Sparsely Annotated Medical Image Segmentation via Cross-SAM of 3D and 2D Networks

Huaqiang Su[1], Zaiyi Liu[2], Lisha Yao[2], Sunyun Li[2], Hun Lin[2], Guoliang Chen[1], Xin Chen[3], Haijun Lei[1⋆], and Baiying Lei[4⋆]

[1] Key Laboratory of Service Computing and Applications, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
[2] Department of Radiology, Guangdong Provincial People's Hospital, Guangzhou 510080, China
[3] Department of Radiology, Guangzhou First People's Hospital, Guangzhou 510180, China
[4] School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen 518060, China
lhj@szu.edu.cn, leiby@szu.edu.cn

Abstract. Medical image segmentation typically relies on large, accurately annotated datasets. However, acquiring pixel-level annotations is a labor-intensive process that demands substantial effort from domain experts, posing significant challenges in obtaining such annotations in real-world clinical settings. To tackle this challenge, we present the SA-Net framework, which leverages cross-supervision from segment anything models (SAM) and 2D segmentation networks to learn from sparse annotations. Specifically, we design an interactive graph learning segmentation network, which employs a bilateral graph convolution (BGC) module to capture more detailed features from multiple perspectives, facilitating the generation of high-quality pseudo-labels, which can serve as direct supervision for semantic segmentation networks and SAM, enabling the synthesis of additional annotations to enhance the training process. The multi-scale attention (MSA) module facilitates cross-layer interaction by partitioning channel label groups and capturing global information across layers, while the recovery module (RM) utilizes deep features and low-level features to fuse global context information and reconstruct lesion boundary regions. Our experimental results on LUNA16, AbdomenCT-1K, and self-collected datasets demonstrate the effectiveness of SA-Net. Our code is available at https://github.com/CTSegPilot/SA-Net.git.

Keywords: Sparse annotation · Segment anything model · Multi-scale attention module · Bilateral graph convolutional module.

1 Introduction

Accurate and robust medical image segmentation is crucial for ensuring reliable clinical diagnosis, as segmenting organs or lesions provides valuable diagnostic information for healthcare professionals [19, 11]. With the progress of deep learning

⋆ Corresponding authors: Haijun Lei and Baiying Lei.

methods, high-performance automatic segmentation algorithms have emerged. Fully convolutional networks [10, 3, 25] and encoder-decoder architectures, such as Swin-SMT [16], CTN [14], HENet [26], EoFormer [20], and ConvUNET [23], are widely adopted for pixel-level or voxel-level segmentation across various medical imaging tasks. A key enabler of these advancements is the availability of high-quality, fully labelled training datasets [27, 22]. However, obtaining annotations for medical images is costly and time-consuming, particularly for 3D volumetric data, requiring specialized expertise to delineate each case.

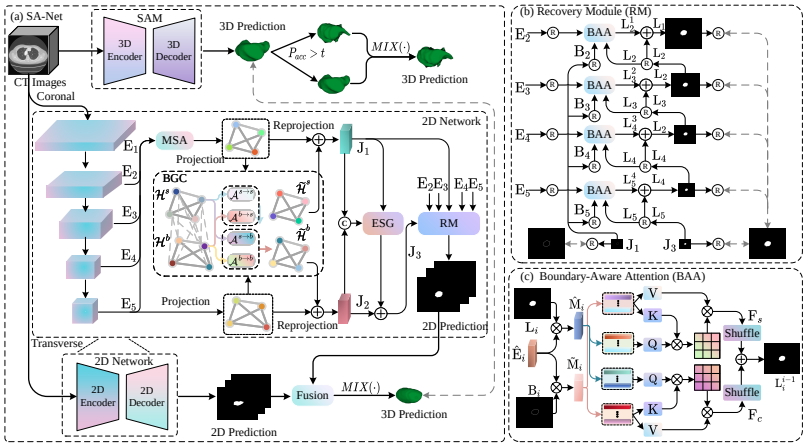


Fig. 1. Overview of the SA-Net architecture for medical image segmentation. SA-Net uses a SAM for preliminary segmentation, a 2D network to predict lesions, and finally, a 2D and SAM network to supervise the segmentation. $\mathbf{J}_1$ and $\mathbf{J}_2$ represent the enhanced feature maps of boundaries and regions. t is the threshold that divides the SAM's predictions. The operation Ⓡ denotes the resizing operation.

The high cost and limited availability of labelled data have emphasized the importance of research in incompletely supervised learning. To enable precise segmentation of medical images with minimal annotations, various strategies, including semi-supervised, self-supervised, weakly supervised, and unsupervised learning, have been explored [15, 7, 4, 2]. Ren et al. [17] introduced a weakly- and semi-supervised method for lesion segmentation, which trains the model using weakly annotated images along with unlabeled ones, thereby reducing dependence on fully annotated data. However, the boundaries between lesions and surrounding tissues are often unclear, and annotation methods typically fail to provide precise object boundary information, leading to a significant performance gap compared to fully supervised approaches. Sparse annotations, which involve labelling only a few slices of each computed tomography (CT) image, have been shown to preserve accurate boundaries for different categories [5]. To increase slice differences, many existing methods generate pseudo labels through image registration. Cai et al. [1] proposed a co-training framework that combines

dense pseudo labels with sparse orthogonal annotations. However, this method relies on the quality of the registration, and performance can degrade significantly when registration fails, such as in the case of small or complex objects.

Transfer learning has become a powerful approach to overcome the challenge of scarce annotated data. This approach utilizes pre-trained models, which have been trained on large, fully annotated datasets, to improve performance on target datasets with fewer annotations. Recently, large language and vision models have demonstrated the potential of modern transformer architectures by training on large, previously unseen datasets [8, 9, 21]. Notably, the segment anything model (SAM) [8, 24] has exhibited exceptional segmentation capabilities in image processing, inspiring its application in this study. We employ SAM [24] to predict ground truth (GT) for CT image slices where annotations are unavailable, leveraging its performance without the need for additional training.

This study introduces a novel, sparsely annotated cross-SAM network, SA-Net, to tackle challenges in medical image segmentation. First, we propose a bilateral graph convolution (BGC) module that captures the inter-task relationship by applying dual constraints between semantics and boundary features through graph interactions. Secondly, the multi-scale attention (MSA) module enables cross-layer interaction by partitioning channel and spatial label groups to capture global information across layers. Lastly, the recovery module (RM) combines deep and low-level features to reconstruct lesion boundaries, enhancing segmentation performance. The key contributions of our work are as follows:

- The BGC module extracts features from semantic segmentation and boundary detection tasks through graph interactions, allowing the network to learn improved semantic and geometric relationships from both labelled and unlabeled data.
- The MSA module enhances local cross-layer interactions across channels and spatial dimensions while capturing global context within each channel and spatial token group.

2 Method

2.1 Cross-SAM of Sparse Annotation

To enhance slice diversity and cope with the problem of limited supervision signals and sparse annotations that make direct training difficult, we learn from two different slices and utilize SAM and two 2D networks to generate pseudo-labels for each other. Specifically, the 3D network directly uses each sample as input. Subsequently, the data were sliced in two directions, generating transverse and coronal plane slices, which were employed to train the 2D segmentation networks. The 2D network uses the MSA module to divide the channels and spatial label groups to realize the cross-layer interaction of the ResNet-50 [6] encoder features to perceive cross-layer global information. It then feeds the fused features into the BGC module to impose dual constraints between semantics and boundaries while globally capturing both intra-task and inter-task relationships. Finally,

the deep features and low-level features are integrated by RM to reconstruct the lesion boundary area and enhance the ability to capture lesion boundary features. To enhance the supervision signal for each training sample, the selected pseudo labels are combined with the sparse GT annotations for guidance. The 2D network learns the foreground and background on different slices and generates pseudo labels for the 3D network by making consistent predictions on the same input sample. It is expressed as follows:

$$\widehat{G} = MIX\left(G, P\right), R_{acc} \approx P_{acc} = \sum_{i=1}^{H\times W\times D} \prod \frac{\widehat{p}_i = y_i}{H \times W \times D} \tag{1}$$

where $MIX\left(\cdot\right)$ is a function that substitutes the labels in P of voxels with GT annotations with labels in G of size $(H \times W \times D)$. $\prod(\cdot)$ denotes the indicator function, and $\widehat{p}_i$ represents the one-hot prediction for voxel i. y_i represents the label corresponding to voxel i in $\widehat{G}$.

Due to the limitations of the supervision signal, SAM's predictions often contain noisy labels. Directly using these as pseudo labels for 2D networks may lead to performance degradation. To address this, a confidence threshold is introduced to identify voxels with a higher likelihood of being accurate. However, the true accuracy R_{acc} of the predictions remains unknown, as dense annotations are unavailable during training. Since R_{acc} and pseudo accuracy P_{acc} are related to the training samples, estimating R_{acc} using P_{acc} is reasonable.

2.2 Bilateral Graph Convolutional Module

In image segmentation, geometric information and the correspondence between semantics and geometry are often overlooked, leading to inconsistent segmentation results, particularly for lesions with blurred boundaries. The BGC module leverages bilateral graph convolution to reinforce dual constraints between semantics and boundaries, thereby enabling the global exploration of task relationships. To project and reproject the semantic-aware graph g_s and boundary-aware graph g_b, we employ graph convolution to propagate information across the graphs. The augmented form of the bilateral graph was defined as follows:

$$\mathcal{H} = \left[\left(\mathcal{H}^s, \left(\mathcal{H}^b\right)^T\right)\right]^T, \mathcal{W} = \left[\left(\mathcal{W}^s, \left(\mathcal{W}^b\right)^T\right)\right]^T \tag{2}$$

where $\mathcal{H}$ and $\mathcal{W}$ represent the augmented form of the bilateral node features and weight matrices. $\mathcal{W}^s$ and $\mathcal{W}^b$ are two trainable weight matrices that adjust the node dimensions of $\mathcal{H}^s$ and $\mathcal{H}^b$, respectively.

The intra-graph reasoning captures the long-range dependencies within each graph. In this study, the adjacency matrix A is composed of the intra-graph matrix ($\mathcal{A}^{intra}$) and the inter-graph matrix ($\mathcal{A}^{inter}$), and is expressed as:

$$\mathcal{A} = \mathcal{A}^{intra} + \mathcal{A}^{inter} = \begin{pmatrix} \mathcal{A}^{s\to s} & 0 \\ 0 & \mathcal{A}^{b\to b} \end{pmatrix} + \begin{pmatrix} \mathcal{A}^{b\to s} & 0 \\ 0 & \mathcal{A}^{s\to b} \end{pmatrix} \tag{3}$$

where $\mathcal{A}^{s\to b} = \{a_{ij}^{s\to b}\}$ represents the correlation weights from the j-th node of g^s to the i-th node of g^b, with s and b denoting semantic and boundary features,

respectively. The coefficients a_{ij} indicate the importance of node j for node i. It is worth emphasizing that the graph constructed here is directional, as the weight vector $\mathcal{W}$ differs when learning a_{ij} and a_{ji}. A single graph convolution layer is defined using the normalized adjacency matrix $\mathcal{A}$, augmented bilateral node features $\mathcal{H}$, and the weight matrix $\mathcal{W}$ as follows:

$$\widetilde{\mathcal{H}} = \mathcal{F}\left(\mathcal{H} \,||\, \alpha\left(\mathcal{A}\left(\mathcal{H} \otimes \mathcal{W}\right)\right)\right), \mathcal{H} \otimes \mathcal{W} = \left[\left(\mathcal{H}^s \mathcal{W}^s\right)^T, \left(\mathcal{H}^b \mathcal{W}^b\right)^T\right]^T \tag{4}$$

where $\mathcal{F}(\cdot)$ combines the original and updated features. The enhanced graph representation is obtained by reprojection to the original coordinate space, and $\mathbf{J}_3$ is derived via feature interaction using ESG-Conv [25]. $\|$ is the concatenation.

2.3 Attention and Recovery Module

To mitigate the semantic gap and prevent information loss that can arise from element-wise summation and layer-by-layer transmission, Fig. 2 utilizes MSA to capture cross-layer global information, enabling feature interaction across layers. The encoded features $\mathbf{E}_i$ are oscillated at different frequencies through MFCA to generate feature maps $\mathbf{F}_i$ to enrich channel information and help the model capture subtle differences in irregular lesion features [13]. CCA facilitates multi-scale interactions across layers in the channel dimension, providing global contextual details across the spatial dimension for each token along the channel. To build the interactive input, CR is first applied to the feature map of each scale to ensure consistent spatial resolution, resulting in $\widehat{\mathbf{F}}_i$. Next, we apply OCP to form channel-wise token groups. Using the feature map from the i-th layer, after acquiring $\widetilde{\mathbf{F}}$, cross-layer consistent multi-head attention is employed to capture global dependencies along the spatial dimension, producing the interaction output $\widetilde{\mathbf{Y}}_i$. The multi-head mechanism is employed to model global dependencies for each token along the channel dimension.

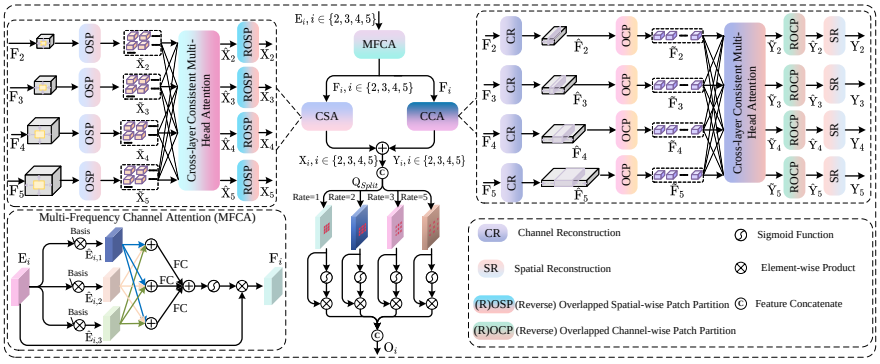


Fig. 2. Overview of MSA module. CCA and CSA promote local cross-layer interactions along the channel and spatial dimensions. FC is a fully connected operation.

After acquiring the interaction outputs $\widetilde{\mathbf{Y}}$ for the feature maps at each scale, ROCP was applied to derive $\widehat{\mathbf{Y}}_{i=0}$. Subsequently, SR is used to generate the final result $\mathbf{Y}_i$, which matches the shape of the input $\mathbf{F}_i$. Similarly, the cross-layer spatial-wise attention (CSA) enables multi-scale interactions between neighboring regions across layers along the spatial dimension, providing global contextual information $\widetilde{\mathbf{X}}_i$ along the channel dimension for each spatial-wise token. We then employ a cross-layer consistent multi-head attention mechanism to capture the global dependencies $\widehat{\mathbf{X}}_i$ in the spatial dimension and apply ROSP to obtain the interaction result $\mathbf{X}_i$. To effectively capture multi-scale information, we utilize four parallel convolutional layers with varying convolution rates to obtain multi-scale features $\mathbf{O}_i$, addressing the issue that a single-scale feature may not adequately cover the diverse sizes and locations of lesions.

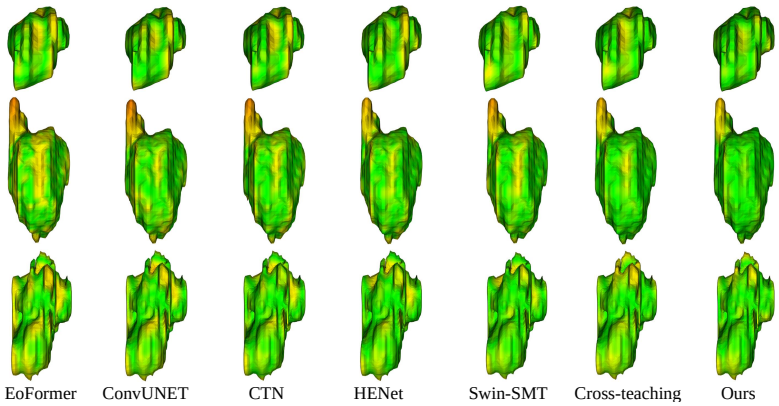


Fig. 3. Example of segmentation results from a self-collected nodule dataset.

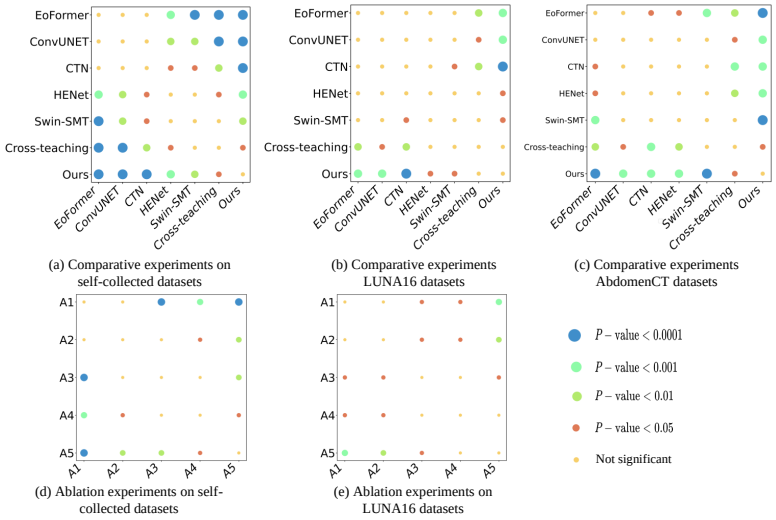
The structure of the boundary-aware attention (BAA) module is shown in Fig. 1 (c). Before being fed into the BAA module, the features $\mathbf{E}_i$ are resized to the corresponding resolution $\hat{\mathbf{E}}_i$ to provide the boundary GT $\mathbf{B}_i$ and the internal lesion region GT $\mathbf{L}_i$. We first perform element-wise product of $\hat{\mathbf{E}}_i$ with $\mathbf{B}_k$ and $\mathbf{L}_k$ to extract the features corresponding to the boundary and interior lesion regions, denoted as $\tilde{\mathbf{M}}_i$ and $\hat{\mathbf{M}}_i$, respectively. $\tilde{\mathbf{M}}_i$ is used as the query ($\mathbf{Q}$) matrix, and $\hat{\mathbf{M}}_i$ as the key ($\mathbf{K}$) and value ($\mathbf{V}$) matrices to compute the interaction feature to extract the characteristics of the internal lesion area more accurately. To enable SA-Net to capture richer global features, we shuffle $\mathbf{F}_s$ and $\mathbf{F}_c$, then combine them element-wise to obtain the output fused features $\mathbf{L}_k^{k-1}$.

SA-Net performs reasoning at low resolution and the direct interpolation of the final estimate results in performance degradation. To overcome this challenge, we propose the RM, which leverages both boundary and region features containing deep semantic information and shallow structural details to progressively refine the lesion area feature (LAF) estimation, as illustrated in Fig. 1(b). This approach enhances LAF prediction by jointly guiding it with the lesion

boundary characteristics (LBC) map $\mathbf{B}_i$, the LAC map $\mathbf{L}_i$.

$$\left\{\mathbf{L}_i^{i-1}\right\}_{i=5}^{2} = f_{BAA}\left(\hat{\mathbf{E}}_i, \mathbf{B}_i, \mathbf{L}_i\right), \mathbf{L}_{i-1} = \mathbf{L}_i + \mathbf{L}_i^{i-1} \tag{5}$$

where $f_{BAA}\left(\hat{\mathbf{E}}_i, \mathbf{B}_i, \mathbf{L}_i\right)$ is formulated to capture contextual information from the side-output feature $\mathbf{L}_i^{i-1}$, guided jointly by the LAC and LBC maps.



(a) Comparative experiments on self-collected datasets (b) Comparative experiments LUNA16 datasets (c) Comparative experiments AbdomenCT datasets

(d) Ablation experiments on self-collected datasets (e) Ablation experiments on LUNA16 datasets

Fig. 4. Bubble chart showing the statistical analysis results of each network.

3 Experiments

Datasets and Preprocessing: We evaluate SA-Net using three different datasets: LUNA16 [18], AbdomenCT-1K [12], and self-collected CT image dataset. The self-collected CT images dataset contains 1299 pulmonary nodule samples. After the preprocessing stage of the CT images, the grayscale images are used to determine the approximate location and size of the nodules. Subsequently, this information is used to select the region of interest (ROI) to include the surrounding area of the nodule. The ROI sizes for the self-collected CT images and the LUNA16 dataset are $160 \times 160 \times 48$ and $32 \times 64 \times 64$, respectively. We divide the CT images in AbdomenCT-1K into $96 \times 96 \times 96$ patches as the input volume and only one-third of the slices are annotated for each case.

Comparative Results Analysis: To demonstrate the proposed SA-Net's effectiveness, we experimentally compare it with other competing segmentation models on three segmentation datasets. The quantitative comparison results are shown in Tables 1 and 2. It can be observed that the collaborative learning

Table 1. Segmentation performance evaluation on three different datasets.

Method	Self-collected		LUNA16		AbdomenCT-1K	
	Dice (%)	BF (%)	Dice (%)	BF (%)	Dice (%)	BF (%)
EoFormer[20]	67.6±2.7	70.4±2.9	68.7±2.9	71.6±3.3	83.3±3.9	81.5±5.9
ConvUNET[23]	68.9±2.6	71.3±2.2	70.8±2.9	74.2±2.7	86.8±3.7	87.6±3.9
CTN[14]	70.7±3.8	73.3±3.5	68.8±1.7	70.3±3.9	87.8±2.2	85.7±6.9
HENet[26]	73.1±2.2	75.4±2.0	70.7±4.3	68.9±6.6	87.6±2.1	88.1±2.2
Swin-SMT[16]	74.7±2.4	76.8±2.2	71.6±3.2	71.0±5.2	88.6±2.6	88.5±2.6
Cross-teaching [2]	76.1±2.0	78.5±1.7	74.3±2.5	76.0±4.1	90.6±1.7	91.3±1.7
Ours	**78.8±2.2**	**80.9±2.1**	**76.0±1.5**	**78.7±1.4**	**93.0±1.4**	**92.7±1.2**

Table 2. The ablation experiments of the SA-Net on the nodule datasets. The baseline consists of SAM and two 2D segmentation networks with ResNet-50 as the encoder.

	Method					Self-collected		LUNA16	
	SAM	Baseline	MSA	BGC	RM	Dice (%)	BF (%)	Dice (%)	BF (%)
A1	✓					72.5±1.4	74.8±1.3	71.2±1.7	74.6±1.6
A2	✓	✓				74.7±2.3	77.1±2.4	71.9±2.4	75.2±2.5
A3	✓	✓	✓			76.0±1.6	78.4±1.5	73.3±2.3	76.3±2.1
A4	✓	✓	✓	✓		76.9±1.3	79.0±1.2	74.8±2.0	77.5±1.3
A5	✓	✓	✓	✓	✓	**78.8±2.2**	**80.9±2.1**	**76.0±1.5**	**78.7±1.4**

framework SA-Net can dynamically adjust SAM to provide more accurate pseudo labels as an additional supervisory signal, even in lesion areas with low contrast or blur. We analyze the Dice coefficient (Dice), Boundary F1 (BF), Jaccard Index (JI), and P-value of SA-Net and other models. Fig. 3 shows the 3D surface distance between the predicted and GT results. As the green area grows, the segmentation improves. These improvements are attributed to the fact that SA-Net is able to promote local cross-layer interactions along the channel and spatial dimensions through the MSA module, thereby building a global receptive field along the spatial and channel dimensions and utilizing different convolution rates to capture features of different scales, so that SA-Net can accurately decode complex and ambiguous areas.

Ablation Study: To assess the performance of the proposed SA-Net architecture, we conducted ablation studies on the pulmonary nodule dataset. Fig. 4 presents a bubble chart obtained by performing a statistical significance analysis on the JI value of each compared network on the dataset. We can see that the MSA module enhances the SA-Net's ability to leverage global contextual information and facilitate cross-layer, multi-scale feature integration. Combining the encoder features of the 2D network enables it to better solve the problems of irregular shapes and blurred boundaries of lesions. The BGC module extracts features from both semantic segmentation and boundary detection tasks through graph interactions, allowing the network to establish more robust semantic and

geometric relationships from both labelled and unlabeled data. RM utilized the characteristics of deep features that are good at capturing and transmitting semantic information. In contrast, low-level features are good at representing complex geometric details to reconstruct the lesion boundary area gradually.

4 Conclusion

We propose a novel SA-Net segmentation framework to address the limitations of existing sparsely annotated methods for medical image segmentation. The framework effectively utilizes SAM to provide pseudo labels as additional supervisory signals to improve segmentation performance. Extensive experiments on segmentation datasets demonstrate the effectiveness of our approach. Further directions will explore unsupervised learning to reduce reliance on large labelled datasets, making models more practical in data-scarce environments.

Acknowledgments. This work was supported partly by the National Natural Science Foundation of China (Nos. 62276172 and 62171312), National Natural Science Foundation of Guangdong Province (Nos. 2023A1515011378 and 2024A1515011950).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cai, H., Li, S., Qi, L., Yu, Q., Shi, Y., Gao, Y.: Orthogonal annotation benefits barely-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3302–3311 (2023)
2. Cai, H., Qi, L., Yu, Q., Shi, Y., Gao, Y.: 3d medical image segmentation with sparse annotation via cross-teaching between 3d and 2d networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 614–624. Springer (2023)
3. Du, Z., Hu, Z., Zhao, G., Jin, Y., Ma, H.: Cross-layer feature pyramid transformer for small object detection in aerial images. arXiv preprint arXiv:2407.19696 (2024)
4. Eyuboglu, S., Angus, G., Patel, B.N., Pareek, A., Davidzon, G., Long, J., Dunnmon, J., Lungren, M.P.: Multi-task weak supervision enables anatomically-resolved abnormality detection in whole-body fdg-pet/ct. Nature communications **12**(1), 1880 (2021)
5. Gao, F., Hu, M., Zhong, M.E., Feng, S., Tian, X., Meng, X., Huang, Z., Lv, M., Song, T., Zhang, X., et al.: Segmentation only uses sparse annotations: Unified weakly and semi-supervised learning in medical images. Medical Image Analysis **80**, 102515 (2022)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Huang, H., Lin, L., Zhang, Y., Xu, Y., Zheng, J., Mao, X., Qian, X., Peng, Z., Zhou, J., Chen, Y.W., et al.: Graph-bas3net: Boundary-aware semi-supervised segmentation network with bilateral graph convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7386–7395 (2021)

8. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
9. Li, X., Ouyang, X., Zhang, J., Ding, Z., Zhang, Y., Xue, Z., Shi, F., Shen, D.: Carotid vessel wall segmentation through domain aligner, topological learning, and segment anything model for sparse annotation in mr images. IEEE Transactions on Medical Imaging (2024)
10. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
11. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
12. Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., et al.: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6695–6714 (2021)
13. Nam, J.H., Syazwany, N.S., Kim, S.J., Lee, S.C.: Modality-agnostic domain generalizable medical image segmentation by multi-frequency in multi-scale attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11480–11491 (2024)
14. Pan, H., Hong, Y., Sun, W., Jia, Y.: Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes. IEEE Transactions on Intelligent Transportation Systems **24**(3), 3448–3460 (2022)
15. Peiris, H., Hayat, M., Chen, Z., Egan, G., Harandi, M.: Uncertainty-guided dual-views for semi-supervised volumetric medical image segmentation. Nature Machine Intelligence **5**(7), 724–738 (2023)
16. Płotka, S., Chrabaszcz, M., Biecek, P.: Swin smt: Global sequential modeling for enhancing 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 689–698. Springer (2024)
17. Ren, G., Lazarou, M., Yuan, J., Stathaki, T.: Towards automated polyp segmentation using weakly-and semi-supervised learning and deformable transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4355–4364 (2023)
18. Setio, A.A.A., Traverso, A., De Bel, T., Berens, M.S., Van Den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. Medical image analysis **42**, 1–13 (2017)
19. Shao, H., Zhang, Y., Hou, Q.: Polyper: Boundary sensitive polyp segmentation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 4731–4739 (2024)
20. She, D., Zhang, Y., Zhang, Z., Li, H., Yan, Z., Sun, X.: Eoformer: Edge-oriented transformer for brain tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 333–343. Springer (2023)
21. Shen, Y., Li, J., Shao, X., Inigo Romillo, B., Jindal, A., Dreizin, D., Unberath, M.: Fastsam3d: An efficient segment anything model for 3d volumetric medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 542–552. Springer (2024)

22. Song, B., Wang, Q.: Sdcl: Students discrepancy-informed correction learning for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 567–577. Springer (2024)
23. Tang, X., Liu, F., Kong, R., Luo, F., Huang, W., Zou, J.: Convunet: a novel depthwise separable convnet for lung nodule segmentation. In: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1443–1450. IEEE (2023)
24. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., et al.: Sam-med3d: towards general-purpose segmentation models for volumetric medical images. arXiv preprint arXiv:2310.15161 (2023)
25. Zhai, Q., Li, X., Yang, F., Jiao, Z., Luo, P., Cheng, H., Liu, Z.: Mgl: Mutual graph learning for camouflaged object detection. IEEE Transactions on Image Processing **32**, 1897–1910 (2022)
26. Zhou, W., Zhang, X., Gu, D., Wang, S., Huo, J., Zhang, R., Jiang, Z., Shi, F., Xue, Z., Zhan, Y., et al.: Henet: Hierarchical enhancement network for pulmonary vessel segmentation in non-contrast ct images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 551–560. Springer (2023)
27. Zhou, X., Sun, Y., Deng, M., Chu, W.C.W., Dou, Q.: Robust semi-supervised multimodal medical image segmentation via cross modality collaboration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 57–67. Springer (2024)