Exploiting Scale-Variant Attention for Segmenting Small Medical Objects

Wei Dai, *Student Member, IEEE*, Rui Liu, Zixuan Wu, Tianyi Wu, Min Wang, *Student Member, IEEE*, Junxian Zhou, Yixuan Yuan, *Senior Member, IEEE*, and Jun Liu, *Senior Member, IEEE*

***Abstract*—Early detection and accurate diagnosis can predict the risk of malignant disease transformation, thereby increasing the probability of effective treatment. Identifying mild syndrome with small pathological regions serves as an ominous warning and is fundamental in the early diagnosis of diseases. While deep learning algorithms, particularly convolutional neural networks (CNNs), have shown promise in segmenting medical objects, analyzing small areas in medical images remains challenging. This difficulty arises due to information losses and compression defects from convolutional and pooling operations in CNNs, which become more pronounced as the network deepens, especially for small medical objects. To address these challenges, we propose a novel scale-variant attention-based network (SvANet) for accurately segmenting small-scale objects in medical images. The SvANet consists of scale-variant attention (SvAttn), cross-scale guidance, Monte Carlo attention (MCAttn), and Vision Transformer (ViT), which incorporates cross-scale features and alleviates compression artifacts for enhancing the discrimination of small medical objects. Quantitative experimental results demonstrate the superior performance of SvANet, achieving 96.12%, 96.11%, 89.79%, 84.15%, 80.25%, 73.05%, and 72.58% in mean Dice (mDice) coefficient for segmenting kidney tumors, skin lesions, hepatic tumors, polyps, surgical excision cells, retinal vasculatures, and sperms, which occupy less than 1% of the image areas in KiTS23, ISIC 2018, ATLAS, PolypGen, TissueNet, FIVES, and SpermHealth datasets, respectively.**

***Index Terms*—Attention mechanisms, medical image segmentation, Monte Carlo method, small object detection, Vision Transformer (ViT).**

Received 23 September 2024; revised 24 April 2025, 10 August 2025, and 7 November 2025; accepted 15 December 2025. This work was supported by the Research Grant Council (RGC) of Hong Kong under Grant GRF 11217922, Grant 11212321, and Grant ECS 21212720. *(Corresponding authors: Jun Liu; Yixuan Yuan.)*

Wei Dai, Zixuan Wu, Tianyi Wu, and Junxian Zhou are with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong, China.

Rui Liu is with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong, China, and also with the School of Automation, Guangdong Polytechnic Normal University, Guangzhou 510642, China.

Min Wang is with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong, China, and also with the School of Mechanical and Electrical Engineering, Central South University, Changsha 410083, China.

Yixuan Yuan is with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, China (e-mail: yxyuan@ee.cuhk.edu.hk).

Jun Liu is with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong, China, and also with the Department of Data and Systems Engineering, The University of Hong Kong, Hong Kong, China (e-mail: djliu@hku.hk).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNNLS.2025.3645355, provided by the authors.

Data is available on-line at https://github.com/anthonyweidai/SvANet

Digital Object Identifier 10.1109/TNNLS.2025.3645355

I. Introduction

IT IS essential to detect and diagnose diseases or conditions at their earliest stages, often prior to the manifestation of symptoms. In the early stages of diseases such as glaucoma [1], skin cancer [2], colorectal cancer [3], hepatocellular carcinoma [4], and renal cancer [5], the pathological areas are comparatively small and challenging to detect. The morphometrics of these infected areas are believed to reflect the risk and progression of diseases (e.g., cancer precursors) [1], [2], [3], [4], [5], [6], [7]. Accurately delineating the boundaries of lesions is crucial for their complete resection. Cell-level imaging analysis is also a cutting-edge field with various clinical applications, such as tumor resection analysis [6] and in vitro fertilization [8]. However, examining cells can be challenging due to differences in size, morphology, and density, especially on a small scale.

A considerable number of images from various modalities contain numerous lesions that occupy less than 10% of the total image area [1], [2], [3], [4], [5], [6], as detailed in Table I. Deep learning algorithms, which employ convolution and pooling, can result in the loss of details for small objects, leading to noticeable compression artifacts. To address the diminished image resolution and information loss, strategies include upscaling input data [9], expanding network architectures [10], [11], [12], [13], [14], [15], tuning loss functions [13], [16], [17], and postprocessing [13]. The attention mechanism is an efficient method for enhancing the focus on the understated region by extending network variants [12], [13], [14], [18], [19], [20], [21], [22]. However, small medical objects pose unique challenges: they not only lack sufficient pixels and information for straightforward local representation extraction, but their relatively small size (e.g., occupying less than 1% of the images) makes them difficult to capture using global operations such as global-average pooling and multihead self-attention.

To effectively analyze those comparatively small objects, it is crucial to understand the changes in feature maps across different levels of compression. Drawing an analogy from animal eyes, which adjust the shape of their crystalline lenses to tune visual perception of objects at varying distances, we introduce a scale-variant attention (SvAttn) method. The SvAttn method is integrated within a cross-scale guidance module for "*tracing*" the behavior of small medical objects by using cross-level features, as demonstrated in Fig. 1(a).

TABLE I

DATASET DETAILS: MEDICAL OBJECTS WITHIN EACH DATASET ARE CATEGORIZED BY AREA RATIOS: BELOW 1% (ULTRASMALL), BELOW 10% (SMALL), AND 100% (ALL)

Dataset	Image capture	Number of image (train + test)	Object area ratio	Number of object		
				ultra small	small	all
FIVES [1]	Oph	600 + 200	0.351% ∼ 52.020%	4	145	798
ISIC 2018 [2]	Derm	2594 + 100	0.288% ∼ 98.575%	52	1084	2694
PolypGen [3]	COL	1230 + 307	0.003% ∼ 85.850%	81	895	1411
ATLAS [4]	MRI	997 + 249	0.001% ∼ 25.826%	274	1084	1464
KiTS23 [5]	CT	1703 + 426	0.001% ∼ 13.790%	665	1533	1539
TissueNet [6]	WSI	2580 + 1324	0.002% ∼ 9.836%	9096	9437	9437
SpermHealth	Microsc	118 + 30	0.042% ∼ 0.651%	1456	1456	1456

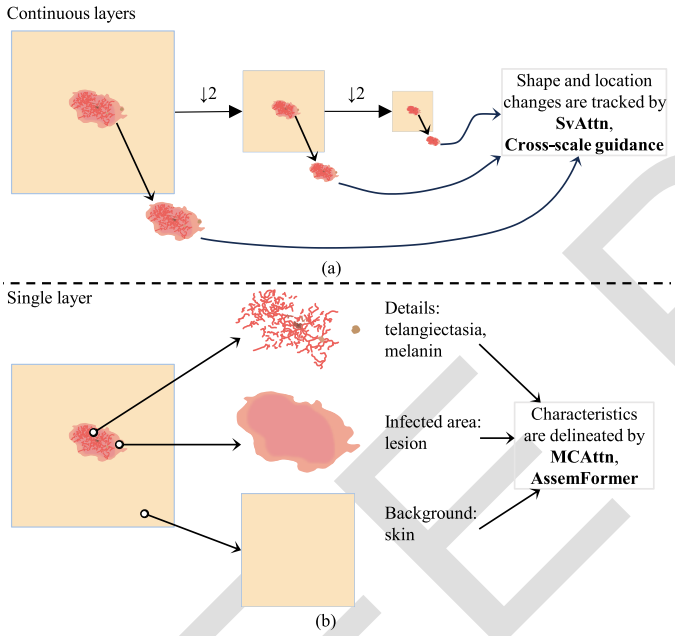

Fig. 1. Illustration of the intuitions of the core components of the proposed methods. The example image depicts a skin lesion. (a) Tracing. (b) Detailing.

The SvAttn method stochastically samples attention maps from different compression stages, enabling the network to discern differences and similarities in object features at these various stages. Concurrently, the cross-scale guidance module leverages high-resolution feature maps from less-compressed stages, enriching supplemental information for small medical objects.

While exploring the evolution of features across different compression stages is critical, it is equally important to accurately identify small regions of interest within a single stage. Traditional attention mechanisms in deep learning typically produce a fixed-dimension attention map [12], [23], [24], [25], which concentrate on central features while often overlooking the extensive contextual information present in the background, vital for clinical interpretation. For instance, in an abdominal slice image, the standard positional relationships among various organs (e.g., stomach, liver, kidneys, spleen, and bone marrow) aid in accurately locating objects of interest within a narrower range. Inspired by this observation, we introduce the Monte Carlo attention (MCAttn) and an assembly-based convolutional Vision Transformer (AssemFormer) to enhance positional relationships for "*detailing*" features of small medical objects, as illustrated in Fig. 1(b). The MCAttn employs different attention map sizes that diversify the receptive field and establish relationships among objects from different regions. The AssemFormer combines convolution with transformer specification to simultaneously extract local and global information, thereby enhancing the capability of feature learning.

The key contributions of this study are highlighted as follows:

1) We propose SvANet, a new network that utilizes two novel attention mechanisms and a Vision Transformer (ViT) to identify small medical objects. To the best of our knowledge, this is the first study to systematically analyze small medical objects across seven medical image modalities and diverse object types (i.e., retinal vessels, skin lesions, polyps, livers, kidneys, tumors, tissue cells, and sperms).
2) We introduce the SvAttn method, which captures the positional and morphological essence of small medical objects by generating attention maps based on the progressively compressed feature maps.
3) We develop the MCAttn module, which generates attention maps at different scales in a single stage by using agnostic pooling output sizes. MCAttn learns the object relations and spatial information of small medical objects with consideration of both their position and morphology.
4) We present AssemFormer, which enables the incorporation of both local spatial hierarchies and interpatch representations, providing a comprehensive understanding of the image data.
5) Equipped with these novel designs, SvANet achieved top-level performance in segmenting medical objects with less than 10% area ratio on seven benchmark datasets, outperforming seven advanced methods. For instance, SvANet achieved the highest mDice of 89.79% and the lowest MAE of 1.6×10^{-3} in distinguishing livers and liver tumors that cover less than 1% regions in abdominal slices.

II. RELATED WORKS

A. Medical Object Segmentation

Surface structures, shapes, and sizes are critical in characterizing medical objects. The morphometric data collected from various devices and patients present a complex and challenging landscape for analysis. In recent years, deep learning algorithms have shown remarkable potential in enhancing diagnostic accuracy, reducing costs, and interpreting images of diverse medical objects across various imaging modalities. Thes modalities include ophthalmoscopy (Oph), dermatoscopy (Derm), colonoscopy (COL), magnetic resonance imaging (MRI), computerized tomography (CT), whole slide imaging (WSI), microscopy (Microsc), electron microscopy (EM), and X-ray [10], [11], [12], [13], [14], [15], [18], [19], [26], [27].

One widely adopted structure for analyzing medical images is the encoder–decoder-based construction, introduced by Long et al. [28]. This approach involves extracting derived features from an encoder and using a decoder to generate the final segmentation mask. Building upon the encoder–decoder structure, Ronneberger et al. [10] introduced "U-shaped" architectures, which connect the limbs by using convolution (U-Net) to disseminate information for segmenting tumor cells or general objects. To further enhance the fusion of multiscale features in analyzing medical images across CT, MRI, and EM modalities, Zhou et al. [11] introduced U-Net++, an extension of U-Net incorporating densely connected links. In addition, Isensee et al. [26] broadened the application of U-Net from 2-D to 3-D medical imaging by self-adaptive configurations (nnUNet).

To improve the performance of encoder–decoder architectures in perceiving medical images, advanced techniques have been suggested. These techniques consist of attention mechanisms [29], multinetwork branches [13], contrastive learning [17], and feature interactions [13], [29]. For example, Fan et al. [18] suggested a parallel reverse attention network (PraNet) by integrating an upsampled feature generated by the medium decoder to discern clearer bounaries of polyps in COL images. Pan et al. [13] introduced a three-branch "U-shaped" framework to ameliorate feature interactions by postprocessing outputs from three branches with the watershed algorithm for examining nuclei. In the study of CT scans of the pancreas, Miao et al. [17] boosted the multibranch architecture by facilitating contrastive learning and a consistency loss function. When assessing polyps from six unique medical centers, Jha et al. [29] integrated transformers with residual connections of convolution to propagate information from the encoder to the decoder. Despite the promising results of the research above in medical image recognition, one aspect overlooked is the size of medical objects, particularly small-scale objects.

B. Small Medical Object Segmentation

The convolutional and pooling operations in deep learning algorithms compress input data, thus damaging the morphological characteristics of medical objects. To mitigate information loss when reducing image resolution, one common method is to upscale the input images to generate high-resolution feature maps of small objects [9]. Another data augmentation method involves concatenating three adjacent 2-D slices to generate a mixed 2-D image, which helps to broaden the sample sizes of small objects [30]. However, these preprocessing methods can be time-consuming during training or testing due to the need for image augmentation and feature dimension enlargement.

Another promising method to reduce compression artifacts involves expanding network variants by incorporating techniques, such as atrous convolution [31], skip connections [10], [11], [26], feature pyramids [32], [33], multiple branches [12], [13], [27], [34], or attention mechanisms [12], [13], [14], [18], [19], which captures cross-scale features and contributes to magnify small objects. For example, Zhao et al. [32] introduced the pyramid scene parsing network (PSPNet), which employs pyramid pooling and concatenates upsampled features from multiple scales to improve context the feature learning. Lou et al. [19] proposed a context axial reverse attention network (CaraNet) to detect small polyps and brain tumors with less than 5% size ratios. However, CaraNet lacks sufficient interpretability regarding its practicality for segmenting small medical objects, appearing more as a general design suited to the segmentation task.

Designing new loss functions is another practical way to boost small object identification. Guo et al. [16] proposed a loss function that adopts the boundary pixel's neighbors to enhance the small object segmentation. In addition, Pan et al. [13] combined six different loss functions for nuclei diagnosis. Instead, Liu et al. [35] conducted backpropagation using only those prioritized losses based on the rank of object pixel counts and the magnitude of loss values. However, the disadvantage of replacing the loss function is that it may not be semantically understandable [16], [17], [35] or it can increase the computational complexity [13]. Postprocessing, such as the watershed algorithm [13], can also enhance small object segmentation. However, postprocessing is a distinct step from the segmentation model, and the network cannot adjust its weights to the postprocessing results.

Previously, object sizes were quantified by object category [16], [35], [36], number of pixels [9], or size ratio [19] in the images. However, the size of the same object can vary based on the distance between the object and the camera, and computer vision algorithms often resize the entire input image, resulting in changes in pixel numbers. Thus, relying solely on the object category or the number of pixels cannot accurately describe the size. This study categorizes medical object sizes using area ratios, providing a more appropriate measure tailored for medical images.

C. Attention Mechanisms

The attention mechanism is extensively employed in semantic segmentation to prioritize salient features. Various approaches have been proposed to incorporate attention in different ways. Hu et al. [23] applied the squeeze–excitation (SE) method to generate channel attention for learning semantic representations. Zhou et al. [12] employed channel attention to capture boundary-aware features for enhancing polyp segmentation. To further extract spatial information, Woo et al. [24] combined channel attention with spatial attention in the convolution block attention module (CBAM). Hou et al. [25] further advanced CBAM by introducing coordinate attention (CoorAttn), which utilizes channelwise average pooling to generate attention maps. Reverse attention is another practical method to mine boundary cues. PraNet [18] extracted fine-grained details by removing the estimated polyp regions using boundary information. Lou et al. [19] enhanced PraNet by decomposing attention maps along height and width axes. Zhou et al. [37] employed channel separation and pooling to adjust the sizes of feature maps for spatial and channel attention; their use of fixed-size attention maps constrained the diversity of attention mechanisms. Moreover, relatively small feature maps (ranging from 1×1 to 44×44) were employed to bridge area and boundary cues, which

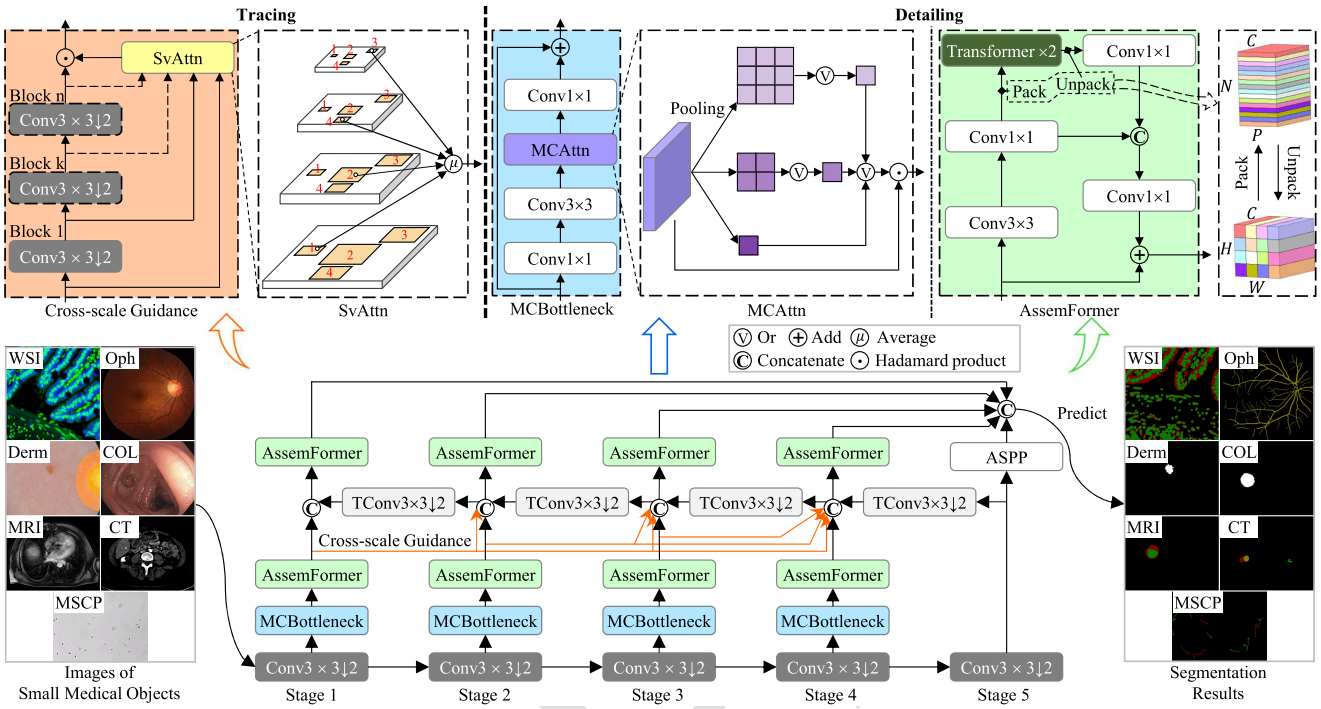

Fig. 2. Architecture of SvANet. Cross-scale guidance and SvAttn techniques, depicted in the top-left dashed boxes, integrate low-level and high-level feature maps to trace the alterations in the shape and location of small medical objects. The modules MCAttn and MCBottleneck, positioned in the top-middle dashed boxes, along with AssemFormer in the top-right dashed boxes, synergistically correlate local and global features to capture intricate object details.

may not adequately capture the structural details of minuscule objects [14], [18], [19], [23], [37].

Moreover, self-attention is an effective attention scheme to obtain dependencies and relationships within input data. Based on self-attention mechanisms, ViT has been introduced to process sequences of image patches to learn the inter-patch representations, which has shown noticeable potential in aggregating and preserving the features of small objects [14]. He et al. [38] proposed a fully transformer-based network that amalgamated spatial pyramid theory and ViT to identify skin lesions. However, the vanilla ViT lacks inherent bias and is susceptible to perturbations [39]. Zhang et al. [27] and Pan et al. [13] employed self-attention to improve the feature correlations in their convolutional neural network (CNN)-based network for polyps and nuclei examination, respectively. To capture long-range information when segmenting cell nuclei, Hörst et al. [40] replaced the CNN encoder with a transformer block in the U-Net architecture. Du et al. [15] incorporated shift-window techniques and a multiscale attention module into a U-shaped architecture to enhance the recognition of polyps and skin lesions. To leverage cross-scale features and improve the capture of contextual connections, Wu et al. [21] embedded feature maps from four stages and further processed them with self-attention modules. However, the aforementioned research overlooks the effect of ViT on the analysis of small medical objects.

III. Methodology

A. Overall Framework

This section introduces the scale-variant attention-based network (SvANet), specifically designed to segment small medical objects. The SvANet model, schematically depicted in Fig. 2, comprises four main components: cross-scale guidance in Section III-B, SvAttn in Section III-C, MCAttn in Section III-D, and the convolution with ViT in Section III-E.

Preserving the features of tiny medical objects, such as sperms and retinal vessels, becomes challenging after multiple pooling or strided convolutional operations. For example, after two strided convolutions, a sperm may be reduced to being represented by only one or two pixels in the image. In this study, cross-scale feature maps are applied to guide the latter stages in learning the features of small medical objects, as indicated by the orange arrows in Fig. 2. The SvAttn and cross-scale guidance are primarily designed to track feature changes, particularly downsizing. Meanwhile, MCAttn and AssemFormer distill multiscale attention maps for improved contextual feature learning. To better comprehend the roles of cross-scale guidance, SvAttn, MCAttn, and AssemFormer in small medical object segmentation, we examined the feature maps, as shown in Figs. 3–5. For simplicity, we selected FIVES, ISIC 2018, KiTS23, and SpermHealth datasets to visualize feature maps. We chose to present the outputs from two cross-scale guidance correlations (see Fig. 3) and the MCAttn-based bottleneck (MCBottleneck) in stage four (see Fig. 4).

In addition, in Fig. 2, each Conv3 $\times$ 3 $\downarrow$ 2, represented by black blocks, contains a single 3×3 convolution with a stride of 2 (strided convolution). Every TConv3 $\times$ 3 $\downarrow$ 2, denoted by gray blocks, consists of three convolution units: a 1×1 convolution, a 3×3 transposed strided convolution, and a 1×1 convolution. The MCBottleneck serves as a compression point in the network, narrowing the tensor channels before

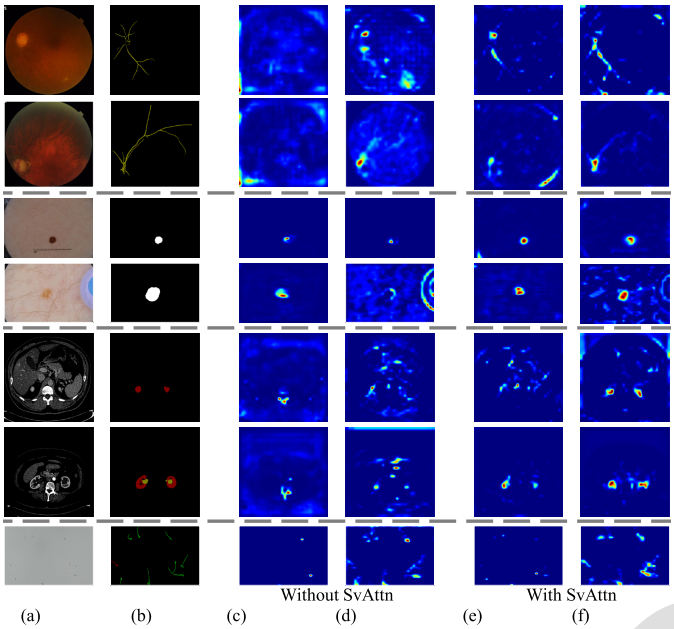

Fig. 3. Output feature maps from cross-scale guidance without or with SvAttn. These feature maps are generated from $g(x_s, t)$, which integrates features at different scales. Example images in odd and even rows include ultrasmall and small medical objects, respectively (GT: ground truth). (a) Input. (b) GT. (c) $s = 1$ and $t = 4$. (d) $s = 3$ and $t = 4$. (e) $s = 1$ and $t = 4$. (f) $s = 3$ and $t = 4$.

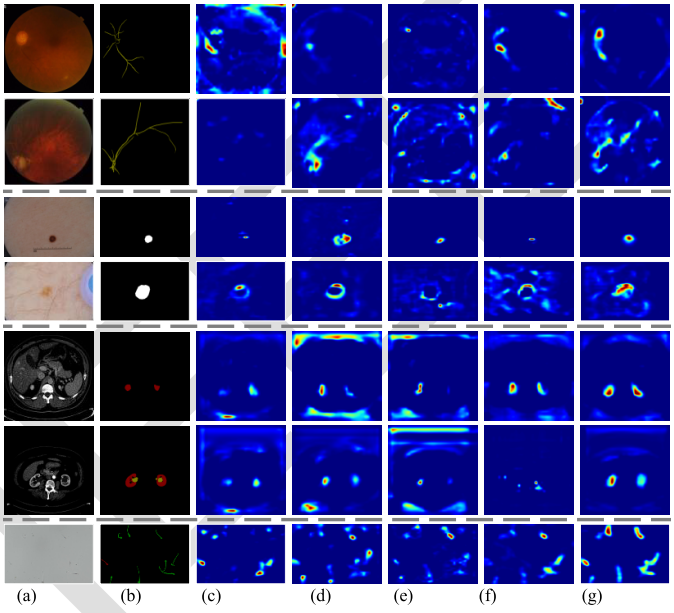

Fig. 4. Output feature maps from the MCBottleneck without or with an attention mechanism. Example images in odd and even rows include ultrasmall and small medical objects, respectively (GT: ground truth). (a) Input. (b) GT. (c) Vanilla. (d) SE. (e) CBAM. (f) CoordAttn. (g) MCAttn.

expanding them to extract salient features by compressing the input information, resembling a "bottleneck" in information theory. To expand the receptive field and capture features at multiple scales, atrous spatial pyramid pooling (ASPP) [31] is integrated after the final stage of our model.

Merely classifying objects based on their category [16], [35], [36] or pixel count [9] does not accurately describe size.

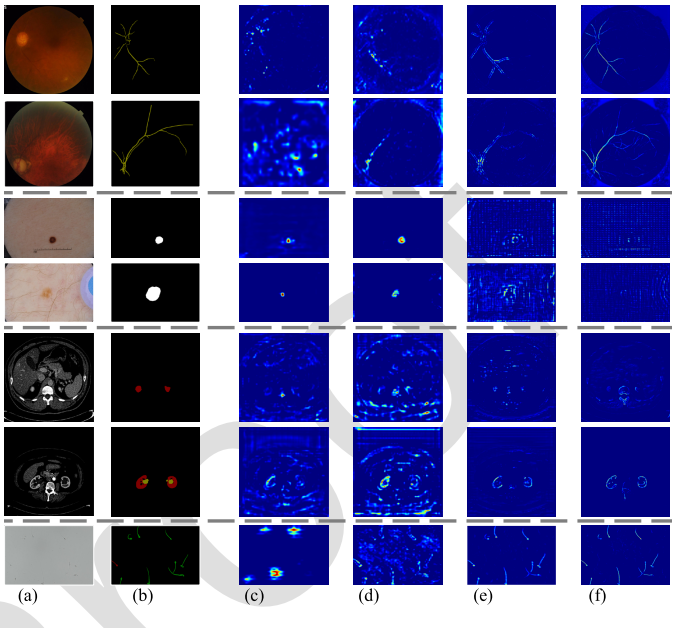

Fig. 5. Output feature maps from the AssemFormer. The feature maps are extracted from four encoder stages individually (GT: ground truth). Since the layer in the fifth stage is directly connected with the ASPP module, no AssemFormer is used at this stage. (a) Input. (b) GT. (c) Stage 1. (d) Stage 2. (e) Stage 3. (f) Stage 4.

This study defines "ultrasmall scale" medical objects as those with an area ratio below 1%, and "small scale" as those below 10% for precise measurement of object sizes.

B. Cross-Scale Feature Guidance

The information content decreases significantly as the size of the medical object reduces, owing to compression artifacts in neural networks. This study introduces a cross-scale guidance module to leverage the higher resolution features from earlier model stages. Assume that t is the target stage, the output y_t can be computed as follows:

$$y_t = \sum_{s=1}^{t-1} g\left(x_s, t\right) \tag{1}$$

where x_s represents the input tensor in stage $s = 1, 2, \ldots, t-1$ and the transformation $g(x_s, t)$ involves $(t - s)$ 3×3 strided convolutions. The function is depicted by the orange arrows and the top-left orange blocks in Fig. 2.

As illustrated in Fig. 3(c) and (d) or (e) and (f), the highlighted region expands as the input stage increases for the same target stage, $t = 4$. This expansion occurs due to an increased total number of strided convolutional operations performed on the data.

C. Scale-Variant Attention

Cross-scale feature guidance is based on convolutional operations, which have inherent limitations in processing global feature representations. While global pooling operations can facilitate learning context representations, it is restricted to handling features uniformly. Given a subregion x_j of an input

tensor x, the output of vanilla global attention, denoted by $\mathcal{A}(x)$, is calculated as follows:

$$\mathcal{A}(x) = \frac{1}{\sigma(x)} \sum_{j=1}^{n} x_j \tag{2}$$

where x_j represents the r^2 neighborhood centered at the jth subregion of x. Here, n denotes the total number of subregions, and $\sigma(x)$ represents the scalar function that normalizes the result. For vanilla global attention, the default values are set as $r = 1$ and $n = \sigma(x) = H_x \times W_x$.

Conventional global attention, as described in (2), fails to capture relationships across subregions and is limited to computing a single-scale size of the feature. To overcome this scale limitation while maintaining long-range correlations, we introduce SvAttn, which processes global dependencies across diverse scales, as depicted by the yellow block in Fig. 2. In SvAttn, multiscale attention maps are calculated across input stages $s = 1, 2, \ldots, t-1$. Assuming that the groupwise correspondence among input tensors is controlled by a probability $P_1(x)$, the output attention map of SvAttn is defined as follows:

$$\mathcal{A}_t(\mathbf{x}) = \frac{1}{\sigma(\mathbf{x})} \sum_{j=1}^{n} \sum_{s=1}^{t-1} P_1\left(x_{s,j}\right) x_{s,j} \tag{3}$$

where $x_{s,j}$ denotes the jth subregion of the input tensor at the sth stage, t is the target stage, and $\mathbf{x} = [x_1, x_2, \ldots, x_{t-1}]^{-1}$ represents the vector of input tensors across various stages. In addition, for the jth subregion, a single input stage is randomly chosen with equal probability across all stages to compute the attention map. For example, if $P_1(x_1, j) = 1$, then $P_1(x_s, j) = 0$ where $s \neq 1$. Therefore, the correspondence probability $P_1(x)$ satisfies the conditions $\sum_{s=1}^{t-1} P_1(x_s, j) = 1$ and $\prod_{s=1}^{t-1} P_1(x_s, j) = 0$, thereby ensuring a weighted sum of attention maps across different scales. Since subregion sampling is realized by masking tensors with random masks, it does not increase model size. The scalar function $\sigma(\mathbf{x})$ is defined by

$$\sigma(\mathbf{x}) = n = \frac{H \vee W}{2^{t+1}} \tag{4}$$

where H and W are the height and width of the input image, respectively. The symbol $\vee$ denotes the logical OR operation.

In conjunction with (1) and (3), the output tensor y'_t of cross-scale guidance using SvAttn can be defined as follows:

$$y'_t = \mathcal{A}_t(\mathbf{x})\, y_t. \tag{5}$$

As indicated by (3), the subregions located at the same proportional scaling position across stages are dynamic. This variability enables the cross-scale guidance module to effectively discern the relationships between the high-level and low-level features. Consequently, SvAttn enhances the network's capability to recognize downsized small medical objects throughout a sequence of stages. As illustrated in Fig. 3(c) and (e) and (d) and (f), for the same source and target stages, the features captured using SvAttn are more detailed and comprehensive for both ultrasmall and small medical objects compared with those obtained without using SvAttn. For example, from top to bottom, there is a more precise delineation of networked retinal vessels, more discernible morphology of nevi, more pronounced instance boundaries of organs such as kidneys, and finer details in sperm morphology. In contrast, without using SvAttn, critical features such as retinal vessels of glaucoma in the first and second rows, the nevus in the third row, and the kidneys and cyst in the sixth row were overlooked. It is noteworthy that ultrasmall objects are harder to perceive compared to small objects without using SvAttn. For example, moving downward from the odd rows of Fig. 3(c) and (d), no retinal vessel was discovered, a relatively small nevus region was highlighted, and the left kidney was missed.

D. Monte Carlo Attention

The MCAttn module, as presented by the purple block in Fig. 2, uses a random-sampling-based pooling operation to generate scale-agnostic attention maps, enabling the network to capture relevant information across different scales, enhancing its ability to identify small medical objects. The MCAttn generates attention maps by randomly selecting a 1×1 attention map from three scales: 3×3, 2×2, and 1×1 (pooled tensors). In conventional methods such as SE, global-average pooling is used to acquire a 1×1 output tensor, which helps calibrate the interdependencies between channels [23]. However, this approach has limited capacity to exploit cross-scale correlations. To address this limitation, MCAttn calculates the attention maps from features across three scales, thereby enhancing long-range semantic interdependencies. Given an input tensor, x, the output attention map of MCAttn, denoted by $\mathcal{A}_m(x)$, is computed as follows:

$$\mathcal{A}_m(x) = \sum_{i=1}^{n} P_2(x, i)\, f(x, i) \tag{6}$$

where i denotes the output size of the attention map, and $f(x, i)$ represents the average pooling function. Similar to (3), the association probability $P_2(x, i)$ satisfies the conditions $\sum_{i=1}^{n} P_2(x, i) = 1$ and $\prod_{i=1}^{n} P_2(x, i) = 0$, ensuring the generation of agnostic and generalizable attention maps. For the input tensor x, a single pool size is randomly selected from all available options, each with equal probability. n represents the number of output pooled tensors and is set to 3 in this study.

The Monte Carlo sampling method described in (6) allows for the random selection of association probabilities, enabling the extraction of both local information (e.g., angle, edge, and color) and context information (e.g., whole image texture, spatial correlation, and color distribution). In Fig. 4(c), the second to the fourth rows and the final row illustrate that MCBottleneck, without using an attention mechanism, struggles to detect the retinas and nevi and often overlooks several sperms. Conversely, when attention mechanisms like SE, CBAM, and CoordAttn are used, localization of densely occupied regions (e.g., optic disk, kidneys, and sperms) is enhanced compared to when no attention mechanism is used. However, sparse regions, such as retinal vasculatures and nevus centers, are often overlooked, especially the ultrasmall ones, as shown in Fig. 4(c)–(f). Instead, using MCAttn in

MCBottleneck, as depicted in Fig. 4(c) and (g), enhances the discernibility of the morphology and precise location of both ultrasmall and small medical objects compared with when MCAttn is not used. For instance, in Fig. 4, moving downward, MCBottleneck coupled with MCAttn emphasizes more apparent retinal vessels for glaucoma, sharper boundaries of nevi, and more perceptible morphology of kidneys, cysts, and sperms. MCAttn also accentuates other medical objects of interest, such as retinas, nevi, kidneys, and sperms, as shown in Fig. 4(g).

E. Convolution With ViT

The proposed AssemFormer is illustrated in the top-right dashed green boxes in Fig. 2. Inspired by [14] and [41], AssemFormer incorporates a 3×3 convolution and a 1×1 convolution, followed by two transformer blocks and two convolutional operations. AssemFormer bridges convolution and transformer operations by stacking and unstacking feature maps. Equipped with this design, AssemFormer tackles the lack of inductive biases for the vanilla transformer.

The functionalities of convolution and transformer operations differ. Convolutional operations focus on learning local and general features, such as corners, edges, angles, and colors of medical objects. In contrast, the transformer module extracts global information, including morphology, depth, and color distribution of medical objects, utilizing multihead self-attention (MHSA). In addition, the transformer module also learns positional associations of medical objects, such as the relationships between a tumor and the kidney, a kidney and the abdomen, and a tumor and the abdomen within an MRI slice image. The ViT algorithm employs a sequence of MHSA and multilayer perceptron (MLP) blocks, each followed by layer normalization [39]. The self-attention mechanism [42] is formulated as follows:

$$\mathcal{A}_{\text{ViT}}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \text{softmax}\left(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{D_h}}\right)\mathbf{v} \tag{7}$$

where $\mathbf{q}$, $\mathbf{k}$, and $\mathbf{v}$ are the query, key, and value vectors of an input sequence $\mathbf{z} \in \mathbb{R}^{N \times D}$, respectivley. N denotes the number of patches, and D represents the patch size. Given m self-attention operations, D_h, the dimension of $\mathbf{q}$ and $\mathbf{k}$, is defined as D/m.

Furthermore, a skip connection and concatenation are incorporated to mitigate the information loss concerning small medical objects. Leveraging the convolution-transformer hybrid structure, the AssemFormer block can simultaneously learn the local and global representation of an input medical image. According to the ablation study presented in Section IV-D2, the AssemFormer significantly improves the segmentation performance of SvANet.

In Fig. 5, progressing from left to right, the AssemFormer increasingly highlights smaller areas that more accurately align with the ground truth (GT), especially notable in scenarios with fewer medical objects. For instance, the first row of Fig. 5 demonstrates how the thin lines of retinal vasculature and light reflections are initially emphasized, becoming progressively thicker. Subsequently, these lines become shorter and more focused on a smaller region corresponding to the optic disk location, as depicted in the first two rows of Fig. 5(e) and (f). Large-scale distortions, such as noise or compression defects, play a role in this observed trend, where the concentration of feature maps intensifies with a deeper layer. The pattern of increased feature map concentration is consistent across the segmentation of various medical objects, including skin lesions, polyps, hepatic tumors, livers, kidneys, tissue cells, and sperm.

The MHSA mechanism of AssemFormer, described in (7), facilitates patch interactions and enriches the context information. In contrast to Fig. 4, from left to right, the feature maps evolve from AssemFormer from coarse to fine representations. As illustrated in Fig. 5(f), the AssemFormer enhances the visibility and precise localization of small medical objects such as glaucoma, nevus, polyp, hepatic tumor, kidney, tissue cell, and sperm, highlighting their morphological details and exact positions.

IV. Experimental Results

A. Evaluation Protocol

1) Dataset: To validate the effectiveness of SvANet, we conducted tests alongside seven state-of-the-art (SOTA) models for small medical object segmentation across seven benchmark datasets: FIVES [1], ISIC 2018 [2], PolypGen [3], ATLAS [4], KiTS23 [5], TissueNet [6], and SpermHealth.

The FIVES dataset comprises 800 fundus photographs taken with ophthalmoscopes featuring age-related macular degeneration, diabetic retinopathy, glaucoma, and healthy fundus types. The ISIC 2018 dataset includes skin lesion images collected by dermatoscopes, encompassing healthy and unhealthy skin areas. PolypGen, sourced from six different hospitals using colonoscopes, focuses exclusively on polyps. The ATLAS dataset consists of 90 MRI scans of livers, detailing two types of medical objects: the liver and the tumor. KiTS23 has 599 CT scans of kidneys, categorized into three semantic classes (i.e., kidney, tumor, and cyst). For experimental comparisons, the ATLAS and KiTS23 datasets were converted into 2-D image sequences. In addition, TissueNet includes images of cells from the pancreas, breast, tonsil, colon, lymph, lung, esophagus, skin, and spleen, derived from humans, mice, and macaques, collected using cell imaging platforms such as CODEX and CyCIF, with annotations for whole cells and their nuclei.

SpermHealth is a customized dataset from the 3rd Affiliated Hospital of Shenzhen University, consisting of low-resolution sperm images (640×480 and 96 DPI) extracted from microscope-captured videos. These images have been meticulously annotated into normal and abnormal categories by experienced fertility doctors. Further details of the datasets used in the tests are presented in Table I.

2) Implementation Details and Evaluation Metric: In this study, the mini-batch size was set to 4. Data augmentation strategies applied to preprocess the input images included random horizontal flips, random cropping to a resolution of 512×512, Gaussian blur, distortion, and rotation. The AdamW optimizer [43] and a cross-entropy loss function were utilized, with the learning rate decaying from 5×10^{-5} to 1×10^{-6} following a cosine schedule [44]. The total training process

spanned 100 epochs. The results were calculated by averaging the outcomes from three times of training and testing cycles. All backbone was pretrained in the ImageNet-1K [45] dataset. In addition, all tested methods followed the configurations above of training, except that nnUNet utilized official settings [26] for training.

The experiments were conducted on an RTX 4090 GPU with an AMD Ryzen 9 7950X CPU. The metrics used to assess the performance of semantic segmentation include the mean Dice (mDice) coefficient, mean intersection over union (mIoU), and mean absolute error (MAE). Given the critical role of sensitivity in medical diagnosis for identifying infected patients among all subjects and facilitating timely treatment, we also incorporated sensitivity and $F2$ score as key performance metrics.

B. Results for Datasets With Diverse Object Sizes

The experimental results for the FIVES, ISIC 2018, PolypGen, ATLAS, KiTS23, and TissueNet datasets are summarized in Table II. These results demonstrate that SvANet outperforms other SOTA methods across all metrics for ultrasmall and small medical object segmentation across six datasets tested.

As presented in Table II, SvANet outperformed other SOTA methods across three object scales in the FIVES, ISIC 2018, and ATLAS datasets, excluding sensitivity of 93.54% and 87.13% in ISIC 2018 and ATLAS datasets and MAE of 5.35×10^{-4} and 6.6×10^{-3} in FIVES and ATLAS datasets for the "all" object scale, as summarized in Table II. In addition, SvANet surpasses other methods with increments in the mDice of at least + 2.95% and + 5.23%, mIoU of + 1.97% and + 5.78%, sensitivity of + 0.19% and + 5.03%, and $F2$ score of + 1.28% and + 5.15% for differentiating ultrasmall and small retinal vessels in the FIVES dataset. However, MAE is comparatively high ($> 7.5 \times 10^{-3}$) in ultrasmall retinal vasculature segmentation across all tested models, potentially due to the minimal number (4) of ultrasmall objects providing insufficient learnable features for deep learning algorithms. In ISIC 2018 and ATLAS datasets, SvANet excelled in segmenting ultrasmall objects (i.e., skin lesions, livers, and hepatic tumors) with mDice of 96.11% and 89.79%, mIoU of 92.76% and 86.06%, sensitivity of 98.35% and 86.68%, and $F2$ score of 97.42% and 87.71%. These results suggest significant potential for SvANet in diagnosing dermatological skin lesions and hepatic tumors in MRI scans, particularly for objects with an area ratio smaller than 1% or 10%. Thus, SvANet can ameliorate therapeutic approaches such as excision therapy, laser therapy, electrosurgery, and radiotherapy for treating these conditions.

Furthermore, the segmentation results for the PolypGen and KiTS23 datasets demonstrate that SvANet delivers superior performance than other SOTA methods across three object scales. Specifically, SvANet achieved the highest mDice of 84.15% and 96.12%, 91.17% and 94.01%, and 93.16% and 94.54% for ultrasmall, small, and all medical object scales in PolypGen and KiTS23 datasets, respectively. Moreover, SvANet delivered up to 14.83% and 2.76%, 6.23% and 6.88%, and 6.93% and 6.33% increments in $F2$ score over other tested methods for ultrasmall, small, and all object scales in PolypGen and KiTS23 datasets, respectively. The $F2$ score, the harmonic mean of sensitivity and precision, underscores the robustness of SvANet in medical object segmentation. SvANet also recorded the lowest MAE, 1.01×10^{-4} and 2.0×10^{-2}, 6.6×10^{-3} and 7.0×10^{-2}, and 8.1×10^{-3} and 8.0×10^{-2} across three object scales for PolypGen and KiTS23 datasets, indicating a high level of precision in the pixel-level recognition of polyps, kidneys, renal tumors, and cysts.

In the TissueNet dataset, which includes only ultrasmall and small cells, Table II reveals that the SvANet leads in segmentation performance, achieving 80.25% and 88.05% in mDice, 71.60% and 79.45% in mIoU, 7.22×10^{-4} and 3.28×10^{-4} in MAE, 83.36% and 88.07% in sensitivity, and 82.00% and 88.06% in $F2$ score, across ultrasmall and small medical object scales, respectively. Notably, SvANet performance is essentially distinguished in the segmentation of ultrasmall tissue cells, surpassing other SOTA models by at least + 9.60% in mIoU, + 8.50% in mDice, and + 6.61% in $F2$ score. This superior performance contrasts with improvements of less than 5% observed in the five other datasets, as shown in Table II, which may be attributed to the relatively large number of ultrasmall objects in TissueNet (i.e., 9096 cells, approximately ten times more objects than other datasets).

Furthermore, the mDice results trends for all tested methods across ultrasmall, small, and all medical object segmentation in FIVES, ISIC 2018, PolypGen, ATLAS, KiTS23, and TissueNet datasets are illustrated in Fig. 6. This figure highlights that the SvANet, represented by the red line, consistently leads across diverse object scales and datasets.

In the FIVES dataset, as shown in Fig. 6(a), only SvANet exhibits an increasing mDice as object scale increases, while other methods' mDice initially increases and then decreases. The subbranches of retinal vessels are relatively thin, and the number of vessels increases as the occupied area expands. Therefore, the subbranches become more difficult to discriminate, decreasing mDice as the object scale range expands from $\leq$ 10% to $\leq$ 100%. However, SvANet maintains a growing trend without decline, demonstrating its effectiveness in recognizing retinal vasculatures, which is crucial for diagnosing blindness-causing diseases. In addition, in ISIC 2018 and KiTS23 datasets, SvANet and over half of the other methods exhibit a mDice trend resembling an "L" shape, as depicted in Fig. 6. Fewer ultrasmall objects in these datasets introduce significant variability, likely contributing to this "L" trend. In the PolypGen, ATLAS, and TissueNet datasets, there is a consistent increase in mDice trends, as shown in Fig. 6(c), (d), and (f). Notably, no change is observed in TissueNet between the small and all object scales, as both categories contain identical medical images. Closer inspection of Fig. 6(d) and Table II reveals that SvANet is the only method that achieved a "V" trend in the ATLAS dataset, with the best mDice of 89.79% for segmenting ultrasmall compared to small and all sizes of livers and tumors, underscoring SvANet's capability to effectively discriminate ultrasmall medical objects.

In addition, U-Net, U-Net++, nnUNet, CFANet, and TransNetR obtained mean standard error (mSE) values

TABLE II

QUANTITATIVE RESULTS IN FIVES, ISIC 2018, POLYPGEN, ATLAS, KITS23, AND TISSUENET DATASETS, DIVIDED BY AREA RATIOS OF MEDICAL OBJECTS: BELOW 1% (ULTRASMALL), BELOW 10% (SMALL), AND 100% (ALL). THE BEST RESULTS ARE UNDERLINED IN BOLD

	Methods	mDice			mIoU			MAE ($\times 10^{-4}$)			Sensitivity			F2 score		
		ultra small	small	all	ultra small	small	all	ultra small	small	all	ultra small	small	all	ultra small	small	all
FIVES	UNet (MICCAI'15) [10]	67.49	72.99	71.12	62.54	65.04	63.10	84.65	14.64	6.22	63.86	71.48	70.13	65.06	72.02	70.42
	UNet++ (TMI'19) [11]	70.10	74.46	70.72	65.06	66.37	62.68	80.64	14.46	5.92	67.02	73.05	69.50	68.08	73.58	69.90
	HRNet (TPAMI'20) [34]	69.46	79.16	75.66	64.67	70.83	67.04	88.84	15.38	5.51	68.33	78.30	74.09	68.68	78.60	74.67
	PraNet (MICCAI'20) [18]	64.47	71.46	68.57	59.42	63.08	60.14	79.64	14.78	5.45	61.20	67.88	65.61	62.27	69.11	66.58
	nnUNet (NM'21) [26]	58.02	74.34	71.73	54.16	66.27	63.59	91.39	14.70	5.90	54.80	73.39	71.03	55.68	73.72	71.27
	CFANet (PR'23) [12]	69.35	76.64	72.23	63.16	68.60	65.27	80.18	15.16	**4.92**	64.86	75.67	73.79	66.29	75.99	72.75
	TransNetR (MIDL'24) [29]	67.87	80.68	78.60	63.59	72.55	69.66	83.54	14.44	5.68	66.90	78.83	77.24	67.26	79.45	77.60
	SvANet (Ours)	**73.05**	**85.91**	**86.29**	**67.03**	**78.33**	**78.39**	**76.82**	**14.42**	5.35	**68.52**	**83.86**	**85.15**	**69.96**	**84.60**	**85.46**
ISIC 2018	UNet (MICCAI'15) [10]	88.00	89.77	90.87	80.98	82.59	83.55	47.51	11.22	7.13	98.23	93.67	90.61	93.35	92.00	90.71
	UNet++ (TMI'19) [11]	81.10	88.80	90.73	72.61	81.20	83.30	93.65	12.59	7.35	98.29	93.64	91.06	89.12	91.54	90.92
	HRNet (TPAMI'20) [34]	88.32	89.83	91.84	81.11	82.67	85.13	43.62	11.53	6.44	97.94	95.23	92.02	93.47	92.87	91.95
	PraNet (MICCAI'20) [18]	95.04	90.66	93.03	90.96	83.90	87.14	15.01	10.46	5.55	97.01	95.56	**93.64**	96.16	93.44	93.39
	nnUNet (NM'21) [26]	89.29	89.74	91.01	82.90	82.60	83.78	41.96	11.47	7.02	98.35	94.19	90.74	94.02	92.26	90.84
	CFANet (PR'23) [12]	94.09	90.34	92.89	89.51	83.41	86.89	18.64	10.93	5.63	97.08	95.68	93.23	95.82	93.35	93.09
	TransNetR (MIDL'24) [29]	88.73	90.43	92.35	82.07	83.56	85.99	42.82	10.69	6.01	96.67	95.21	92.38	92.92	93.14	92.36
	SvANet (Ours)	**96.11**	**91.63**	**93.24**	**92.76**	**85.36**	**87.50**	**11.90**	**9.18**	**5.35**	**98.35**	**95.71**	93.54	**97.42**	**93.96**	**93.42**
PolypGen	UNet (MICCAI'15) [10]	73.40	84.81	87.14	70.96	76.50	78.85	2.94	1.13	1.45	79.18	84.09	84.06	75.58	84.36	85.19
	UNet++ (TMI'19) [11]	74.87	85.79	88.43	71.86	77.67	80.60	2.81	1.07	1.33	82.84	85.44	85.85	77.89	85.58	86.82
	HRNet (TPAMI'20) [34]	70.58	85.34	89.32	68.88	77.10	81.85	5.72	1.14	1.26	76.36	85.77	87.38	71.93	85.58	88.12
	PraNet (MICCAI'20) [18]	81.11	90.69	92.60	76.41	84.22	86.83	1.34	0.68	0.88	86.99	89.03	90.71	83.97	89.67	91.44
	nnUNet (NM'21) [26]	78.52	87.94	89.33	77.50	80.41	81.87	4.06	0.94	1.24	84.95	88.52	87.04	79.52	88.29	87.91
	CFANet (PR'23) [12]	79.44	90.65	92.71	75.08	84.16	87.00	1.75	0.70	0.86	87.76	88.79	90.70	83.16	89.51	91.47
	TransNetR (MIDL'24) [29]	79.51	90.67	92.49	75.16	84.19	86.65	1.80	0.70	0.89	88.08	90.04	90.68	83.29	90.29	91.37
	SvANet (Ours)	**84.15**	**91.17**	**93.16**	**78.95**	**84.90**	**87.71**	**1.01**	**0.66**	**0.81**	**89.21**	**90.21**	**91.47**	**86.76**	**90.59**	**92.12**
ATLAS	UNet (MICCAI'15) [10]	82.09	83.55	85.45	79.89	76.60	77.89	0.41	0.75	0.88	81.98	81.24	83.40	81.95	82.05	84.12
	UNet++ (TMI'19) [11]	81.70	83.91	84.75	79.58	76.98	77.33	0.47	0.73	0.84	82.59	82.33	83.00	82.17	82.86	83.57
	HRNet (TPAMI'20) [34]	85.86	84.98	86.66	82.56	78.53	79.68	0.26	0.56	**0.65**	84.74	83.70	85.12	85.14	84.09	85.59
	PraNet (MICCAI'20) [18]	86.04	86.70	88.02	82.69	80.00	81.12	0.30	0.63	0.76	85.39	85.22	86.80	85.62	85.77	87.25
	nnUNet (NM'21) [26]	86.22	85.01	85.65	85.79	77.93	77.97	0.23	0.82	1.01	86.46	83.74	84.83	86.35	84.19	85.13
	CFANet (PR'23) [12]	86.24	87.04	88.25	82.96	80.52	81.46	0.26	0.57	0.70	84.89	85.51	86.44	85.34	86.03	87.06
	TransNetR (MIDL'24) [29]	86.28	86.53	88.69	82.93	80.37	82.05	0.27	0.51	0.66	85.29	85.42	**87.15**	85.61	85.75	87.68
	SvANet (Ours)	**89.79**	**87.60**	**89.14**	**86.06**	**81.29**	**82.56**	**0.16**	**0.51**	0.66	**86.68**	**85.86**	87.13	**87.71**	**86.43**	**87.82**
KiTS23	UNet (MICCAI'15) [10]	95.28	91.27	91.94	93.42	87.58	88.30	0.03	0.07	0.09	96.21	91.08	91.54	95.76	91.11	91.65
	UNet++ (TMI'19) [11]	95.38	93.94	93.73	93.50	89.14	89.86	0.04	0.08	0.09	96.40	92.79	93.28	95.90	92.89	93.45
	HRNet (TPAMI'20) [34]	96.00	93.48	93.98	94.25	89.72	90.35	0.02	0.07	0.09	96.16	93.08	93.52	96.09	93.21	93.67
	PraNet (MICCAI'20) [18]	95.58	92.84	93.39	93.67	88.71	89.34	0.03	0.08	0.10	96.25	92.38	92.84	95.95	92.53	93.03
	nnUNet (NM'21) [26]	93.05	87.26	88.46	91.12	83.77	84.68	0.03	0.07	0.09	94.26	86.76	87.73	93.74	86.92	87.98
	CFANet (PR'23) [12]	95.33	93.66	94.14	93.57	89.80	90.39	0.02	0.07	0.09	96.10	93.15	93.60	95.76	93.32	93.78
	TransNetR (MIDL'24) [29]	95.82	93.12	93.76	94.20	89.51	90.31	0.02	0.07	0.08	96.65	92.79	93.42	96.30	92.88	93.52
	SvANet (Ours)	**96.12**	**94.01**	**94.54**	**94.51**	**90.38**	**91.05**	**0.02**	**0.07**	**0.08**	**96.81**	**93.70**	**94.20**	**96.50**	**93.80**	**94.31**
TissueNet	UNet (MICCAI'15) [10]	65.89	86.36		56.03	76.99		28.26	3.34		80.69	86.29		71.29	86.32	
	UNet++ (TMI'19) [11]	64.43	85.89		52.73	76.25		41.76	3.39		78.36	85.89		67.19	85.88	
	HRNet (TPAMI'20) [34]	64.66	86.99		54.96	77.84		31.61	3.35		78.18	86.92		69.15	86.94	
	PraNet (MICCAI'20) [18]	60.30	85.96		50.81	76.36		47.43	3.37		77.41	85.98		64.84	85.97	
	nnUNet (NM'21) [26]	61.35	86.65		55.36	77.38		45.41	3.33		82.53	86.74		66.39	86.70	
	CFANet (PR'23) [12]	71.75	87.48		62.00	78.59		16.15	3.31		80.20	87.43		75.39	87.45	
	TransNetR (MIDL'24) [29]	65.74	86.85		56.07	77.69		24.28	3.34		61.70	86.94		71.46	86.90	
	SvANet (Ours)	**80.25**	**88.05**		**71.60**	**79.45**		**7.22**	**3.28**		**83.36**	**88.07**		**82.00**	**88.06**	

exceeding 1 in one or two of FIVES, ISIC 2018, and TissueNet datasets, suggesting potential instability of these methods. In contrast, all tested methods in the PolypGen, ATLAS, and KiTS23 datasets exhibited mSE values lower than 1, demonstrating consistent segmentation performance across these three datasets. Moreover, SvANet consistently achieves

TABLE III

QUANTITATIVE ANALYSIS OF THE SPERMHEALTH DATASET HIGHLIGHTING MODEL SIZE AND INFERENCE TIME COMPARISONS. ALL SPERMS IN THIS DATASET OCCUPY LESS THAN 1% OF THE IMAGES' AREA. THE BEST RESULTS ARE UNDERLINED IN BOLD

	Methods	# Parameters /Million	MACs /Billion	Speed /FPS	mDice/%	mIoU/%	MAE/$\times 10^{-4}$	Sensitivity/%	F2 score/%
SpermHealth	UNet (MICCAI'15) [10]	34.53	262.21	93	58.47	50.18	13.54	57.52	57.65
	UNet++ (TMI'19) [11]	9.16	139.68	100	57.94	49.58	15.09	56.15	56.63
	PraNet (MICCAI'20) [18]	30.34	25.65	82	60.56	50.82	16.59	57.51	58.61
	HRNet (TPAMI'20) [34]	63.60	65.80	78	64.25	54.01	14.65	62.68	63.23
	nnUNet (NM'21) [26]	30.60	232.19	83	65.74	55.28	13.43	67.51	66.77
	CFANet (PR'23) [12]	25.71	115.63	56	60.58	50.88	20.01	59.13	59.63
	TransNetR (MIDL'24) [29]	27.27	44.71	100	70.28	59.19	14.66	69.70	69.89
	LiteSvANet (Ours)	53.04	224.18	77	70.88	59.84	14.33	69.96	70.23
	SvANet (Ours)	177.64	312.76	55	**72.58**	**61.44**	**13.06**	**72.50**	**72.51**

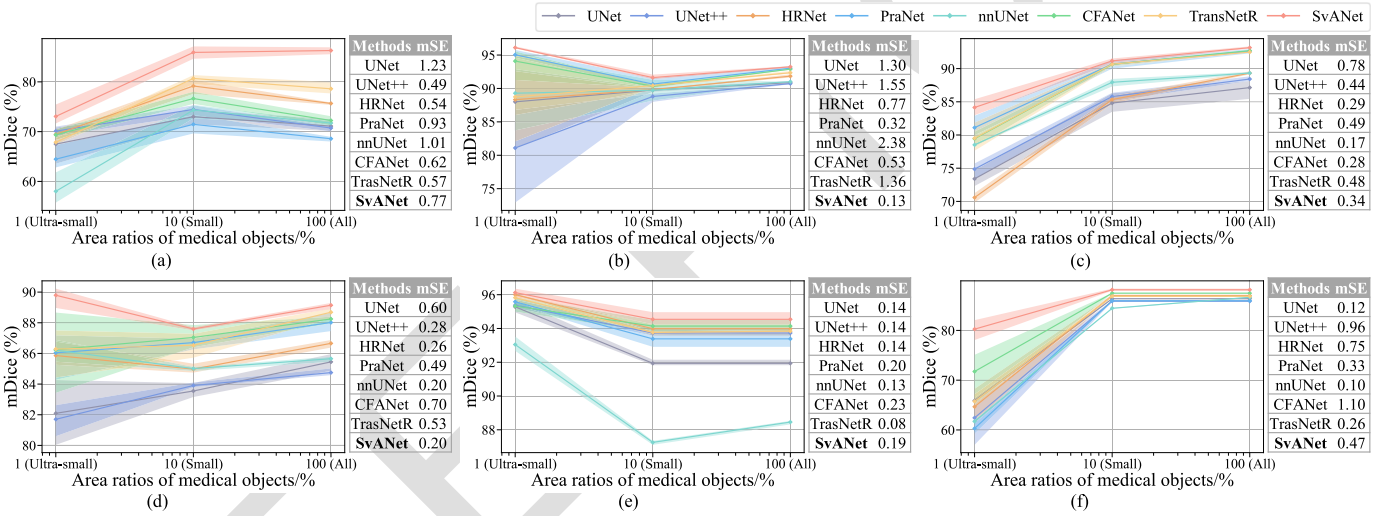

Fig. 6. Segmentation mDice across different area ratios of medical objects in (a) FIVES, (b) ISIC 2018, (c) PolypGen, (d) ATLAS, (e) KiTS23, and (f) TissueNet datasets. mSE refers to the mean standard error across three object scales.

mSEs of less than 0.8 across all datasets and displays narrow error bars (shown as color bands) across three object scales, indicating its robustness in accurately recognizing medical objects.

C. Results for the Dataset With Only Ultrasmall Objects

To further evaluate the performance of SvANet in distinguishing ultrasmall medical objects, experiments were conducted in the SpermHealth dataset, which exclusively has sperms with an area ratio of less than 1%. As shown in Table III, SvANet secured top performance in sperm segmentation within the SpermHealth dataset, achieving 72.58% in mDice, 61.44% in mIoU, 13.06×10^{-4} in MAE, 72.50% in sensitivity, and 72.51% $F2$ score. SvANet's performance in sperm segmentation notably exceeded that of other models, surpassing them by up to 15.88% in $F2$ score, 14.99% in sensitivity, 14.64% in mDice, and 11.86% in mIoU. In addition, the performance metrics (mDice, mIoU, sensitivity, and $F2$ score) gained in the SpermHealth dataset are significantly lower than those observed in ISIC 2018, PolypGen, ATLAS, and KiTS23 for all tested models, with a gap of >10%, because

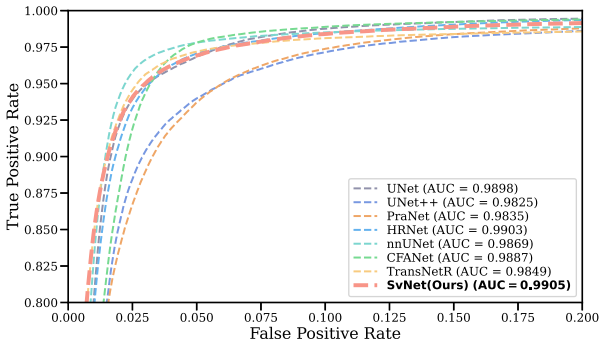

Fig. 7. ROC curves for tested models in the SpermHealth dataset.

all sperms have an area lower than 1%, presenting limited learnable features and posing more significant challenges for differentiation.

To quantify the robustness and adaptability of SvANet versus other SOTA methods, receiver operating characteristic (ROC) curves of seven tested methods in the SpermHealth dataset are employed and illustrated in Fig. 7. The ROC curve

TABLE IV

ABLATION STUDY RESULTS ON THE MAIN COMPONENTS OF SVANET. THE BEST RESULTS ARE UNDERLINED IN BOLD. ✗: CANCEL THE SETTING AND ✓: USE THE SETTING

Item	Ablation settings					mDice	p-value
	MCBottleneck	MCAttn	Cross-scale guidance	SvAttn	AssemFormer		
(a)	✗	✗	✗	✗	✗	71.10	-
(b)	✓	✗	✗	✗	✗	+0.15	0.042
(c)	✓	✓	✗	✗	✗	+0.67	0.033
(d)	✗	✗	✓	✗	✗	+0.27	0.002
(e)	✗	✗	✓	✓	✗	+0.47	0.002
(f)	✗	✗	✗	✗	✓	+0.38	0.002
(g)	✓	✗	✓	✗	✗	+0.46	0.001
(h)	✓	✓	✓	✓	✗	+1.32	0.013
(i)	✓	✗	✓	✓	✗	+0.54	0.001
(j)	✓	✓	✗	✗	✓	+0.50	0.001
(k)	✓	✓	✓	✓	✓	**+1.48**	0.001

TABLE V

COMPARISON OF MCATTN WITH OTHER ADVANCED ATTENTION METHODS. THE BEST RESULTS ARE UNDERLINED IN BOLD

Attention module	# Parameters /Million	mDice	Sensitivity	p-value
-	174.87	71.81	70.37	-
SE [23]	+2.77	71.19	71.39	0.048
CBAM [24]	+0.70	70.98	69.96	0.041
CoorAttn [25]	+0.10	71.43	70.14	0.049
MCAttn (Ours)	+2.77	**72.58**	**72.51**	0.022

of SvANet, represented by the red line in Fig. 7, blends nearest toward the top-left corner, with the highest area under the curve (AUC) of 0.9905, surpassing other SOTA methods by up to AUC of + 0.008. In addition, the ROC curve of U-Net++ is close to the lower-right corner and under all other curves, with the lowest AUC of 0.9825. The ROC and AUC results of U-Net++ are consistent with Table III, demonstrating that U-Net++ struggled to recognize sperms.

D. Ablation Studies

Unless otherwise specified, all ablation studies were conducted in the SpermHealth dataset for the sake of simplicity.

1) Inference Time: This section quantifies the inference characteristics of the tested networks, including the number of parameters (# Parameters), multiply–accumulate operations (MACs), and inferencing speed. The unit of inference speed is frames per second (FPS). The number of classes was set to eight, and other configurations were consistent with those described in Section IV-A2. The inference speed results, averaged over 1000 runs, are presented in Table III.

Table III illustrates that SvANet achieved a real-time analysis of medical images with 55 FPS. Notably, while SvANet consumed 312.76 billion MACs—two times more than CFANet—it performed at only 1 FPS lower than CFANet. This discrepancy highlights that MACs, as theoretical indicators of computational cost, do not fully capture the effects of hardware or software optimizations for inference. Despite the high computational load, SvANet's performance remains well-suited for self-examination and clinical diagnostic applications.

TABLE VI

SIZE COMBINATIONS OF POOLED FEATURE MAPS IN MCATTN. THE BEST RESULTS ARE UNDERLINED IN BOLD

Pooled tensor sizes	MACs/Million	mDice/%	Sensitivity/%
-	312,732.67	-	-
(1, 2)	+23.08	69.72	68.50
(1, 2, 3)	+23.09	**72.58**	**72.51**
(2, 3)	+23.09	71.76	70.24
(1, 2, 3, 4)	+23.10	71.39	70.85

In addition, a streamlined version of SvANet, named LiteSvANet, was developed by omitting the fifth encoder stage while retaining the ASPP on the fourth encoder stage, reducing the parameter count by 70%. Subsequent tests, under identical conditions (described in Section IV-A2), demonstrated that LiteSvANet achieved a mDice of 70.88% and a sensitivity of 69.96% in the SpermHealth dataset, surpassing the performance of the second-best method, TransNetR, as shown in Table III. Moreover, LiteSvANet significantly enhanced the inference speed to 77 FPS, which, while 23% lower than the fastest models (U-Net++ and TransNetR), represents a considerable improvement over the standard SvANet model. For straightforward applications, implementing LiteSvANet is advantageous for examining small medical objects.

2) Main Components Ablation: To investigate the influence of each core module of SvANet (i.e., MCBottleneck, MCAttn, cross-scale guidance, SvAttn, and AssemFormer), ablation studies were conducted and discussed in this section.

To investigate the influence of specific modules, we conducted experiments in which each module was individually included. MCAttn and SvAttn are part of MCBottleneck and cross-scale guidance, respectively. As presented in Table IV—(b)–(f), the inclusion of MCBottleneck, MCAttn, cross-scale guidance, SvAttn, and AssemFormer led to mDice improvements of + 0.15%, + 0.67%, + 0.27%, + 0.47%, and + 0.38%. Such results underscore the prominent contributions of MCAttn, SvAttn, cross-scale guidance, and AssemFormer in enhancing mDice. Specifically, SvAttn and MCAttn each contributed over 0.4% improvements in mDice.

Compatibility analysis of module combinations was conducted, with the results presented in Table IV(g)–(j). The combinations of all modules, excluding any attention mechanisms as shown in Table IV—(g), resulted in a + 0.46% increase in mDice, indicating that while nonattention modules are essential, they alone are insufficient to effectively enhance the learning capabilities of SvANet. In addition, Table IV—(h), which included MCAttn and SvAttn in a nontransformer version of SvANet, showed a notable improvement of + 1.32% mDice.

To elucidate the specific contributions of the "detailing" and "tracing" concepts introduced in Fig. 1, separate ablation studies were performed by excluding combinations of MCAttn + AssemFormer and cross-scale guidance + SvAttn, respectively. As shown in Table IV—(i) and (j), both "detailing" and "tracing" modules contributed comparable contributions to the mDice. Without "detailing" modules

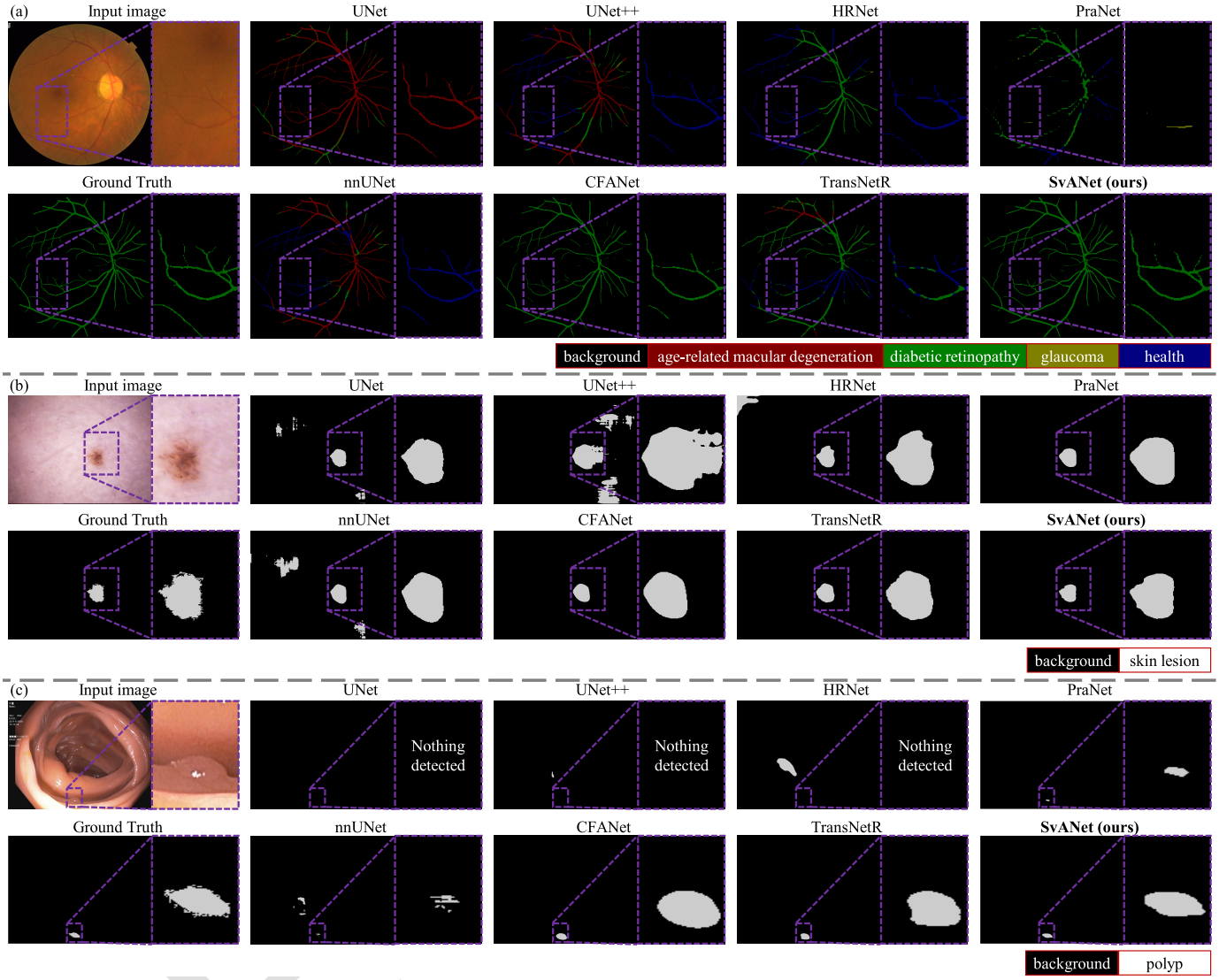

Fig. 8. Examples of segmentation results across tested methods in (a) FIVES, (b) ISIC 2018, and (c) PolypGen datasets for error analysis. Examples contain ultrasmall objects (i.e., polyp), small objects (i.e., nevus), and objects with >10% area ratio (i.e., retinal vessel).

(MCAttn + AssemFormer), SvANet registered 71.64% in mDice. Conversely, when the "tracing" modules (cross-scale guidance + SvAttn) were omitted, the performance decreased to 71.60% in mDice.

By integrating all five modules, SvANet obtained the highest improvement in mDice of + 1.48%, revealing the necessity of each module. All p-values for mDice are below 0.05, confirming the result's reliability.

3) MCAttn Versus Other Advanced Attention Methods: To assess the impact of different attention mechanisms within the MCBottleneck, three advanced attention modules, including SE, CBAM, and CoorAttn, were utilized as the control group. According to the results shown in Table V, MCAttn achieved performance improvements of over + 1.15% in mDice and + 1.12% sensitivity compared to these alternatives. Notably, the control group's attention methods resulted in reduced performance, with decreases of up to −0.83% in mDice and −0.41% in sensitivity, underscoring the superior efficacy of MCAttn in enhancing medical image segmentation within a bottleneck structure. The p-value for mDice is below 0.05, affirming the reliability of the result.

4) Number of Pooled Tensors for MCAttn: The selection of the size and number of pooled tensors for MCAttn is crucial for expanding network variants. We tested combinations (1, 2), (1, 2, 3), (2, 3), and (1, 2, 3, 4). The results, shown in the first, second, and fourth rows of Table VI, reveal that the (1, 2, 3) combination of pooled tensors outperformed (1, 2) and (1, 2, 3, 4) combinations, with improvements exceeding 2.86% and 1.19% in mDice and 4.01% and 1.66%, respectively. Further analysis, as indicated in the second and third rows of Table VI, highlights the necessity of a pool size of 1, leading to an increase of 0.82% in mDice and + 2.27% in sensitivity. These findings emphasize the importance of maintaining an optimal level of variation in the network. An insufficient pooled tensor can limit performance, whereas an excessive number can introduce too much stochasticity. Thus, striking the right balance is critical for maximizing the effectiveness of MCAttn within the model.

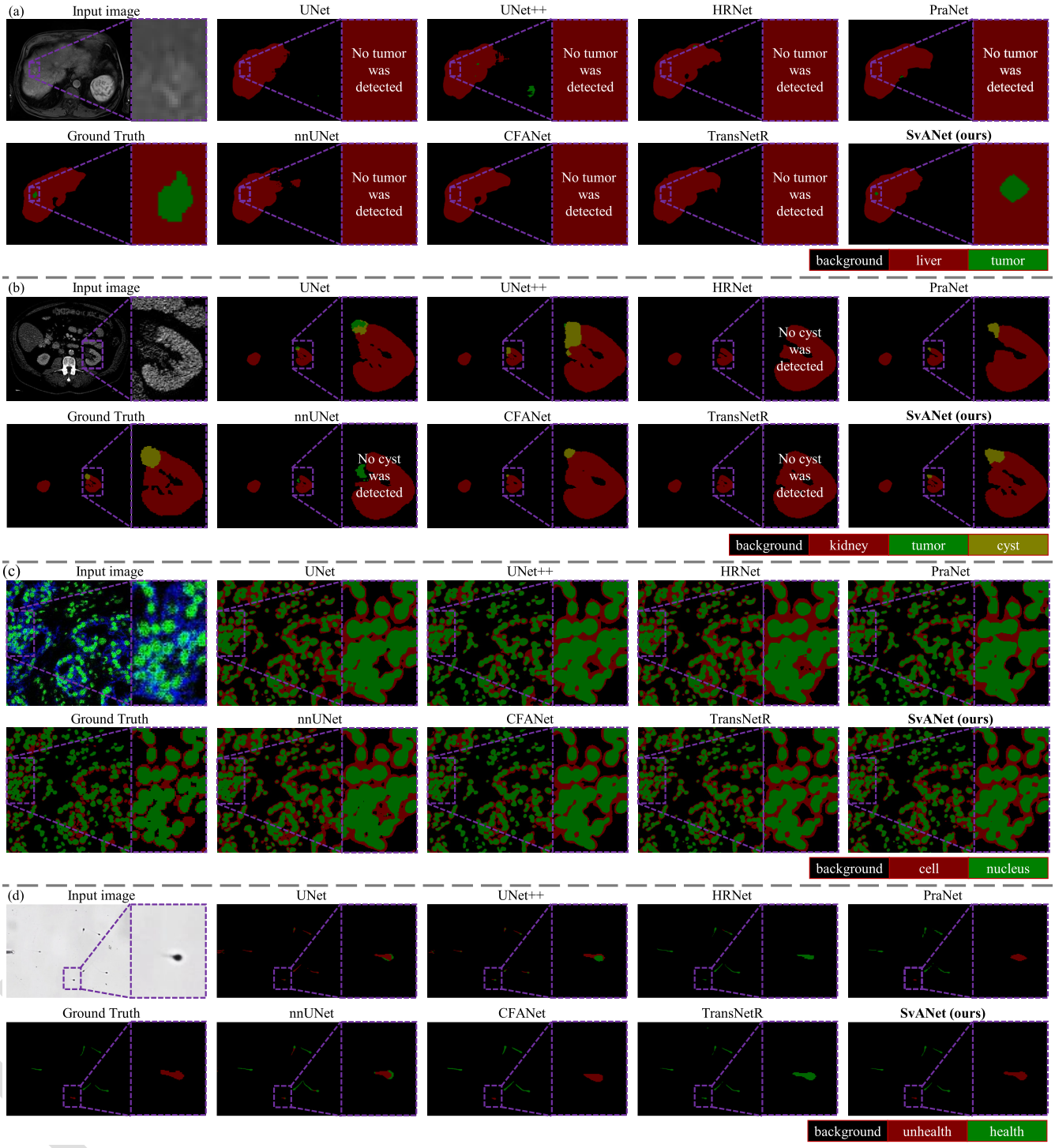

Fig. 9. Examples of segmentation results across tested methods in (a) ATLAS, (b) KiTS23, (c) TissueNet, and (d) SpermHealth datasets for error analysis. Examples contain ultrasmall objects (i.e., tumor, cyst, tissue cell, nucleus, and sperm), small objects (i.e., kidney), and objects with >10% area ratio (i.e., liver).

Furthermore, variations in pooled tensor sizes result in less than a 0.05% difference in MACs, indicating a negligible influence on computational complexity. Given that pooling operations do not add parameters to the model size and increase the MACs by less than 0.01%, MCAttn's impact on the network's computational performance is minimal.

E. Negative Case Studies

Examples of visualization results for ultrasmall and small medical objects in the FIVES, PolypGen, ISIC 2018, ATLAS, KiTS23, TissueNet, and SpermHealth datasets are presented in Figs. 8 and 9. As illustrated in Fig. 9(a), U-Net misclassified

diabetic retinopathy (green region) as age-related macular degeneration (red region). Similarly, U-Net++, HRNet, nnUNet, and TransNetR misclassified diabetic retinopathy as healthy retinal vessels (blue region), while PraNet misclassified it as glaucoma (yellow region). In addition, PraNet struggled to detect retinal vasculature in the zoomed-in region. Furthermore, none of the SOTA methods in the control group could accurately recover the retinal vessels at the bottom of the zoomed-in region. In contrast, SvANet not only correctly classified diabetic retinopathy but also effectively detected the position and shape of retinal vessels.

In the skin lesion examination illustrated in Fig. 9(b), U-Net, U-Net++, HRNet, and nnUNet misclassified normal skin as a nevus. Moreover, U-Net, PraNet, nnUNet, CFANet, and TransNetR represented the nevus region as a relatively smooth circle, while U-Net++ and HRNet captured a larger region encompassing the GT annotations, leading to an underestimation of the lesion boundary. In contrast, SvANet accurately identified a skin lesion of similar size to the GT and delineated its sawtooth-shaped boundary. For the polyp diagnosis as presented in Fig. 9(c), SOTA methods such as PraNet and nnUNet either detected a smaller polyp area than the GT, or other methods, including CFANet and TransNetR, regarded a larger region than the GT. In addition, U-Net, U-Net++, and HRNet failed to detect the polyp in the example image. Furthermore, the detected regions from the methods in the control group significantly deviated from the GT. However, SvANet recognized an area close to the GT and maintained shapes akin to GT annotations.

For MRI and CT image modalities analysis, as shown in Fig. 9(a) and (b), it is possible to overlook the overlapping medical objects, particularly ultrasmall ones. For instance, all tested models in the control group failed to identify an ultrasmall tumor inside the liver. In addition, HRNet, nnUNet, and TransNetR missed an ultrasmall cyst at the edge of the kidney. Moreover, U-Net++ and CFANet incorrectly emphasized the background as a tumor or liver region in the example image, and U-Net misclassified a cyst as a tumor. Although the organ region (e.g., liver and kidney) detected by SOTA methods in the control group appeared complete, the pathological regions, such as the tumor and cyst, were either larger (hepatic tumor) or smaller (cyst) than the GT in the example image. However, SvANet accurately differentiated between organs and their pathological regions. Furthermore, SvANet captured the morphological details of the liver, hepatic tumor, kidney, and cyst in the example image, closely aligning with the GT annotations.

For tissue cell recognition in the TissueNet dataset, as shown in Fig. 9(c), both TransNetR and SvANet effectively delineated cell boundaries and accurately labeled the cells and nuclei regions, closely resembling the GT. In contrast, other SOTA methods struggled to categorize cells and nuclei, leading to difficulties in differentiating cell boundaries and merging several cells. For sperm cell analysis, as presented by the final image in Fig. 9(d), SvANet precisely located all sperm positions and effectively recognized the region of the short tail of an abnormal sperm. Conversely, tested methods like PraNet and CFANet struggled to differentiate the head and tail of the unhealthy sperm, as illustrated in the zoomed-in region of Fig. 9(d). Moreover, U-Net, U-Net++, HRNet, nnUNet, and TransNetR misclassified an unhealthy sperm head as healthy, as indicated by a green subregion in Fig. 9(d).

These visualization results align with the findings discussed in Sections IV-B and IV-C, suggesting that SvANet holds significant potential for application in general small medical object recognition across various medical imaging modalities for disease diagnostics and surgeries.

V. Conclusion

This article introduces SvANet, a novel network to enhance the segmentation of small medical objects, aiding in the detection of life-threatening diseases and supporting in vitro fertilization. The experimental results demonstrate that the SvANet is significantly effective in distinguishing medical objects of various sizes. SvANet consistently outperformed other SOTA methods, achieving up to 19.95%, 15.03%, 15.01%, 14.64%, 13.57%, 8.09%, and 3.07% increments in mDice for segmenting objects occupying less than 1% image area across TissueNet, FIVES, ISIC 2018, SpermHealth, PolypGen, ATLAS, and KiTS23 datasets. Furthermore, the visualization results confirm that SvANet accurately identifies the locations and morphologies of all medical objects, demonstrating its exceptional capability in segmenting small medical objects. These findings underscore the potential of SvANet as a significant advancement in medical imaging.

In addition, SvANet features a substantial model size of over 150 million parameters and a computational burden of over 300 billion MACs, which is best suited for scenarios that can accommodate its high computational demands and require enhanced recognition accuracy. In contrast, LiteSvANet, streamlined to around 53 million parameters, offers a viable alternative for integration into low-performance devices, balancing computational efficiency with performance needs.

References

[1] K. Jin et al., "FIVES: A fundus image dataset for artificial intelligence based vessel segmentation," *Sci. Data*, vol. 9, no. 1, p. 475, Aug. 2022.
[2] N. Codella et al., "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," 2019, *arXiv:1902.03368*.
[3] S. Ali et al., "A multi-centre polyp detection and segmentation dataset for generalisability assessment," *Sci. Data*, vol. 10, no. 1, p. 75, Feb. 2023.
[4] F. Quinton et al., "A tumour and liver automatic segmentation (ATLAS) dataset on contrast-enhanced magnetic resonance imaging for hepatocellular carcinoma," *Data*, vol. 8, no. 5, p. 79, Apr. 2023.
[5] N. Heller et al., "The KiTS21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase CT," 2023, *arXiv:2307.01984*.
[6] N. F. Greenwald et al., "Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning," *Nature Biotechnol.*, vol. 40, no. 4, pp. 555–565, Apr. 2022.
[7] D. Bouget, A. Pedersen, J. Vanel, H. O. Leira, and T. Langø, "Mediastinal lymph nodes segmentation using 3D convolutional neural network ensembles and anatomical priors guiding," *Comput. Methods Biomechanics Biomed. Engineering: Imag. Visualizat.*, vol. 11, no. 1, pp. 44–58, Jan. 2023.
[8] W. Dai et al., "Automated non-invasive analysis of motile sperms using sperm feature-correlated network," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 3960–3970, 2025.

[9] J. Ding et al., "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7778–7796, Nov. 2022.
[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.
[11] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.
[12] T. Zhou et al., "Cross-level feature aggregation network for polyp segmentation," *Pattern Recognit.*, vol. 140, Aug. 2023, Art. no. 109555.
[13] X. Pan et al., "SMILE: Cost-sensitive multi-task learning for nuclear segmentation and classification with imbalanced annotations," *Med. Image Anal.*, vol. 88, Aug. 2023, Art. no. 102867.
[14] W. Dai, R. Liu, T. Wu, M. Wang, J. Yin, and J. Liu, "Deeply supervised skin lesions diagnosis with stage and branch attention," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 2, pp. 719–729, Feb. 2024.
[15] H. Du, J. Wang, M. Liu, Y. Wang, and E. Meijering, "SwinPA-Net: Swin transformer-based multiscale feature pyramid aggregation network for medical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 4, pp. 5355–5366, Apr. 2024.
[16] D. Guo, L. Zhu, Y. Lu, H. Yu, and S. Wang, "Small object sensitive segmentation of urban street scene with spatial adjacency between object classes," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2643–2653, Jun. 2019.
[17] J. Miao et al., "SC-SSL: Self-correcting collaborative and contrastive co-training model for semi-supervised medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 43, no. 4, pp. 1347–1364, Apr. 2024.
[18] D.-P. Fan et al., "PraNet: Parallel reverse attention network for polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 263–273.
[19] A. Lou, S. Guan, and M. Loew, "CaraNet: Context axial reverse attention network for segmentation of small medical objects," *J. Med. Imag.*, vol. 10, no. 1, Feb. 2023, Art. no. 014005.
[20] Y. Yuan, Y. Wu, X. Fan, M. Gong, W. Ma, and Q. Miao, "EGST: Enhanced geometric structure transformer for point cloud registration," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 9, pp. 6222–6234, Sep. 2024.
[21] Y. Wu, J. Liu, M. Gong, Z. Liu, Q. Miao, and W. Ma, "MPCT: Multiscale point cloud transformer with a residual network," *IEEE Trans. Multimedia*, vol. 26, pp. 3505–3516, 2024.
[22] Y. Wu et al., "Evolutionary multitasking descriptor optimization for point cloud registration," *IEEE Trans. Evol. Comput.*, vol. 29, no. 4, pp. 1239–1253, Aug. 2025.
[23] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 7132–7141.
[24] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.
[25] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13713–13722.
[26] F. Isensee et al., "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2020.
[27] W. Zhang, C. Fu, Y. Zheng, F. Zhang, Y. Zhao, and C.-W. Sham, "HSNet: A hybrid semantic network for polyp segmentation," *Comput. Biol. Med.*, vol. 150, Nov. 2022, Art. no. 106173.
[28] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[29] D. Jha, N. K. Tomar, V. Sharma, and U. Bagcı, "TransNetR: Transformer-based residual network for polyp segmentation with multi-center out-of-distribution testing," in *Proc. Med. Imag. Deep Learn.*, 2023, pp. 1372–1384.
[30] L. Zhao et al., "A novel framework for segmentation of small targets in medical images," *Sci. Rep.*, vol. 15, no. 1, p. 9924, Mar. 2025.
[31] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.
[32] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
[33] Y. Mei et al., "Pyramid attention network for image restoration," *Int. J. Comput. Vis.*, vol. 131, no. 12, pp. 3207–3225, Dec. 2023.
[34] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.
[35] W. Liu, X. Kang, P. Duan, Z. Xie, X. Wei, and S. Li, "SOSNet: Real-time small object segmentation via hierarchical decoding and example mining," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 2, pp. 3071–3083, Feb. 2025.
[36] S. Sang, Y. Zhou, M. T. Islam, and L. Xing, "Small-object sensitive segmentation using across feature map attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 6289–6306, May 2023.
[37] Q. Zhou, H. Shi, W. Xiang, B. Kang, and L. J. Latecki, "DPNet: Dual-path network for real-time object detection with lightweight attention," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 3, pp. 4504–4518, Mar. 2025.
[38] X. He, E.-L. Tan, H. Bi, X. Zhang, S. Zhao, and B. Lei, "Fully transformer network for skin lesion analysis," *Med. Image Anal.*, vol. 77, Apr. 2022, Art. no. 102357.
[39] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–21.
[40] F. Hörst et al., "CellViT: Vision transformers for precise cell segmentation and classification," *Med. Image Anal.*, vol. 94, May 2024, Art. no. 103143.
[41] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–26.
[42] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2025, pp. 5998–6008.
[43] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. 7th Int. Conf. Learn. Represent. (ICLR)*, May 2019, pp. 1–18.
[44] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–16.
[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[46] K. Roman. (2019). *Human Segmentation Dataset—TikTok Dances*. [Online]. Available: https://www.kaggle.com/datasets/tapakah68/segmentation-full-body-tiktok-dancing-dataset