and intra-layer multi-scale subtraction structures. Thirdly, we propose a LossNet to automatically supervise the extracted feature maps from bottom layer to top layer, which can optimize the segmentation from detail to structure with a simple L2-loss function.

Our main contributions are summarized as follows:

- We present a new segmentation framework by replacing traditional addition or concatenation feature fusion with an efficient subtraction aggregation.
- We propose a simple yet general multi-scale in multi-scale subtraction network (M<sup>2</sup>SNet) for diverse medical image segmentation. With multi-scale in multi-scale module, the multi-scale complementary information from lower order to higher order among different levels can be effectively obtained, thereby comprehensively enhancing the perception of organs or lesion areas.
- We design an efficient intra-layer multi-scale subtraction unit (MSU). Due to the low parameters and computation of MSU, it can be equipped for all cross-layer aggregations in our M<sup>2</sup>SNet.
- We build a general training-free loss network to implement the detail-to-structure supervision in the feature levels, which provides the important supplement to the loss design based on the prediction itself.
- We verify the effectiveness of the M<sup>2</sup>SNet on four challenge medical segmentation tasks: polyp segmentation, breast cancer segmentation, lung infection and OCT layer segmentation corresponding to the color colonoscopy imaging, ultrasound imaging, computed tomography (CT), and optical coherence tomography (OCT) image input modality, respectively. In addition, M<sup>2</sup>SNet won the second place in the MICCAI2022 GOALS International Ophthalmology Challenge.

## 2 Related Work

### 2.1 Medical Image Segmentation Network

According to the characteristics of different organs or lesions, we classify existing medical image segmentation methods into two types: medical-general and medical-specific one.

**Medicine-general Methods.** With the U-Net [2] achieving stable performance in the medical image segmentation field, the U-shape structure with encoder-decoder has become the basic segmentation baseline. U-Net++ [12] integrates both the long connection and short connection, which can reduce the semantic gap between the feature maps of the encoder and decoder sub-networks. For attention U-Net [30], an attention gate is embedded in each transition layer between the encoder and decoder block, which can automatically learn to focus on target structures of varying shapes and sizes. More recently, Transformer-based approaches [31–33] have gained prominence by exploiting self-attention to capture long-range dependencies beyond the limited receptive fields of CNNs. SegFormer [32] and SwinUNet [33] combine hierarchical feature extraction with attention-driven decoding, while hybrid designs such as UTNet [9] and TransUNet [11] embed Transformer modules into U-shaped frameworks to infuse global context into local representations. Despite their improved

expressiveness, these models often suffer from quadratic complexity and limited local inductive bias, which can hinder scalability and fine-grained predictions. In response, state-space-based architectures have emerged as an efficient alternative. Methods built on Mamba [34], such as Sigma [35], SegMamba [36] and U-Mamba [37], replace attention with structured recurrence, achieving linear time complexity and reduced memory footprint. By combining the inherent structural bias of state-space models with CNNs' local detail preservation, hybrid CNN–Mamba designs offer a promising route to scalable, high-resolution, and real-time medical image segmentation.

**Medicine-specific Methods.** In the polyp segmentation task, SFA [38] and PraNet [4], focus on recovering the sharp boundary between a polyp and its surrounding mucosa. The former proposes a selective feature aggregation structure and a boundary-sensitive loss function under a shared encoder and two mutually constrained decoders. The latter utilizes a reverse attention module to establish the relationship between the region and boundary cues. UM-Net [39] introduces color transfer to reduce color–polyp dependency and incorporating uncertainty estimation with variance correction to enhance the reliability of segmentation results. In addition, Ji *et al.* [40] utilize spatio-temporal information to build the video polyp segmentation model. In the lung infection task, Paluru *et al.* [41] propose an anamorphic depth embedding-based lightweight CNN to segment anomalies in chest CT images. Inf-Net [42] builds the implicit reverse attention and explicit edge attention to model the boundaries. BCS-Net [43] has three progressive boundary context-semantic reconstruction blocks, which can help the decoder to capture the piecemeal region for lung infection. In the breast segmentation task, Byra *et al.* [44] develop a selective kernel via an attention mechanism to adjust the receptive fields of the U-Net, which can further improve the segmentation accuracy of breast tumors. Chen *et al.* [45] propose a nested U-net to achieve robust representation of breast tumors by exploiting different depths and sharing weights.

We can see that the medicine-general methods are usually towards general challenges (i.e.,

rich feature representation, multi-scale information extraction and cross-level feature aggregation). And, the medicine-specific methods propose targeted solutions based on the characteristics of the current organ or lesion, such as designing a series of attention mechanisms, edge enhancement modules, uncertainty estimation, etc. However, both general medicine-general and medicine-specific models rely on a large number of addition or concatenation operations to achieve feature fusion, which weakens the specificity parts among complementary features. Our proposed multi-scale subtraction module naturally focuses on extracting difference information, thus providing the decoder with efficient targeted features.

## 2.2 Multi-scale Feature Extraction

Scale cues play an important role in capturing contextual information of objects. Inspired by the scale-space theory that has been widely validated as an effective and theoretically sound framework, more and more multi-scale methods are proposed. Compared with single-scale features, multi-scale features are beneficial to address naturally occurring scale variations. This characteristic can help the medical segmentation models perceive lesions with different scales. According to the form, current multi-scale based methods can be roughly divided into two categories, namely, the inter-layer multi-scale structure and the intra-layer multi-scale structure. Inter-layer multi-scale structure is based on features with different scales extracted by the feature encoder and progressively aggregates them in decoder, such as the U-shape [1, 2, 4, 12–14, 46, 47] architecture. Among them, dense skip connections are widely used in decoder because of the advantages in gradient backpropagation and information aggregation. U-Net++ [12] undergoes a dense convolution block and re-designs skip pathways to transform the connectivity of the encoder and decoder sub-networks. ICUNet++ [48] replaces the convolutional layer of U-Net++ with inception structure and add attention gate module before each skip connection to filter interference information. DSSNet [49] and CPFPP [50] use a fluid pyramid integration strategy to make better use of multi-scale cross-modal/level features. MINet [14]

proposes the self-interaction and aggregate interaction strategy to avoid the interference in feature fusion caused by large resolution differences. Intra-layer multi-scale structure usually equips the multi-scale pluggable modules, such as ASPP [17], DenseASPP [18], FoldASPP [8], and PAFEM [51] to construct the parallel multi-branch convolution layers with different dilated rates to obtain a rich combination of receptive fields. In this work, we apply our subtraction unit to both inter-layer and intra-layer multi-scale feature fusion, which can fully show the flexibility and effectiveness of the subtraction unit, and its general gains will be more convincing. It should be emphasized that we naturally aggregate multi-scale subtractive features in the decoder without focusing on skip connection designs.

### 2.3 Loss Method

Most loss functions in image segmentation are based on cross-entropy or coincidence measures. The traditional cross-entropy loss treats the categories information equally. Long *et al.* [27] propose a weighted cross-entropy loss (WCE) for each class to offset the class imbalance in the data. Lin *et al.* [52] introduce the weights of difficult and easy samples to propose the Focal loss. Dice loss [53] is proposed as the loss function of coincidence measurement in V-Net, which can effectively suppress problems caused by category imbalance. Tversky loss [54] is a regularized version of Dice loss to control the contribution of accuracy and recall to the loss function. Wong *et al.* [55] propose exponential logarithmic loss (EL Loss) through the weighted summation of Dice loss and WCE loss to improve the segmentation accuracy of small structure objects. Taghanaki *et al.* [56] find that there is a risk in using the loss function based on overlap alone, and propose the como-loss to combine Dice loss as a regularization term with WCE loss to deal with the problem of input and output imbalance. Although these various loss functions have different effects at different levels, it is indeed time-consuming and laborious to manually design these complex functions. To this end, we propose the automatic and comprehensive segmentation loss structure, coined as the LossNet.

## 3 Method

The M<sup>2</sup>SNet architecture is shown in Fig. 2, in which there are five encoder blocks ( $\mathbf{E}^i$ ,  $i \in \{1, 2, 3, 4, 5\}$ ), a multi-scale in multi-scale subtraction module (MMSM) and four decoder blocks ( $\mathbf{D}^i$ ,  $i \in \{1, 2, 3, 4\}$ ). We adopt the Res2Net-50 as the backbone to extract five levels of features. First, we separately adopt a  $3 \times 3$  convolution for feature maps of each encoder block to reduce the channel to 64, which can decrease the number of parameters for subsequent operations. Next, these different level features are fed into the MMSM and output five complementarity enhanced features ( $CE^i$ ,  $i \in \{1, 2, 3, 4, 5\}$ ). Finally, each  $CE^i$  progressively participates in the decoder and generates the final prediction. In the training phase, both the prediction and ground truth are input into the LossNet to achieve supervision. We describe the multi-scale in multi-scale subtraction module in Sec. 3.1 and give the details of LossNet in Sec. 3.2.

### 3.1 Multi-scale in Multi-scale Subtraction Module

We use  $F_A$  and  $F_B$  to represent adjacent level feature maps. They all have been activated by the ReLU operation. We define a basic subtraction unit (SU):

$$SU = Conv(|F_A \ominus F_B|), \quad (1)$$

where  $\ominus$  is the element-wise subtraction operation,  $|\cdot|$  calculates the absolute value and  $Conv(\cdot)$  denotes the convolution layer. Directly performing single-scale subtraction on the features of element positions is only to establish the difference relationship on the isolated pixel level, without considering that the lesion may have the characteristics of regional clustering. Compared to the MICCAI version [57] of MSNet with the single-scale subtraction unit, we design a powerful intra-layer multi-scale subtraction unit (MSU) and improve MSNet to M<sup>2</sup>SNet. As shown in Fig. 3, we utilize the multi-scale convolution filters with fixed full one weights of size  $1 \times 1$ ,  $3 \times 3$  and  $5 \times 5$  to calculate the detail and structure difference values according to the pixel-pixel and region-region pattern. Using multi-scale filters with fixed parameters not only can directly

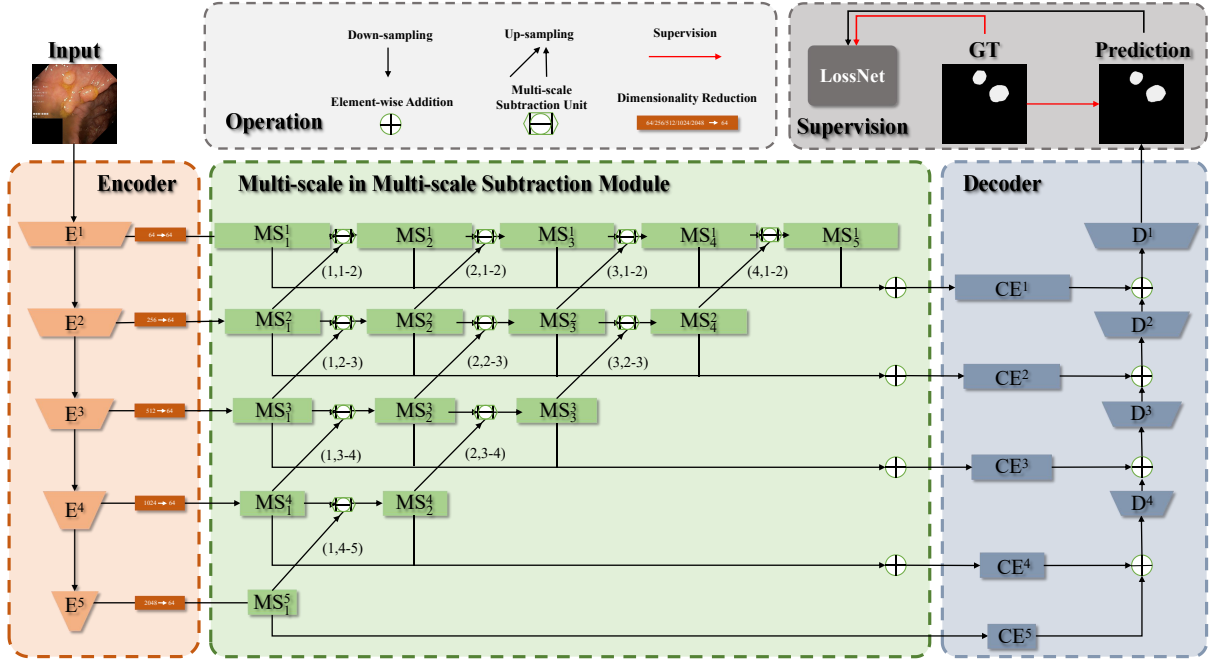


**Fig. 2** Overview of the proposed multi-scale subtraction network.

capture the multi-scale difference clues between initial feature pairs at matched spatial locations, but also achieve efficient training without introducing additional parameter burdens. Therefore, M<sup>2</sup>SNet can maintain the same low computation as MSNet and achieve higher precision performance. The entire multi-scale subtraction process can be formulated as:

$$MSU = Conv( \begin{aligned} &|Filter(F_A)_{1 \times 1} \ominus Filter(F_B)_{1 \times 1}| + \\ &|Filter(F_A)_{3 \times 3} \ominus Filter(F_B)_{3 \times 3}| + \\ &|Filter(F_A)_{5 \times 5} \ominus Filter(F_B)_{5 \times 5}|, \end{aligned} \quad (2)$$

where  $Filter(\cdot)_{n \times n}$  represents the full one filter of size  $n \times n$ . The MSU can capture the complementary information of  $F_A$  and  $F_B$  and highlight their differences from texture to structure, thereby providing richer information for the decoder.

To obtain higher-order complementary information across multiple feature levels, we horizontally and vertically concatenate multiple MSUs to calculate a series of differential features with different orders and receptive fields. The detail of the multi-scale in multi-scale subtraction module can be found in Fig. 2. We aggregate the scale-specific feature ( $MS_1^i$ ) and cross-scale differential

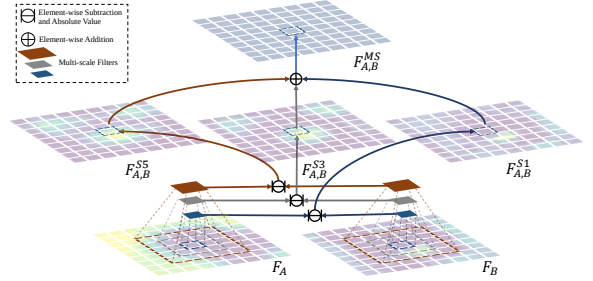


**Fig. 3** Detailed diagram of multi-scale subtraction unit.

features ( $MS_{n \neq 1}^i$ ) between the corresponding level and any other levels to generate complementarity enhanced feature ( $CE^i$ ). This process can be formulated as follows:

$$CE^i = Conv(\sum_{n=1}^{6-i} MS_n^i) \quad i = 1, 2, 3, 4, 5. \quad (3)$$

Finally, all  $CE^i$  participate in decoding and then the polyp region is segmented.

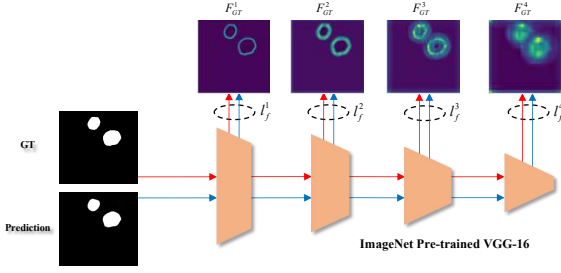


Fig. 4 Illustration of LossNet.

### 3.2 LossNet

In the proposed model, the total training loss can be written as:

$$\mathcal{L}_{total} = \mathcal{L}_{IoU}^w + \mathcal{L}_{BCE}^w + \mathcal{L}_f, \quad (4)$$

where  $\mathcal{L}_{IoU}^w$  and  $\mathcal{L}_{BCE}^w$  represent the weighted IoU loss and binary cross-entropy (BCE) loss which have been widely adopted in segmentation tasks. We use the same definitions as in [4, 58, 59] and their effectiveness has been validated in these works. Different from them, we extra use a LossNet to further optimize the segmentation from detail to structure. Specifically, we use an ImageNet pre-trained classification network, such as VGG-16, to extract the multi-scale features of the prediction and ground truth, respectively. Then, their feature difference is computed as loss  $\mathcal{L}_f$ :

$$\mathcal{L}_f = l_f^1 + l_f^2 + l_f^3 + l_f^4. \quad (5)$$

Let  $F_P^i$  and  $F_G^i$  separately represent the  $i$ -th level feature maps extracted from the prediction and ground truth. The  $l_f^i$  is calculated as their Euclidean distance (L2-Loss), which is supervised at the pixel level:

$$l_f^i = \|F_P^i - F_G^i\|_2, \quad i = 1, 2, 3, 4. \quad (6)$$

As can be seen from Fig 4, the low-level feature maps contain rich boundary information and the high-level ones depict location information. Thus, the LossNet can generate comprehensive supervision at the feature levels.

## 4 Experiments

### 4.1 Datasets

Extensive experiments are conducted to verify the effectiveness of the proposed framework on four different types of medical segmentation tasks with data from varied image modalities, including color colonoscopy imaging, ultrasound imaging, computed tomography (CT), and optical coherence tomography (OCT).

**Polyp Segmentation.** According to GLOBOCAN 2020 data, colorectal cancer is the third most common cancer worldwide and the second most common cause of death. It usually begins as small, noncancerous (benign) clumps of cells called polyps that form on the inside of the colon. We evaluate the proposed model on five benchmark datasets: CVC-ColonDB [60], ETIS [61], Kvasir [62], CVC-T [63] and CVC-ClinicDB [64]. We adopt the same training set as the latest image polyp segmentation method [4], that is, 900 samples from the Kvasir and 550 samples from the CVC-ClinicDB [64] are used for training. The remaining images and the other three datasets are used for testing. Besides, there are some video-based polyp datasets, including the CVC-300[65] and CVC-612[64]. We follow the latest video polyp segmentation method to split the videos from CVC-300 (12 clips) and CVC-612 (29 clips) into 60% for training, 20% for validation, and 20% for testing.

**Lung Infection.** A novel viral pneumonia that emerged in early 2020 rapidly spread worldwide, creating an unprecedented public-health challenge. At present, only a few public lung CT datasets are available for infection-area segmentation. To obtain a relatively sufficient sample size for training, we slice one publicly available dataset [66] and merge it with another public dataset [67], resulting in 1,277 high-quality CT images through uniform sampling. These are then split into 894 images for training and 383 images for testing.

**Breast Ultrasound Segmentation.** Breast cancer is one of the most dreaded cancers in women [68]. Segmenting the lesion region from breast ultrasound images is essential for tumor diagnosis. BUSI [69] dataset contains 780 images of 600 female patients. Among them, there are 133

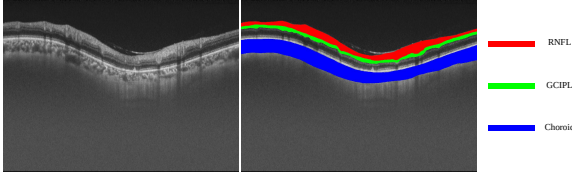


**Fig. 5** Visualization of the fundus OCT layer segmentation.

normal cases, 437 benign tumors, and 210 malignant tumors. We follow the popular breast ultrasound segmentation methods [44, 45] to perform four-fold cross-validation on BUSI.

**OCT Layer Segmentation.** The OCT images are often used to diagnose and monitor retinal diseases more accurately based on abnormality quantification and retinal layer thickness computation both in research centers and clinic routines. At present, many scholars have been studying the segmentation of fundus structure in macular OCT scans, but few focus on parapapillary circular scans. To fully show the generalization of M<sup>2</sup>SNet in different medical tasks, we take our M<sup>2</sup>SNet to participate in the **MICCAI 2022 Challenge: Glaucoma Oct Analysis and Layer Segmentation (GOALS)**<sup>1</sup>. It requests participants to segment three layers, which has positive significance for the diagnosis of glaucoma, including retinal nerve fiber layer (RNFL), ganglion cell-inner plexiform layer (GCIPL), and choroid layer, as shown in Fig. 5. The GOALS2022 [70] dataset [71] contains 300 circumpapillary OCT. There are three equal groups with 100 OCT images for the training process, the preliminary competition process and the final process, respectively. The GOALS2022 challenge attracts 100 teams from all over the world to participate, and we finally won **the second place (2/100)**<sup>2</sup>.

## 4.2 Evaluation Metrics

There are many popular metrics used in different medical segmentation branches. mean Dice (mDice), mean IoU (mIoU), the weighted F-measure ( $F_{\beta}^w$ ) [72], S-measure ( $S_{\alpha}$ ) [73], E-measure ( $E_{\phi}^{max}$ ) [74] and mean absolute error (MAE) are widely used in polyp segmentation.

**Table 1** Quantitative comparisons on image polyp segmentation datasets. Top 2 scores are highlighted in **red** and **blue**, respectively. “†” represents the medicine-specific method.

|          | Methods             | Backbone       | mDice ↑      | mIoU ↑       | $F_{\beta}^w$ ↑ | $S_{\alpha}$ ↑ | $E_{\phi}^{max}$ ↓ | MAE ↓        |
|----------|---------------------|----------------|--------------|--------------|-----------------|----------------|--------------------|--------------|
| ColonDB  | U-Net [2]           | R2-50          | 0.519        | 0.449        | 0.498           | 0.711          | 0.763              | 0.061        |
|          | U-Net++ [12]        | R2-50          | 0.490        | 0.413        | 0.467           | 0.691          | 0.762              | 0.064        |
|          | Atten-U-Net [30]    | R2-50          | 0.466        | 0.385        | 0.431           | 0.670          | 0.724              | 0.071        |
|          | UTNet [9]           | R-50 + ViT-B16 | 0.676        | 0.600        | 0.656           | 0.799          | 0.855              | 0.041        |
|          | TransUNet [77]      | R-50 + ViT-B16 | 0.717        | 0.645        | 0.685           | 0.824          | 0.841              | 0.044        |
|          | SFA† [38]           | R2-50          | 0.467        | 0.351        | 0.379           | 0.634          | 0.648              | 0.094        |
|          | PraNet† [4]         | R2-50          | 0.716        | 0.645        | <b>0.699</b>    | 0.820          | 0.847              | 0.043        |
|          | MSNet [57]          | R2-50          | <b>0.755</b> | <b>0.678</b> | <b>0.737</b>    | <b>0.836</b>   | <b>0.883</b>       | <b>0.041</b> |
|          | M <sup>2</sup> SNet | R2-50          | <b>0.758</b> | <b>0.685</b> | <b>0.737</b>    | <b>0.842</b>   | <b>0.869</b>       | <b>0.038</b> |
|          | U-Net [2]           | R2-50          | 0.406        | 0.343        | 0.366           | 0.682          | 0.645              | 0.036        |
| ETIS     | U-Net++ [12]        | R2-50          | 0.413        | 0.342        | 0.390           | 0.681          | 0.704              | 0.035        |
|          | Atten-U-Net [30]    | R2-50          | 0.382        | 0.308        | 0.372           | 0.641          | 0.670              | 0.050        |
|          | UTNet [9]           | R-50 + ViT-B16 | 0.556        | 0.489        | 0.522           | 0.749          | 0.772              | 0.022        |
|          | TransUNet [77]      | R-50 + ViT-B16 | 0.573        | 0.512        | 0.517           | 0.765          | 0.707              | 0.029        |
|          | SFA† [38]           | R2-50          | 0.297        | 0.219        | 0.231           | 0.557          | 0.515              | 0.109        |
|          | PraNet† [4]         | R2-50          | 0.630        | 0.576        | 0.600           | 0.791          | 0.792              | 0.031        |
|          | MSNet [57]          | R2-50          | <b>0.719</b> | <b>0.664</b> | <b>0.678</b>    | <b>0.840</b>   | <b>0.830</b>       | <b>0.020</b> |
|          | M <sup>2</sup> SNet | R2-50          | <b>0.749</b> | <b>0.678</b> | <b>0.712</b>    | <b>0.846</b>   | <b>0.872</b>       | <b>0.017</b> |
|          | U-Net [2]           | R2-50          | 0.821        | 0.756        | 0.794           | 0.858          | 0.901              | 0.055        |
|          | U-Net++ [12]        | R2-50          | 0.824        | 0.753        | 0.808           | 0.862          | 0.907              | 0.048        |
| Kvasir   | Atten-U-Net [30]    | R2-50          | 0.769        | 0.683        | 0.730           | 0.828          | 0.859              | 0.062        |
|          | UTNet [9]           | R-50 + ViT-B16 | 0.862        | 0.803        | 0.843           | 0.886          | 0.911              | 0.042        |
|          | TransUNet [77]      | R-50 + ViT-B16 | 0.869        | 0.816        | 0.847           | 0.899          | 0.920              | 0.040        |
|          | SFA† [38]           | R2-50          | 0.725        | 0.619        | 0.670           | 0.782          | 0.828              | 0.075        |
|          | PraNet† [4]         | R2-50          | 0.901        | 0.848        | 0.885           | <b>0.915</b>   | 0.943              | 0.030        |
|          | MSNet [57]          | R2-50          | <b>0.907</b> | <b>0.862</b> | <b>0.893</b>    | <b>0.922</b>   | <b>0.944</b>       | <b>0.028</b> |
|          | M <sup>2</sup> SNet | R2-50          | <b>0.912</b> | <b>0.861</b> | <b>0.901</b>    | <b>0.922</b>   | <b>0.953</b>       | <b>0.025</b> |
|          | U-Net [2]           | R2-50          | 0.717        | 0.639        | 0.684           | 0.842          | 0.867              | 0.022        |
|          | U-Net++ [12]        | R2-50          | 0.714        | 0.636        | 0.687           | 0.838          | 0.884              | 0.018        |
|          | Atten-U-Net [30]    | R2-50          | 0.603        | 0.511        | 0.554           | 0.760          | 0.819              | 0.024        |
| CVC4-T   | UTNet [9]           | R-50 + ViT-B16 | 0.806        | 0.733        | 0.778           | 0.882          | 0.924              | 0.016        |
|          | TransUNet [77]      | R-50 + ViT-B16 | 0.828        | 0.757        | 0.785           | 0.906          | 0.901              | 0.015        |
|          | SFA† [38]           | R2-50          | 0.465        | 0.332        | 0.341           | 0.640          | 0.604              | 0.065        |
|          | PraNet† [4]         | R2-50          | <b>0.873</b> | 0.804        | 0.843           | 0.924          | 0.938              | <b>0.010</b> |
|          | MSNet [57]          | R2-50          | 0.869        | <b>0.807</b> | <b>0.849</b>    | <b>0.925</b>   | <b>0.943</b>       | <b>0.010</b> |
|          | M <sup>2</sup> SNet | R2-50          | <b>0.903</b> | <b>0.842</b> | <b>0.881</b>    | <b>0.939</b>   | <b>0.965</b>       | <b>0.009</b> |
|          | U-Net [2]           | R2-50          | 0.824        | 0.767        | 0.811           | 0.889          | 0.917              | 0.019        |
|          | U-Net++ [12]        | R2-50          | 0.797        | 0.741        | 0.785           | 0.872          | 0.908              | 0.022        |
|          | Atten-U-Net [30]    | R2-50          | 0.866        | 0.809        | 0.856           | 0.908          | 0.960              | 0.015        |
|          | UTNet [9]           | R-50 + ViT-B16 | 0.860        | 0.818        | 0.856           | 0.910          | 0.963              | 0.017        |
| ClinicDB | TransUNet [77]      | R-50 + ViT-B16 | 0.847        | 0.798        | 0.831           | 0.907          | 0.920              | 0.020        |
|          | SFA† [38]           | R2-50          | 0.698        | 0.615        | 0.647           | 0.793          | 0.816              | 0.042        |
|          | PraNet† [4]         | R2-50          | 0.902        | 0.858        | 0.896           | 0.935          | 0.958              | <b>0.009</b> |
|          | MSNet [57]          | R2-50          | <b>0.921</b> | <b>0.879</b> | <b>0.914</b>    | <b>0.941</b>   | <b>0.972</b>       | <b>0.008</b> |

Following [42], five metrics are employed for quantitative evaluation, including Precision, Recall, Dice Similarity Coefficient (DSC) [75], S-measure and MAE. Jaccard, Precision, Recall, Dice and Specificity [45, 76] are more commonly used for breast tumor segmentation. For OCT layer segmentation, GOALS2022 [71] adopts the Dice coefficient and mean Euclidean distance (MED) to evaluate segmentation bodies and edges, respectively. The lower value is better for the MAE and MED, and higher is better for others.

## 4.3 Implementation Details

Our model is implemented based on the PyTorch framework and trained on a single 2080Ti GPU with mini-batch size 16. We resize the inputs to  $352 \times 352$  and employ a general multi-scale training strategy as most methods [4, 26, 57, 58, 78, 79]. Random horizontally flipping and random rotate data augmentation are used to avoid overfitting. For the optimizer, we adopt the stochastic gradient descent (SGD). The momentum and weight decay are set as 0.9 and 0.0005, respectively. Maximum learning rate is set to 0.005 for backbone

<sup>1</sup><https://conferences.miccai.org/2022/en/MICCAI2022-CHALLENGES.html>

<sup>2</sup><https://github.com/Xiaoqi-Zhao-DLUT/MSNet>

**Table 2** Quantitative comparisons on video polyp segmentation datasets. Top 2 scores are highlighted in **red** and **blue**, respectively. “†” and “★” represent the medicine-specific method and the video polyp method, respectively.

|            | Methods             | Backbone | mDice ↑      | mIoU ↑       | $S_a$ ↑      | $E_{\phi}^{max}$ ↑ | MAE ↓        |
|------------|---------------------|----------|--------------|--------------|--------------|--------------------|--------------|
| CVC-300-TV | U-Net [2]           | R2-50    | 0.639        | 0.525        | 0.793        | 0.826              | 0.027        |
|            | U-Net++ [12]        | R2-50    | 0.649        | 0.539        | 0.796        | 0.831              | 0.024        |
|            | ResUNet++† [3]      | R2-50    | 0.535        | 0.412        | 0.703        | 0.718              | 0.052        |
|            | ACSNet† [46]        | R2-50    | 0.738        | 0.632        | 0.837        | 0.871              | 0.016        |
|            | PraNet† [4]         | R2-50    | 0.739        | 0.645        | 0.833        | 0.852              | 0.016        |
|            | PNS-Net†★ [40]      | R2-50    | <b>0.863</b> | <b>0.805</b> | <b>0.909</b> | <b>0.921</b>       | <b>0.013</b> |
|            | MSNet [57]          | R2-50    | 0.837        | 0.755        | 0.895        | 0.942              | 0.014        |
| CVC-612-V  | M <sup>2</sup> SNet | R2-50    | <b>0.876</b> | <b>0.805</b> | <b>0.918</b> | <b>0.963</b>       | <b>0.010</b> |
|            | U-Net [2]           | R2-50    | 0.725        | 0.610        | 0.826        | 0.855              | 0.023        |
|            | U-Net++ [12]        | R2-50    | 0.684        | 0.570        | 0.805        | 0.830              | 0.025        |
|            | ResUNet++† [3]      | R2-50    | 0.752        | 0.648        | 0.829        | 0.877              | 0.023        |
|            | ACSNet† [46]        | R2-50    | 0.804        | 0.712        | 0.847        | 0.887              | 0.054        |
|            | PraNet† [4]         | R2-50    | 0.869        | 0.799        | 0.915        | 0.936              | 0.013        |
|            | PNS-Net†★ [40]      | R2-50    | 0.859        | 0.804        | 0.923        | 0.944              | 0.012        |
| CVC-612-T  | MSNet [57]          | R2-50    | <b>0.889</b> | <b>0.834</b> | <b>0.931</b> | <b>0.959</b>       | <b>0.009</b> |
|            | M <sup>2</sup> SNet | R2-50    | <b>0.897</b> | <b>0.838</b> | <b>0.936</b> | <b>0.966</b>       | <b>0.010</b> |
|            | U-Net [2]           | R2-50    | 0.729        | 0.635        | 0.810        | 0.836              | 0.058        |
|            | U-Net++ [12]        | R2-50    | 0.740        | 0.635        | 0.800        | 0.817              | 0.059        |
|            | ResUNet++† [3]      | R2-50    | 0.617        | 0.514        | 0.727        | 0.758              | 0.084        |
|            | ACSNet† [46]        | R2-50    | 0.782        | 0.700        | 0.838        | 0.864              | 0.053        |
|            | PraNet† [4]         | R2-50    | <b>0.852</b> | <b>0.786</b> | 0.886        | <b>0.904</b>       | <b>0.038</b> |
|            | PNS-Net†★ [40]      | R2-50    | 0.841        | <b>0.788</b> | <b>0.903</b> | 0.903              | <b>0.038</b> |
|            | MSNet [57]          | R2-50    | 0.824        | 0.761        | 0.879        | 0.904              | 0.040        |
|            | M <sup>2</sup> SNet | R2-50    | <b>0.846</b> | 0.782        | <b>0.894</b> | <b>0.921</b>       | <b>0.037</b> |

**Table 3** Quantitative comparisons on the lung infection CT dataset. Top 2 scores are highlighted in **red** and **blue**, respectively. “†” represents the medicine-specific method.

| Methods             | Backbone       | DSC ↑        | Precision ↑  | Recall ↑     | $S_a$ ↑      | MAE ↓        |
|---------------------|----------------|--------------|--------------|--------------|--------------|--------------|
| U-Net [2]           | R2-50          | 0.736        | 0.782        | 0.793        | 0.834        | <b>0.007</b> |
| U-Net++ [12]        | R2-50          | 0.592        | 0.637        | 0.748        | 0.806        | 0.010        |
| Atten-UNet [30]     | R2-50          | 0.650        | 0.755        | 0.715        | 0.801        | 0.010        |
| UTNet [9]           | R-50 + ViT-B16 | 0.735        | 0.782        | 0.786        | 0.836        | <b>0.007</b> |
| TransUNet [77]      | R-50 + ViT-B16 | 0.710        | 0.770        | 0.776        | 0.831        | <b>0.007</b> |
| Inf-Net† [42]       | R2-50          | <b>0.783</b> | 0.774        | <b>0.852</b> | <b>0.853</b> | 0.007        |
| BCS-Net† [43]       | R2-50          | 0.763        | <b>0.775</b> | 0.763        | 0.840        | <b>0.007</b> |
| MSNet [57]          | R2-50          | 0.779        | <b>0.815</b> | 0.802        | 0.846        | <b>0.007</b> |
| M <sup>2</sup> SNet | R2-50          | <b>0.795</b> | <b>0.825</b> | <b>0.813</b> | <b>0.855</b> | <b>0.006</b> |

**Table 4** Quantitative comparisons on the breast ultrasound dataset. Top 2 scores are highlighted in **red** and **blue**, respectively. “†” represents the medicine-specific method.

| Methods             | Backbone       | Jaccard ↑         | Precision ↑       | Recall ↑          | Specificity ↑     | Dice ↑            |
|---------------------|----------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| U-Net [2]           | R2-50          | 51.54±1.17        | 56.84±1.35        | 71.18±2.87        | 95.20±0.33        | 57.61±1.26        |
| U-Net++ [12]        | R2-50          | 50.17±1.71        | 57.34±2.29        | 70.81±1.43        | 95.85±0.36        | 58.54±1.84        |
| Atten-UNet [30]     | R2-50          | 46.91±4.40        | 55.84±3.81        | 70.77±2.41        | 95.56±0.62        | 57.39±2.78        |
| UTNet [9]           | R-50 + ViT-B16 | 67.46±1.78        | 79.88±1.22        | 74.82±1.95        | <b>98.37±0.36</b> | 74.41±1.39        |
| TransUNet [77]      | R-50 + ViT-B16 | <b>71.17±0.93</b> | 81.66±1.52        | <b>80.75±1.35</b> | 98.05±0.25        | <b>78.00±0.49</b> |
| SKU-Net† [44]       | SKs            | 64.48±2.37        | 75.37±3.22        | 78.56±3.27        | 97.21±1.02        | 74.03±2.21        |
| NU-Net† [45]        | Deeper U-Net   | 68.86±1.99        | 78.90±2.26        | <b>82.48±2.14</b> | 97.79±0.87        | 77.79±1.88        |
| MSNet [57]          | R2-50          | 70.57±3.38        | <b>82.73±2.72</b> | 78.50±1.56        | 98.29±0.44        | 78.18±3.75        |
| M <sup>2</sup> SNet | R2-50          | <b>71.52±1.19</b> | <b>84.00±0.55</b> | 79.97±1.65        | <b>98.35±0.20</b> | <b>79.21±1.28</b> |

**Table 5** The leaderboard of MICCAI2022 Challenge: Glaucoma Oct Analysis and Layer Segmentation (GOALS). Top 2 scores are highlighted in **red** and **blue**, respectively.

| Teams                         | RNFL          |               | GCPL          |               | Choroid       |               |
|-------------------------------|---------------|---------------|---------------|---------------|---------------|---------------|
|                               | Dice ↑        | MED ↓         | Dice ↑        | MED ↓         | Dice ↑        | MED ↓         |
| LaTITM                        | 0.9547        | 1.1554        | 0.8950        | 1.3804        | 0.9557        | 1.7120        |
| OPTIMA-MUW                    | 0.9559        | 1.1084        | 0.8955        | 1.2838        | 0.9550        | 1.7599        |
| WRMT                          | 0.9560        | 1.1020        | <b>0.8980</b> | 1.2719        | 0.9567        | 1.6769        |
| Miracle-boyi                  | 0.9562        | 1.1065        | 0.8904        | 1.3346        | 0.9542        | 1.7904        |
| SZUMed                        | 0.9555        | 1.1289        | 0.8921        | 1.3565        | 0.9553        | 1.7628        |
| MedicalExplorer               | 0.9565        | 1.0888        | 0.8941        | 1.2972        | 0.9563        | 1.7452        |
| AUTOMATE                      | 0.9561        | 1.1014        | 0.8966        | 1.2848        | 0.9569        | 1.6767        |
| SJMED                         | 0.9565        | 1.0899        | 0.8970        | 1.2801        | <b>0.9578</b> | <b>1.6456</b> |
| <b>IIAU-Segmentors (Ours)</b> | <b>0.9576</b> | <b>1.0827</b> | 0.8953        | <b>1.2322</b> | <b>0.9578</b> | 1.6468        |
| Vision Wise                   | <b>0.9569</b> | <b>1.0833</b> | <b>0.8992</b> | <b>1.2566</b> | <b>0.9576</b> | <b>1.6295</b> |

**Table 6** The FLOPs, parameters and speed of different methods. The best and worst results are shown in **red** and **blue**, respectively.

| Metrics       | U-Net       | U-Net++      | Atten-UNet | UTNet        | TransUNet    | PraNet | Inf-Net | BCSNet | M <sup>2</sup> SNet |
|---------------|-------------|--------------|------------|--------------|--------------|--------|---------|--------|---------------------|
| FLOPs (GB) ↓  | ~12.2       | <b>~39.3</b> | ~9.7       | ~27.1        | ~24.7        | 12.0   | 12.4    | 14.8   | <b>~9.0</b>         |
| Params (MB) ↓ | <b>~7.2</b> | ~19.2        | ~12.6      | <b>~34.0</b> | <b>~93.2</b> | 30.5   | 31.1    | 44.8   | ~27.7               |
| Speed (FPS) ↑ | <b>~111</b> | ~82          | ~106       | <b>~44</b>   | ~59          | ~74    | ~68     | ~49    | ~90                 |

and 0.05 for other parts. Warm-up and linear decay strategies are used to adjust the learning rate. For any medical image sub-tasks, the above training strategy is used for all the multi-scale subtraction models involved in this paper. The difference among these models is only in the number of training epochs due to different convergence speeds. Specifically, the number of training epochs settings in the polyp segmentation, lung infection, breast tumor segmentation and OCT layer segmentation are 50, 200, 100 and 100, respectively.

#### 4.4 Comparisons with Medicine-general and Medicine-specific Methods

For a fair comparison, we compare not only with medicine-specific methods but also with representative medicine-general methods, including UNet [2], UNet++ [12], Attention U-Net [30], UTNet [9] and TransUNet [77]. Based on the open-source codes, we retrain these medicine-general methods on the same training sets as our models.

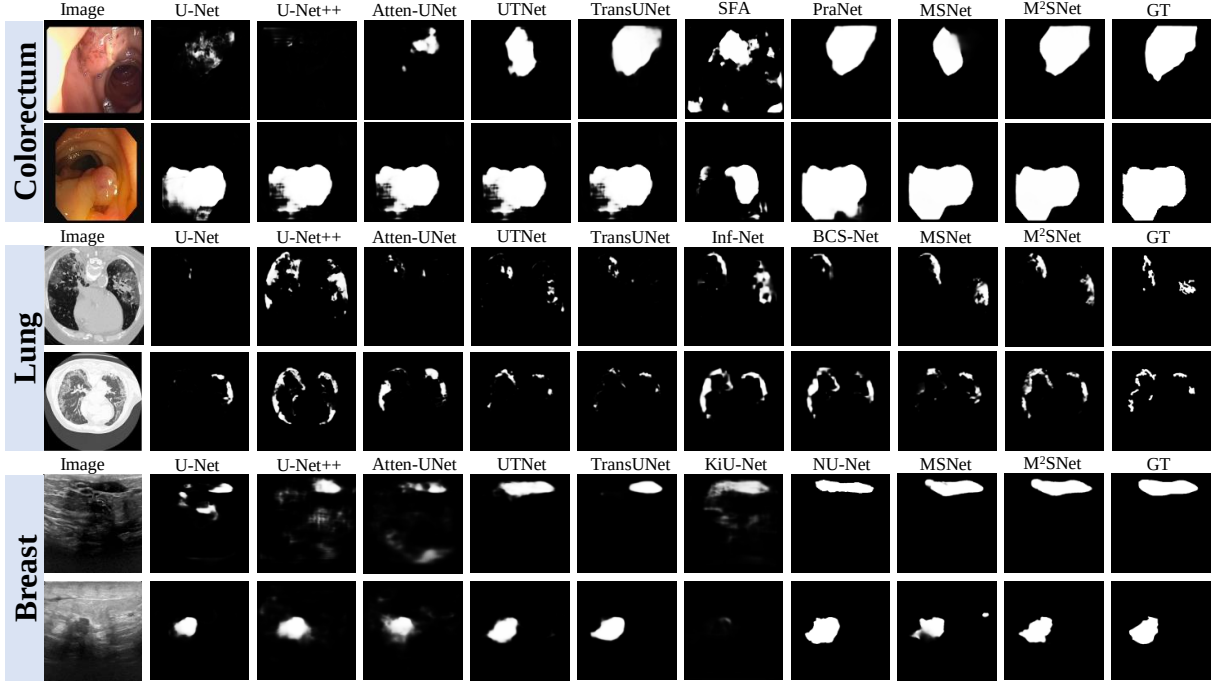
- In Tab. 1, among 30 scores of all *image polyp* datasets, our multi-scale subtraction models (MSNet + M<sup>2</sup>SNet) achieve the best performance in terms of all six metrics. The M<sup>2</sup>SNet even outperforms the video-based polyp segmentation method PNS-Net [40] on the *video polyp* datasets, as shown in Tab. 2.

- Tab. 3 shows performance comparisons on the *lung infection CT* datasets. Compared to the second best method (Inf-Net [42]), M<sup>2</sup>SNet achieves an important improvement of 1.5%, 6.6% and 14.3% in terms of DSC, Precision and MAE, respectively.

- Tab. 4 shows performance comparisons with *breast tumor segmentation* methods. Following most methods [44, 45] in this field, we adopt the four-fold cross-validation strategy. M<sup>2</sup>SNet achieves the best performance in terms of the Jaccard, Precision and Dice metrics, which outperforms representative transformer-based methods, UTNet and TransUNet. Moreover, M<sup>2</sup>SNet has

**Table 7** Quantitative results at different training and inference codes frameworks. “▲” and “▼” represent the model trained on the MSNet [57] and nnU-Net [80] code framework, respectively.

| Methods              | ETIS    |        |                 |                    | CVC-300-TV |        |                |       | Lung Infection |                |       | BUSI       |               |            |
|----------------------|---------|--------|-----------------|--------------------|------------|--------|----------------|-------|----------------|----------------|-------|------------|---------------|------------|
|                      | mDice ↑ | mIoU ↑ | $F_{\beta}^w$ ↑ | $E_{\phi}^{max}$ ↑ | mDice ↑    | mIoU ↑ | $S_{\alpha}$ ↑ | MAE ↓ | DSC ↑          | $S_{\alpha}$ ↑ | MAE ↓ | Jaccard ↑  | Specificity ↑ | Dice ↑     |
| nnU-Net▼             | 0.554   | 0.477  | 0.510           | 0.730              | 0.742      | 0.670  | 0.848          | 0.016 | 0.819          | 0.854          | 0.005 | 71.30±2.29 | 97.46±0.46    | 78.96±2.09 |
| M <sup>2</sup> SNet▲ | 0.749   | 0.678  | 0.712           | 0.872              | 0.876      | 0.805  | 0.918          | 0.010 | 0.795          | 0.855          | 0.006 | 71.52±1.19 | 98.35±0.20    | 79.21±1.38 |
| M <sup>2</sup> SNet▼ | 0.780   | 0.711  | 0.754           | 0.899              | 0.897      | 0.832  | 0.930          | 0.009 | 0.848          | 0.855          | 0.005 | 74.43±1.74 | 98.94±0.22    | 80.26±1.26 |

**Fig. 6** Visual comparison of different medicine-general and medicine-specific methods.

the smallest mean standard deviation (0.97) under the five metrics, which indicates its performance stability.

- Generally speaking, different training and inference frameworks will produce different final performance. For a fair comparison, we train two versions of M<sup>2</sup>SNet, one version follows the MSNet [57] framework and the other based on the popular nnU-Net [80] framework. Benefiting from the nnU-Net framework with many effective tricks, the performance of M<sup>2</sup>SNet can be further improved, as shown in Tab. 7.

- In Tab. 5, we list the top 10 *OCT layer segmentation* results in MICCAI2022 GOALS Challenge. Based on the M<sup>2</sup>SNet, we won the second place (2/100) according to the weighted score of six results in three different layers. It is worth noting that our method ranks the top 1 in four out of six metrics.

- As can be seen from Tab. 1 - Tab. 4, our M<sup>2</sup>SNet consistently surpasses other medicine-general methods in all medical segmentation sub-branches. In Tab. 6, we list the FLOPs and parameters of different medicine-general methods. It can be seen that our method has only 9GB in FLOPs, which has obvious advantages in terms of computational efficiency.

- In Tab. 6, we compare the model efficiency in terms of FLOPs, parameters and inference speed. It can be seen that our method ranks first in terms of FLOPs, which has obvious advantages among both medicine-general and medicine-specific methods. As can be seen from Tab. 1 - Tab. 4, our M<sup>2</sup>SNet consistently surpasses other methods in all medical segmentation sub-branches. Therefore, the proposed M<sup>2</sup>SNet achieves good balance on accuracy and efficiency.

- Fig. 6 depicts a qualitative comparison with

**Table 8** Quantitative comparison of unified models on three different medical lesion segmentation tasks. The best results are shown in **red**.

| Methods             | Backbone        | Lung Infection |              | Breast Lesion |              | Polyp        |              |
|---------------------|-----------------|----------------|--------------|---------------|--------------|--------------|--------------|
|                     |                 | Dice ↑         | mIoU ↑       | Dice ↑        | mIoU ↑       | Dice ↑       | mIoU ↑       |
| UniverSeg [81]      | ResNet-101 [82] | 0.673          | 0.368        | 0.775         | 0.600        | 0.553        | 0.261        |
| SegGPT [83]         | ViT-L [84]      | 0.131          | 0.553        | 0.336         | 0.603        | 0.568        | 0.707        |
| SAM 2 [85]          | Hiera-L [86]    | 0.382          | 0.655        | 0.539         | 0.712        | 0.499        | 0.706        |
| Spider [16]         | ConvNeXt-B [87] | 0.696          | 0.813        | 0.838         | 0.866        | 0.824        | 0.866        |
| M <sup>2</sup> SNet | ConvNeXt-B [87] | <b>0.702</b>   | <b>0.818</b> | <b>0.844</b>  | <b>0.870</b> | <b>0.832</b> | <b>0.866</b> |

**Table 9** Ablation experiments of the subtraction unit, inter-layer multi-scale aggregation and LossNet.

| Methods               | ColonDB |        |                 |                    |  | ETIS    |        |                 |                    |  |
|-----------------------|---------|--------|-----------------|--------------------|--|---------|--------|-----------------|--------------------|--|
|                       | mDice ↑ | mIoU ↑ | $F_{\beta}^w$ ↑ | $E_{\phi}^{max}$ ↑ |  | mDice ↑ | mIoU ↑ | $F_{\beta}^w$ ↑ | $E_{\phi}^{max}$ ↑ |  |
| baseline ( $SU_1^i$ ) | 0.678   | 0.607  | 0.659           | 0.825              |  | 0.588   | 0.549  | 0.532           | 0.707              |  |
| + $SU_2^i$            | 0.731   | 0.652  | 0.703           | 0.861              |  | 0.642   | 0.579  | 0.586           | 0.745              |  |
| + $SU_3^i$            | 0.733   | 0.659  | 0.712           | 0.861              |  | 0.642   | 0.580  | 0.581           | 0.745              |  |
| + $SU_4^i$            | 0.750   | 0.676  | 0.729           | 0.872              |  | 0.643   | 0.580  | 0.585           | 0.757              |  |
| + $SU_5^i$            | 0.749   | 0.676  | 0.729           | 0.878              |  | 0.643   | 0.582  | 0.600           | 0.787              |  |
| + $\mathcal{L}_f$     | 0.755   | 0.678  | 0.737           | 0.883              |  | 0.719   | 0.664  | 0.678           | 0.830              |  |
| $SU \rightarrow AU$   | 0.697   | 0.630  | 0.676           | 0.839              |  | 0.680   | 0.621  | 0.636           | 0.820              |  |

other methods. It can be seen that the results of M<sup>2</sup>SNet have greater advantages in terms of detection accuracy, completeness, and sharpness across different image modalities.

## 4.5 Comparisons with Unified and Generalist Methods

In recent years, with the rapid development of large-scale and foundation models, there has been a growing interest in building *unified* and *generalist* models that can solve multiple tasks with a single set of model parameters. These models [81, 83, 85] aim to break the traditional paradigm of training task-specific networks by introducing shared representations, prompt mechanisms, or in-context learning strategies so that one model can flexibly adapt to diverse downstream tasks. Following the evaluation protocol of SAM-Eva [88], we evaluate our proposed M<sup>2</sup>SNet against representative unified and generalist methods on three challenging cross-modality, cross-lesion segmentation tasks, including lung infection, breast lesion, and polyp segmentation.

As shown in Tab. 8, UniverSeg [81], SegGPT [83], and SAM 2 [85] exhibit clear performance drops when directly applied to cross-modality and cross-lesion scenarios. These models mainly rely on prompt embeddings or in-context examples to handle unseen tasks but still struggle with the strong context-dependence and heterogeneous appearances in medical images. Spider [16], a unified context-dependent segmentation model,

achieves better performance than the above generalist models, demonstrating the effectiveness of high-level concept matching mechanisms. Our M<sup>2</sup>SNet consistently surpasses all the competitors across all three tasks, obtaining the highest Dice and mIoU scores. In our experiments, we jointly train M<sup>2</sup>SNet on the training sets of the three tasks. Thanks to the implicit prompts naturally embedded in modality and lesion types, M<sup>2</sup>SNet can perform joint learning of all data under a single parameter set, thus simultaneously achieving unified modelling and superior performance. By contrast, other methods are trained with a much broader range of tasks and domains but still underperform on these specific cross-modality, cross-lesion settings. Although the scope of M<sup>2</sup>SNet is narrower than that of some generalist models, the results provide an important insight: future universal and unified models can be organized into multiple levels or hierarchies of unification to better balance task coverage and performance for real-world applications.

## 4.6 Ablation Study

We take the common FPN network as the baseline to analyze the contribution of each component.

### 4.6.1 Effectiveness of the subtraction unit, inter-layer multi-scale subtraction aggregation

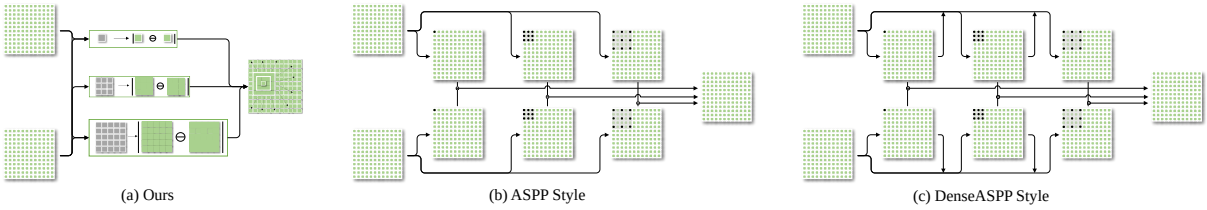
The results are shown in Tab. 9. These defined feature subscripts are the same as those in Fig 2. First, we apply the basic subtraction unit (SU) to the baseline to get a series of  $SU_2^i$  features to participate in the feature aggregation calculated by Equ. 1. The gap between the “+  $SU_2^i$ ” and the baseline demonstrates the effectiveness of the SU. It can be seen that the usage of SU has a significant improvement on the ColonDB dataset compared to the baseline, with the gain of 7.8%, 7.4%, 6.7% and 4.4% in terms of mDice, mIoU,  $F_{\beta}^w$ , and  $E_{\phi}^{max}$ , respectively. Next, we gradually add  $SU_3^i$ ,  $SU_4^i$  and  $MS_5^i$  to achieve inter-layer multi-scale aggregation. The gap between the “+  $SU_5^i$ ” and the “+  $SU_2^i$ ” quantitatively demonstrates the effectiveness of inter-layer multi-scale subtraction strategy. Next, we evaluate the benefit of  $\mathcal{L}_f$ . Compared to the “+  $SU_5^i$ ” model, the “

**Table 10** Ablation experiments of the loss functions.

| Methods                                                                | ETIS    |        |                 |                    | CVC-300-TV |        |                |       | Lung Infection |                |       | BUSI       |               |            |
|------------------------------------------------------------------------|---------|--------|-----------------|--------------------|------------|--------|----------------|-------|----------------|----------------|-------|------------|---------------|------------|
|                                                                        | mDice ↑ | mIoU ↑ | $F_{\beta}^w$ ↑ | $E_{\phi}^{max}$ ↓ | mDice ↑    | mIoU ↑ | $S_{\alpha}$ ↑ | MAE ↓ | DSC ↑          | $S_{\alpha}$ ↑ | MAE ↓ | Jaccard ↑  | Specificity ↑ | Dice ↑     |
| M <sup>2</sup> SNet                                                    | 0.749   | 0.678  | 0.712           | 0.872              | 0.876      | 0.805  | 0.918          | 0.010 | 0.795          | 0.855          | 0.006 | 71.52±1.19 | 98.35±0.20    | 79.21±1.38 |
| M <sup>2</sup> SNet (w/o $\mathcal{L}_f$ )                             | 0.725   | 0.668  | 0.690           | 0.854              | 0.856      | 0.783  | 0.905          | 0.012 | 0.783          | 0.850          | 0.007 | 71.22±1.90 | 98.28±0.22    | 79.02±1.94 |
| M <sup>2</sup> SNet (w/o $\mathcal{L}_{IoU}^w + \mathcal{L}_{BCE}^w$ ) | 0.700   | 0.649  | 0.678           | 0.836              | 0.830      | 0.762  | 0.895          | 0.013 | 0.771          | 0.838          | 0.007 | 70.82±1.85 | 98.01±0.25    | 78.66±1.97 |

**Table 11** Ablation experiments of the kernel scales and weights in intra-layer multi-scale subtraction.

| Methods                                                | ETIS    |        |                 |                    | CVC-300-TV |        |                |       | Lung Infection |                |       | BUSI       |               |            |
|--------------------------------------------------------|---------|--------|-----------------|--------------------|------------|--------|----------------|-------|----------------|----------------|-------|------------|---------------|------------|
|                                                        | mDice ↑ | mIoU ↑ | $F_{\beta}^w$ ↑ | $E_{\phi}^{max}$ ↓ | mDice ↑    | mIoU ↑ | $S_{\alpha}$ ↑ | MAE ↓ | DSC ↑          | $S_{\alpha}$ ↑ | MAE ↓ | Jaccard ↑  | Specificity ↑ | Dice ↑     |
| M <sup>2</sup> SNet (Kernel Scales = 1)                | 0.719   | 0.664  | 0.678           | 0.830              | 0.837      | 0.755  | 0.895          | 0.014 | 0.779          | 0.846          | 0.007 | 70.57±3.38 | 98.29±0.44    | 78.18±3.75 |
| M <sup>2</sup> SNet (Kernel Scales = 1,3)              | 0.738   | 0.672  | 0.695           | 0.858              | 0.862      | 0.787  | 0.906          | 0.012 | 0.789          | 0.850          | 0.007 | 71.08±1.43 | 98.33±0.51    | 78.88±1.89 |
| M <sup>2</sup> SNet (Kernel Scales = 1,3,5)            | 0.749   | 0.678  | 0.712           | 0.872              | 0.876      | 0.805  | 0.918          | 0.010 | 0.795          | 0.855          | 0.006 | 71.52±1.19 | 98.35±0.20    | 79.21±1.38 |
| M <sup>2</sup> SNet (Kernel Scales = 1,3,5,7)          | 0.748   | 0.678  | 0.713           | 0.868              | 0.875      | 0.808  | 0.921          | 0.010 | 0.799          | 0.861          | 0.007 | 71.56±1.18 | 98.40±0.21    | 79.25±1.36 |
| M <sup>2</sup> SNet (Average Kernel → Gaussian Kernel) | 0.670   | 0.605  | 0.650           | 0.817              | 0.827      | 0.750  | 0.886          | 0.014 | 0.753          | 0.830          | 0.007 | 68.18±2.34 | 97.74±0.83    | 77.52±2.42 |

**Fig. 7** Illustration of different multi-scale modules applied in the subtraction unit.**Table 12** Quantitative comparisons of different multi-scale styles in intra-layer multi-scale subtraction design. Positive and negative gains are highlighted in **red** and **blue**, respectively.

| Methods                         | Efficiency |          | ETIS    |        |                 |                    | CVC-300-TV |        |                |        | Lung Infection |                |        | BUSI       |               |            |
|---------------------------------|------------|----------|---------|--------|-----------------|--------------------|------------|--------|----------------|--------|----------------|----------------|--------|------------|---------------|------------|
|                                 | FLOPs ↓    | Params ↓ | mDice ↑ | mIoU ↑ | $F_{\beta}^w$ ↑ | $E_{\phi}^{max}$ ↓ | mDice ↑    | mIoU ↑ | $S_{\alpha}$ ↑ | MAE ↓  | DSC ↑          | $S_{\alpha}$ ↑ | MAE ↓  | Jaccard ↑  | Specificity ↑ | Dice ↑     |
| MSNet                           | 9.0 G      | 27.7 M   | 0.719   | 0.664  | 0.678           | 0.830              | 0.837      | 0.755  | 0.895          | 0.014  | 0.779          | 0.846          | 0.007  | 70.57±3.38 | 98.29±0.44    | 78.18±3.75 |
| M <sup>2</sup> SNet (ASPP)      | 13.8 G     | 29.2 M   | 0.711   | 0.644  | 0.671           | 0.838              | 0.827      | 0.752  | 0.890          | 0.014  | 0.809          | 0.855          | 0.006  | 60.23±1.94 | 98.14±0.42    | 68.52±1.82 |
| Δ gains                         | -53.3%     | -5.4%    | -1.1%   | -3.0%  | -1.0%           | -1.0%              | -1.2%      | -0.4%  | -0.6%          | 0.0%   | +3.9%          | +1.1%          | +14.3% | -14.7%     | -0.2%         | -12.4%     |
| M <sup>2</sup> SNet (DenseASPP) | 17.5 G     | 29.6 M   | 0.709   | 0.645  | 0.668           | 0.833              | 0.869      | 0.798  | 0.916          | 0.010  | 0.798          | 0.853          | 0.006  | 63.11±4.63 | 98.42±0.59    | 71.71±4.33 |
| Δ gains                         | -94.4%     | -6.9%    | -1.4%   | -2.9%  | -1.5%           | -0.4%              | +3.8%      | +5.7%  | +2.3%          | +28.6% | +2.4%          | +0.8%          | +14.3% | -10.6%     | +0.1%         | -8.3%      |
| M <sup>2</sup> SNet (Ours)      | 9.0 G      | 27.7 M   | 0.749   | 0.678  | 0.712           | 0.872              | 0.876      | 0.805  | 0.918          | 0.010  | 0.795          | 0.855          | 0.006  | 71.52±1.19 | 98.35±0.20    | 79.21±1.38 |
| Δ gains                         | 0.0%       | 0.0%     | +4.2%   | +2.1%  | +5.0%           | +5.1%              | +4.7%      | +6.6%  | +2.6%          | +28.6% | +2.1%          | +1.1%          | +14.3% | +1.3%      | +0.1%         | +1.3%      |

+  $\mathcal{L}_f$  ” achieves significant performance improvement on the ETIS dataset, with the gain of 11.8%, 14.1%, 13.0% and 5.5% in terms of mDice, mIoU,  $F_{\beta}^w$ , and  $E_{\phi}^{max}$ , respectively. Besides, we replace all subtraction units with the element-wise addition units (AU) and compare their performance. It can be seen that our subtraction units have significant advantage and no additional parameters are introduced.

#### 4.6.2 Effectiveness of loss function

In Tab. 10, we thoroughly verify the effectiveness of the loss function used in M<sup>2</sup>SNet. The gap between M<sup>2</sup>SNet and M<sup>2</sup>SNet (w/o  $\mathcal{L}_f$ ) demonstrates the general effectiveness of LossNet for different medical segmentation tasks. At the same time, without the auxiliary of LossNet, the performance of M<sup>2</sup>SNet itself is good enough to surpass most of the methods in Tab. 1 - Tab. 4. The gap between M<sup>2</sup>SNet and M<sup>2</sup>SNet (w/o  $\mathcal{L}_{IoU}^w +$

$\mathcal{L}_{BCE}^w$ ) shows the necessity of the weighted IoU and BCE loss, which provide a basic foreground region guidance for LossNet to focus on supervising multi-level lesion regions without distracting in the background area.

#### 4.6.3 Effectiveness of the intra-layer multi-scale subtraction design

Compared to the previous MSNet, the M<sup>2</sup>SNet replace all the original single-scale subtraction unit with the stronger multi-scale subtraction unit. As shown in Tab. 11, more scales of feature fusion can help improve the overall performance and multi-scale filters of size [1, 3, 5] and [1, 3, 5, 7] have close performance. Next, we replace the fixed full one weights with Gaussian weights. It can be seen that Gaussian weights significantly degrade the performance of the model in all tasks. We think that the subtraction unit should try to maintain the characteristic distribution of the input

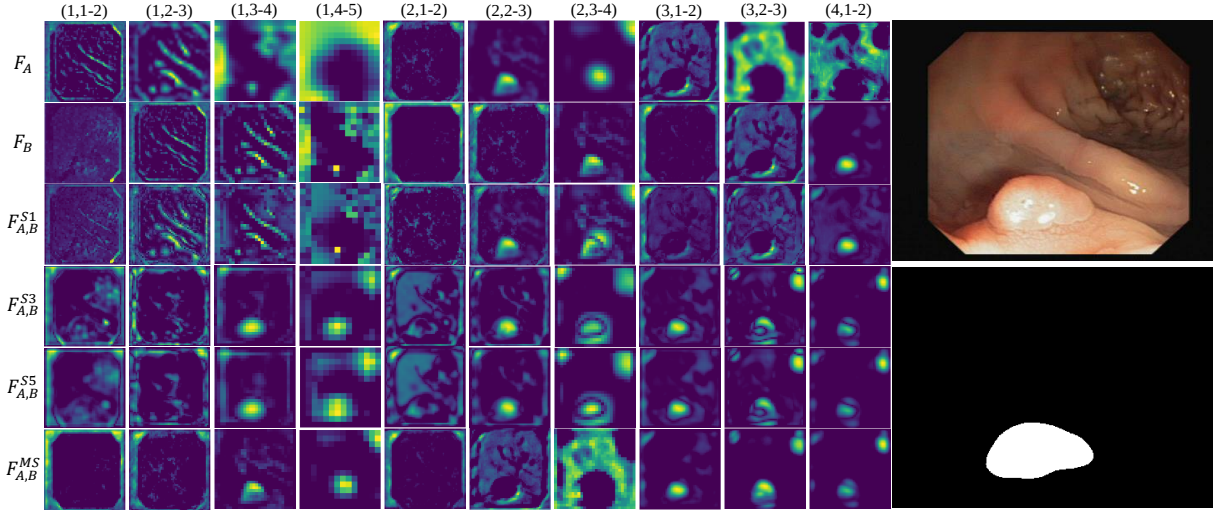


**Fig. 8** Visualization of the feature maps in the multi-scale in multi-scale subtraction module.

features itself, but the Gaussian weights rigidly changes the spatial distribution of the original features, causing an extra burden for the subsequent decoder. Therefore, we choose the multi-scale convolution filters with fixed full one weights of size  $[1, 3, 5]$  as the final setting. To further show the advantages of our intra-layer multi-scale design, we apply other popular multi-scale modules (i.e., ASPP [17] and DenseASPP [18]) to the subtraction unit and these architectures are shown in Fig. 7. In Tab. 12, we thoroughly compare both the efficiency and accuracy of these three structures. It can be seen that “M<sup>2</sup>SNet (Ours)” has a significant performance gain in terms of fourteen metrics on four challenges datasets under different tasks. However, the other two models not only increase the computational burden by more than 50%, but also produce negative gains in multiple datasets. Therefore, the proposed intra-layer multi-scale subtraction design can be taken as a new baseline for future research in subtraction family.

To more intuitively show the differential information from different scales, we visualize all the features of the multi-scale in multi-scale subtraction module, as shown in Fig. 8. We can see that the multi-scale in multi-scale subtraction module can clearly highlight the difference between high-level features and other level features and propagate its localization effect to the low-level ones. At the same level, the intra-layer multi-scale aggregation design can comprehensively capture

both the subtle and regional difference features. Thus, both the global structural information and local boundary information is well depicted in the enhanced features of different levels.

## 5 Discussion

**Multi-scale Subtraction Unit:** Different from previous addition and concatenation operations, using subtraction in multi-level structure make resulted features input to the decoder have much less redundancy among different levels and their level-specific properties are significantly enhanced. In this work, we further explore the potential of the subtraction unit in intra-layer multi-scale fusion. How to improve the accuracy while maintaining the same efficiency as the single-scale one is the key challenge. We provide the solution of using multi-scale convolution filters with fixed parameters. Compared to the single-scale design, multi-scale subtraction unit can enable the network to collect more complementary information both in pixel-pixel and neighbor-neighbor levels. The advantages of multi-scale subtraction unit in terms of efficiency and accuracy can be seen in Tab. 12. Multi-scale information extraction and feature aggregation are two general problems in the field of computer vision. Our multi-scale subtraction unit can solve both of them at once. We think this new paradigm can drive more researches on the subtraction operation in the future.

**LossNet:** LossNet is similar in form to perception

loss [89] that has been applied in many tasks, such as style transfer and inpainting. While in those vision tasks, the perception-like loss is mainly used to speed the convergence of GAN and obtain high frequency information and ease checkerboard artifacts, but it does not bring obvious accuracy improvement. In our paper, the inputs are binary segmentation masks, LossNet can directly target the geometric features of the lesion and perform joint supervisions from the contour to the body, thereby improving the overall segmentation accuracy.

## 6 Conclusion

In this paper, we rethink previous addition-based or concatenation-based methods and present a simple yet general multi-scale in multi-scale subtraction network (M<sup>2</sup>SNet) for more efficient medical image segmentation. Based on the proposed intra-layer multi-scale subtraction unit, we pyramidally aggregate adjacent levels to extract lower-order and higher-order cross-level complementary information and combine with level-specific information to enhance multi-scale feature representation. Besides, we design a loss function based on a training-free network to supervise the prediction from different feature levels, which can optimize the segmentation on both structure and details during the backward phase. Experimental results on 11 benchmark datasets towards 4 medical segmentation tasks demonstrate that the proposed model outperforms various state-of-the-art methods.

## Acknowledgments

This work was supported by Dalian Science and Technology Innovation Foundation under Grant 2023JJ12GX015, and by the National Natural Science Foundation of China under Grant 62276046 and 62431004. (Corresponding author: Lihe Zhang.)

## Conflicts of Interests

The authors declared that they have no conflicts of interest to this work.

## References

- [1] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017.
- [2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.
- [3] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D Johansen. Resunet++: An advanced architecture for medical image segmentation. In *IEEE ISM*, pages 225–2255, 2019.
- [4] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *MICCAI*, pages 263–273, 2020.
- [5] Xiaoqi Zhao, Youwei Pang, Chenyang Yu, Lihe Zhang, Huchuan Lu, Shijian Lu, Georges El Fakhri, and Xiaofeng Liu. Unimrseg: Unified modality-relax segmentation via hierarchical self-supervised compensation. In *NeurIPS*, 2025.
- [6] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. A bi-directional message passing model for salient object detection. In *CVPR*, pages 1741–1750, 2018.
- [7] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In *ECCV*, pages 35–51, 2020.
- [8] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Towards diverse binary segmentation via a simple yet general gated network. *IJCV*, 132(10):4157–4234, 2024.

- [9] Yunhe Gao, Mu Zhou, and Dimitris N Metaxas. Utnet: a hybrid transformer architecture for medical image segmentation. In *MICCAI*, pages 61–71, 2021.
- [10] Yanfeng Zhou, Lingrui Li, Le Lu, and Minfeng Xu. nnwnet: Rethinking the use of transformers in biomedical image segmentation and calling for a unified evaluation benchmark. In *CVPR*, pages 20852–20862, 2025.
- [11] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024.
- [12] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE TMI*, 39(6):1856–1867, 2019.
- [13] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106:107404, 2020.
- [14] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *CVPR*, pages 9413–9422, 2020.
- [15] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoomnext: A unified collaborative pyramid network for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9205–9220, 2024.
- [16] Xiaoqi Zhao, Youwei Pang, Wei Ji, Baicheng Sheng, Jiaming Zuo, Lihe Zhang, and Huchuan Lu. Spider: A unified framework for context-dependent concept segmentation. In *ICML*, pages 60906–60926, 2024.
- [17] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40:834–848, 2017.
- [18] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, pages 3684–3692, 2018.
- [19] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In *CVPR*, pages 8582–8591, 2020.
- [20] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *CVPR*, pages 8772–8781, 2021.
- [21] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *ECCV*, pages 121–136, 2018.
- [22] Jinhee Kim and Wonjun Kim. Attentive feedback feature pyramid network for shadow detection. *IEEE SPL*, 27:1964–1968, 2020.
- [23] Zijun Deng, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Jing Qin, Guoqiang Han, and Pheng-Ann Heng. R3net: Recurrent residual refinement network for saliency detection. In *IJCAI*, pages 684–690, 2018.
- [24] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, pages 7254–7263, 2019.
- [25] Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate rgb-d salient object detection via collaborative learning. In *ECCV*, pages 52–69, 2020.

- [26] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *CVPR*, pages 11591–11601, 2021.
- [27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.
- [28] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thirteenth Asilomar Conference on Signals, Systems & Computers*, volume 2, pages 1398–1402, 2003.
- [29] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *CVPR*, pages 2160–2170, 2022.
- [30] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
- [31] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segformer: Transformer for semantic segmentation. In *ICCV*, pages 7262–7272, 2021.
- [32] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, pages 12077–12090, 2021.
- [33] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *ECCV*, pages 205–218, 2022.
- [34] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *COLM*, 2024.
- [35] Zifu Wan, Pingping Zhang, Yuhao Wang, Silong Yong, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Sigma: Siamese mamba network for multi-modal semantic segmentation. In *WACV*, pages 1734–1744, 2025.
- [36] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In *MICCAI*, pages 578–588, 2024.
- [37] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.
- [38] Yuqi Fang, Cheng Chen, Yixuan Yuan, and Kai-yu Tong. Selective feature aggregation network with area-boundary constraints for polyp segmentation. In *MICCAI*, pages 302–310, 2019.
- [39] Xiuquan Du, Xuebin Xu, Jiajia Chen, Xuejun Zhang, Lei Li, Heng Liu, and Shuo Li. Um-net: Rethinking icgnet for polyp segmentation with uncertainty modeling. *Medical Image Analysis*, 99:103347, 2025.
- [40] Ge-Peng Ji, Yu-Cheng Chou, Deng-Ping Fan, Geng Chen, Huazhu Fu, Debesh Jha, and Ling Shao. Progressively normalized self-attention network for video polyp segmentation. In *MICCAI*, pages 142–152, 2021.
- [41] Naveen Paluru, Aveen Dayal, Håvard Bjørke Jenssen, Tomas Sakinis, Linga Reddy Cenkeramaddi, Jaya Prakash, and Phaneendra K Yalavarthy. Anam-net: Anamorphic depth embedding-based lightweight cnn for segmentation of anomalies in covid-19 chest ct images. *IEEE TNNLS*, 32(3):932–946, 2021.
- [42] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE TMI*, 39(8):2626–2637, 2020.
- [43] Runmin Cong, Haowei Yang, Qiuping Jiang, Wei Gao, Haisheng Li, Cong Wang, Yao

- Zhao, and Sam Kwong. Bcs-net: Boundary, context, and semantic for automatic covid-19 lung infection segmentation from ct images. *IEEE TIM*, 71:1–11, 2022.
- [44] Michal Byra, Piotr Jarosik, Aleksandra Szubert, Michael Galperin, Haydee Ojeda-Fournier, Linda Olson, Mary O’Boyle, Christopher Comstock, and Michael Andre. Breast mass segmentation in ultrasound with selective kernel u-net convolutional neural network. *BSPC*, 61:102027, 2020.
- [45] Gong-Ping Chen, Lei Li, Yu Dai, and Jian-Xun Zhang. Nu-net: An unpretentious nested u-net for breast tumor segmentation. *arXiv preprint arXiv:2209.07193*, 2022.
- [46] Ruifei Zhang, Guanbin Li, Zhen Li, Shuguang Cui, Dahong Qian, and Yizhou Yu. Adaptive context selection for polyp segmentation. In *MICCAI*, pages 253–262, 2020.
- [47] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *CVPR*, pages 3203–3212, 2017.
- [48] Lei Li, Juan Qin, Lianrong Lv, Mengdan Cheng, Biao Wang, Dan Xia, and Shike Wang. Icnnet++: an inception-cbam network based on unet++ for mr spine image segmentation. *International Journal of Machine Learning and Cybernetics*, pages 1–13, 2023.
- [49] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *CVPR*, pages 3203–3212, 2017.
- [50] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. Contrast prior and fluid pyramid integration for rgbd salient object detection. In *CVPR*, pages 3922–3931, 2019.
- [51] Xiaoqi Zhao, Lihe Zhang, Youwei Pang, Huchuan Lu, and Lei Zhang. A single stream network for robust and real-time rgb-d salient object detection. In *ECCV*, pages 646–662, 2020.
- [52] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.
- [53] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.
- [54] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International workshop on machine learning in medical imaging*, pages 379–387, 2017.
- [55] Ken CL Wong, Mehdi Moradi, Hui Tang, and Tanveer Syeda-Mahmood. 3d segmentation with exponential logarithmic loss for highly unbalanced object sizes. In *MICCAI*, pages 612–619, 2018.
- [56] Saeid Asgari Taghanaki, Yefeng Zheng, S Kevin Zhou, Bogdan Georgescu, Puneet Sharma, Daguang Xu, Dorin Comaniciu, and Ghassan Hamarneh. Combo loss: Handling input and output imbalance in multi-organ segmentation. *CMIG*, 75:24–33, 2019.
- [57] Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Automatic polyp segmentation via multi-scale subtraction network. In *MICCAI*, pages 120–130, 2021.
- [58] Jun Wei, Shuhui Wang, and Qingming Huang. F<sup>3</sup>net: Fusion, feedback and focus for salient object detection. In *AAAI*, pages 12321–12328, 2020.
- [59] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, pages 7479–7489, 2019.
- [60] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection

in colonoscopy videos using shape and context information. *IEEE TMI*, 35(2):630–644, 2015.

- [61] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *IJCARS*, 9(2):283–293, 2014.
- [62] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MMM*, pages 451–462, 2020.
- [63] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *JHE*, 2017, 2017.
- [64] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilarino. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *CMIG*, 43:99–111, 2015.
- [65] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012.
- [66] Covid-19 ct segmentation dataset. In <https://medicalsegmentation.com/COVID19/>, accessed April, 2020.
- [67] Covid-19 ct lung and infection segmentation dataset. In <https://zenodo.org/record/3757476>, accessed April 20, 2020.
- [68] Seung Yeon Shin, Soochahn Lee, Il Dong Yun, Sun Mi Kim, and Kyoung Mu Lee. Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images. *IEEE TMI*, 38(3):762–774, 2018.
- [69] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.
- [70] Huihui Fang, Fei Li, Huazhu Fu, Junde Wu, Xiulan Zhang, and Yanwu Xu. Dataset and evaluation algorithm design for goals challenge. *arXiv preprint arXiv:2207.14447*, 2022.
- [71] Huihui Fang, Fei Li, Huazhu Fu, Junde Wu, Xiulan Zhang, and Yanwu Xu. Dataset and evaluation algorithm design for goals challenge. In *International Workshop on Ophthalmic Medical Image Analysis*, pages 135–142, 2022.
- [72] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, pages 248–255, 2014.
- [73] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, pages 4548–4557, 2017.
- [74] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In *IJCAI*, 2018.
- [75] Fei Shan, Yaozong Gao, Jun Wang, Weiya Shi, Nannan Shi, Miaofei Han, Zhong Xue, Dinggang Shen, and Yuxin Shi. Lung infection quantification of covid-19 in ct images with deep learning. *arXiv preprint arXiv:2003.04655*, 2020.
- [76] Gongping Chen, Lei Li, Yu Dai, Jianxun Zhang, and Moi Hoon Yap. Aau-net: An adaptive attention u-net for breast lesions segmentation in ultrasound images. *IEEE TMI*, 2022.
- [77] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

- [78] Zuyao Chen, Qianqian Xu, Runmin Cong, and Qingming Huang. Global context-aware progressive aggregation network for salient object detection. In *AAAI*, pages 10599–10606, 2020.
- [79] Tao Zhou, Huazhu Fu, Geng Chen, Yi Zhou, Deng-Ping Fan, and Ling Shao. Specificity-preserving rgb-d saliency detection. In *ICCV*, pages 4681–4691, 2021.
- [80] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.
- [81] Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. In *ICCV*, pages 21438–21451, 2023.
- [82] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
- [83] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Towards segmenting everything in context. In *ICCV*, pages 1130–1140, 2023.
- [84] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.
- [85] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *ICLR*, 2025.
- [86] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *ICML*, pages 29441–29454, 2023.
- [87] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022.
- [88] Xiaoqi Zhao, Youwei Pang, Shijie Chang, Yuan Zhao, Lihe Zhang, Chenyang Yu, Hanqi Liu, Jiaming Zuo, Jinsong Ouyang, Weisi Lin, et al. Inspiring the next generation of segment anything models: Comprehensively evaluate sam and sam 2 with diverse prompts towards context-dependent concepts under different scenes. *arXiv preprint arXiv:2412.01240*, 2024.
- [89] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016.