DUR-Net+: Semi-supervised abdominal CT pheochromocytoma segmentation via dynamic uncertainty rectified and prior knowledge from SAM-Med3D

Chuanbo Qin, Zhuyuan Chen, Dong Wang, Bin Zheng, Jun Luo, Junying Zeng, Xudong Jia, Jin Wen, Maoqing Hu, Yikui Zhai, Pasquale Coscia, *Senior Member, IEEE*, and Angelo Genovese, *Senior Member, IEEE*

***Abstract*— Pheochromocytoma is a rare urological adrenal tumor disease. Automated segmentation of pheochromocytomas from computed tomography (CT) is essential for diagnosis and treatment. However, this task is a challenging one due to issues such as blurred boundaries, irregular shapes, variations in location and size, and the lack of annotated images for training. To address these issues, we propose a semi-supervised framework for pheochromocytoma segmentation that primarily consists of a dynamic uncertainty rectification mechanism and a supervised strategy based on SAM-Med3D prior knowledge. First, we design a semi-supervised segmentation model comprising a shared encoder and multiple independent decoders that dynamically select pseudo labels from the different decoder outputs. To mitigate the risk of unreliable predictions caused by sparse annotations during training, we introduce uncertainty estimation to prioritize reliable outputs. Additionally, an Attentional Convolution Block (ACB) is designed in the encoding stage to fully utilize both global and local features, improving tumor recognition in segmentation. Furthermore, SAM-Med3D prior knowledge is incorporated into the framework as supplementary supervisory information, aiding the model in learning from limited labeled data. To eliminate the labor-intensive requirement for manual prompts in SAM-Med3D, we leverage pseudo labels to generate high-quality mask prompts, thus transforming the clinical workflow. Experiments on two datasets of pheochromocytoma from different centers demonstrate that our proposed method produces competitive performance.**

***Index Terms*— Pheochromocytoma segmentation, semi-supervised learning, dynamic rectification, uncertainty estimation, SAM-Med3D.**

This work was supported in part by the National High Level Hospital Clinical Research Funding (2022-PUMCH-C-044), in part by the 2024 Key Research Platform Project for Ordinary Universities in Guangdong Province (2024ZDZX1008), in part by the Guangdong Higher Education Innovation and Strengthening School Project (No. 2022ZDZX1032, No. 2023ZDZX1029, 2024ZDZX1009), in part by the Wuyi University Hong Kong and Macao Joint Research and Development Fund (No. 2022WGALH19), in part by the Guangdong Jiangmen Science and Technology Research Project (No. 2220002000246, No. 2023760300070008390), in part by the 2022 Guangdong Provincial Education Department Graduate Education Innovation Project (2022 No.2). (Chuanbo Qin, Zhuyuan Chen, Bin Zheng, Dong Wang, Jun Luo, and Yikui Zhai contributed equally to this work). (Corresponding authors: Junying Zeng; Xudong Jia; Jin Wen; Maoqing Hu.)

This work involved human subjects in its research. Approval Document of the Ethics Review Committee of Peking Union Medical College Hospital, Chinese Academy of Medical Sciences(23PJ1015).

Chuanbo Qin, Zhuyuan Chen, Bin Zheng, Junying Zeng, and YikuiZhai are with the School of Electronics and Information Engineering, Wuyi University, Jiangmen 529020, China (e-mail: tenround@163.com; chenzy183@163.com; ferzera@163.com; zengjunying@126.com; yikuizhai@163.com).

Jun Luo is with the School of Economics & Management , WuYi Univeristy, Jiangmen 529020, China (e-mail: 33007957@qq.com).

Dong Wang, Jin Wen are with the Department of Urology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China (e-mail: wangdongaaa777@163.com; wjpumch@163.com).

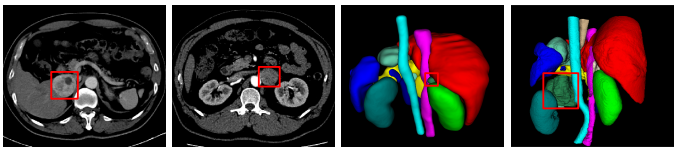
Maoqing Hu is with the Department of Radiology, Jiangmen Central Hospital, Jiangmen, 529020, China (e-mail: hmq7401@163.com).

Xudong Jia is with the College of Engineering and Computer Science, California State University, Northridge, California, 91330, USA (e-mail: Xudong.Jia@csun.edu).

Pasquale Coscia, Angelo Genovese, are with the Departiment of Computer Science and Dipartimento di Informatica, Universita Degli Studi di Milano, 20133, Milan, Italy (e-mail: pasquale.coscia@unimi.it; angelo.genovese@unimi.it).

I. Introduction

PHEOCHROMOCYTOMA is a rare neuroendocrine tumor with an annual incidence of approximately 0.8 per 100,000 people. Precise diagnosis and timely surgery can prevent tumor progression and metastasis, thereby reducing mortality [1]. However, the clinical diversity of pheochromocytomas makes diagnosis challenging for physicians. And pheochromocytomas are often located in abdominal region, surrounded by multi-organ anatomical structures. Tumor deterioration and enlargement can easily invade adjacent organs, further increasing the risk of surgery. Therefore, accurate contouring of pheochromocytomas from medical images is crucial for clinical diagnosis and surgical planning. Furthermore, annotating 3D pheochromocytoma data for fully supervised training is a time-consuming and costly task. This constraint has sparked interest in semi-supervised learning (SSL) methods that exploit large amounts of unlabeled data to improve model performance.



(a) (b) (c) (d)

Fig. 1. CT scan of the pheochromocytoma and its surrounding abdominal organs, with pheochromocytoma in the red box. (a) (b) shows the different abdominal locations where the pheochromocytoma is located, (c) and (d) show the variation in the size of the pheochromocytoma and its association with the surrounding organs. In addition, there is significant grey level variation within the pheochromocytoma in (a).

The current semi-supervised learning (SSL) methods mainly use two types of strategies to deal with medical image segmentation tasks: one is consistency regularization, such as the Mean Teacher (MT) framework, which ensures model output invariance against input perturbations through collaborative optimization between student-teacher networks. However, its reliance on exponential moving average (EMA) for teacher network updates inherently accumulates cognitive biases, especially in the gray similar regions of tumors and abdominal organs. The second is the pseudo-label strategy, which leverages models trained on limited labeled data to generate supervisory signals for unlabeled samples, followed by retraining the model with labeled data and supervisory signals. Although this expands the training corpus, it is easy to introduce a large number of noisy pseudo-labels in the scenario of low annotation rate. Moreover, the model performs self-rectification based on pseudo-labels, and such a learning paradigm that relies on internal supervisory signals tends to make the model acquire conservative information, thereby resulting in a state of stagnation.

In addition, the inherent properties of pheochromocytoma make automated segmentation challenging. First, tumor size varies greatly with the severity of pheochromocytoma. A small tumor lesion may contain very few voxels, while a large tumor can take a large body space, leading to a class imbalance problem. Second, tumors are not fixed in location. They appear in various areas of the abdomen and often have a complex spatial distribution. Third, tumor morphology is highly varied, smaller tumors are frequently homogeneous, whereas larger tumors may be heterogeneous in grayscale and exhibit central necrosis, which can lead to under-segmentation. Moreover, the proximity of tumors to abdominal organs may cause models to misinterpret the morphological variations of these organs as tumor boundaries and incorrectly segment them as part of the tumors, resulting in over-segmentation. These challenges, along with low soft tissue contrasts in abdominal CT images, further exacerbate the difficulty of pheochromocytoma segmentation. Fig. 1 shows four example CT scans with a pheochromocytoma, which demonstrates characteristics of the pheochromocytoma.

In the existing work, Tang et al. [2] formulated an adaptive energy function to achieve pheochromocytoma segmentation. This is a traditional method due to its manual parameter tuning operation. Wang et al. [3] transferred fingerprint information from target domains to source domains in training a model for abdominal multi-organ and pheochromocytoma segmentation, but pheochromocytoma labels are still required for supervision. Oluigbo et al. [4] adopted a weakly supervised approach to detect pheochromocytomas in CT images through a proxy segmentation task. The task focuses on obtaining the approximate location of the pheochromocytoma rather than pixel-level segmentation. Santra et al. [5] utilized information about the anatomical location of tumors to identify pheochromocytoma genetic clusters. Overall, very few deep learning methods have showcased their effectiveness in pheochromocytoma segmentation.

The recently introduced new AI model, SAM [6], has received significant attention thanks to its ability of utilizing prompts to generate fine-grained segmentation masks. Several studies [7] have made notable progress in employing SAM for medical image segmentation, with SAM-Med3D [8] being a representative work among them. It abandons SAM's inherent 2D architecture and instead processes 3D medical data directly. Extensively trained on large-scale medical datasets, SAM-Med3D efficiently segments medical images with only prompts. Considering SAM-Med3D's feature extraction capabilities, we propose utilizing its prior knowledge to enhance semantic feature representation for assisting pheochromocytoma segmentation. However, SAM-Med3D relies on manual prompt, and generating prompts requires expertise that is often inaccessible to non-specialist users, which reduces the utility of prompt-based approaches.

To address these issues, we propose a semi-supervised framework for pheochromocytoma segmentation, which consists primarily of a semi-supervised segmentation model and SAM-Med3D. Following [9], the semi-supervised segmentation model is a V-Net-based [10] model comprising a shared encoder and multiple independent decoders. In view of the intrinsic characteristics of pheochromocytomas that make target feature extraction difficult, we designed an Attentional Convolution Block (ACB) within the encoding stage of the model. The ACB fully leverages captured local and global information to accurately localize the region of interest, enabling the model to distinguish the tumor from surrounding organs. Additionally, we developed a dynamic uncertainty rectification mechanism to optimize the rectification process. Specifically, the optimal feature representation is initially selected as a pseudo label from multiple decoder outputs. Subsequently, uncertainty estimation is applied to emphasize reliable information and reduce noise in the pseudo labels. Furthermore, we propose a supervised strategy based on SAM-Med3D prior knowledge to enhance the utilization of unlabeled data, achieved by directly leveraging SAM-Med3D pre-training weights. In response to the manual prompt requirements of SAM-Med3D, we introduce a method for filtering pseudo labels to generate high-quality mask prompts, thereby eliminating the prompt requirements of labour-intensive pheochromocytoma annotations.

In summary, our work contributes the following:

- An innovative segmentation framework combining a semi-supervised segmentation model with SAM-Med3D is developed to efficiently integrate SAM-Med3D's prior knowledge in pheochromocytoma segmentation. Moreover, SAM-Med3D utilizes pseudo labels generated by

the semi-supervised model as prompt information, eliminating the need for labor-intensive manual prompts.

- In strengthening the rectification capacity of pseudo labels, a dynamic uncertainty rectification mechanism is proposed in the framework, which first dynamically selects pseudo labels from multiple decoder outputs, and then incorporates uncertainty estimation to emphasize reliable information, thereby improving the framework's ability to perceive pheochromocytoma. Additionally, an attention convolution block was designed to focus the semi-supervised segmentation model on the region of interest. As a result, the segmentation performance is significantly enhanced.

- The experimental results on two independent private pheochromocytoma datasets consistently validate that our method outperforms state-of-the-art semi-supervised approaches, thereby showcasing its robust and reliable capability in real-world scenarios.

II. RELATED WORK

A. The SAM in Medical Imaging

The Segment Anything Model (SAM) [6], which consists of an image encoder, a prompt encoder, and a mask decoder, has demonstrated significant segmentation advantages in a variety of scenarios. Chen et al. [11] used accurate masks extracted from SAM output to guide the training of student models, alleviating the noise impact in semi-supervised referring expression segmentation. Liu et al. [12] proposed using SAM to generate pseudo labels for remote sensing segmentation.

The introduction of SAM into medical image segmentation represents a significant research topic. Zhang et al. [13] proposed a novel iterative pseudo label correction based on the SAM framework to facilitate source-free domain adaptation medical image segmentation. Siblini et al. [14] explored prompt based medical image segmentation strategies in SAM under few sample and weakly supervised scenarios. Ma et al. [15] demonstrated the potential of SAM in the field of medical images by training SAM using bounding box prompts. Given that SAM is a 2D structure, a slice-by-slice approach is frequently employed in the processing of volumetric images. This approach fails to consider the contextual relationships between slices, which ultimately leads to suboptimal performance [16]. Some researchers have achieved 2D to 3D adaptation by freezing 2D layers and training 3D adapters, thereby enabling SAM to learn from 3D images [17]. Nevertheless, the intrinsic limitations of the 2D structure continue to constrain the applicability of SAM to 3D images. In response to this challenge, Wang et al. [8] proposed a novel approach, SAM-Med3D, which employs 3D medical images directly in the training and fine-tuning of SAM, thereby enabling it to achieve remarkable performance in the processing of contextual information.

In addition, manual prompt is a limitation of SAM, which diminishes the utility of SAM. To solve this, Li et al. [18] developed a lightweight model to replace SAM's original prompt encoder, training the feature map output from the adapter as prompt. Chen et al. [19] generated high-quality prompts by fusing multi-scale features from the image encoder. Although these methods automatically generate prompts, they still bring several problems, including: 1) the prompt generated by SAM's image encoder may contain noise, which can lead to incorrect segmentation guidance; 2) the input to the mask decoder is essentially taken from SAM's image encoder, exacerbating the learning of conserved information, which could make the mask decoder susceptible to overfitting; 3) the prompt generator is designed to be trainable in order to learn from various modes of image, which requires an additional computational load. In this paper, we propose utilising pseudo labels as prompts, which avoids the need for additional computation and ensures that the prompts do not originate from the image encoder in SAM.

B. Semi-Supervised Medical Image Segmentation

Popular semi-supervised methods can be grouped into two main categories: pseudo labeling and consistency regularization. Pseudo labeling methods are employed to generate pseudo labels for unlabeled data, thereby expanding the training dataset. Lee et al. [20] proposed to use outputs from the fully supervised segmentation model as pseudo labels. However, simple pseudo-labelling introduce noise effects to the training. To solve this, Sohn et al. [21] employed a threshold for the filtering of pseudo-labels with high confidence, thereby reducing the mislabelling rate. Additionally, Zhao et al. [22] propose to rectify pseudo supervision by prediction uncertainty estimation and consistency regularization to mitigate the impact of noise. These methods have significantly improved the quality of the pseudo labels. In terms of consistency regularisation, the objective is to encourage the model to produce outputs that are consistent with the inputs. For instance, MT [23] optimises the model by imposing consistency constraints on the predictions derived from both the original and perturbed samples. Luo et al. [24] introduced a dual-task consistency between the directly predicted segmentation map and the segmentation map derived from the level set. Furthermore, Wu et al.[9] enhanced the model's capacity to segment challenging regions by reinforcing the consistency of learning among the three decoders.

It is noted that there exist some other methods utilizing unlabeled data. Xu et al. [25] presents an ingenious approach to integrating deep reinforcement learning (DRL) and generative adversarial networks (GANs) in a unified pipeline, thereby optimising both detection and segmentation tasks. [26] adopted entropy minimisation for segmentation maps of unlabeled data to improve reliability. Furthermore, uncertainty estimation [27], [28], [29] is generally used to improve model confidence.

C. Attention Mechanism

In the field of medical image segmentation, convolutional neural networks (CNNs) incorporating attention mechanisms have demonstrated superior feature extraction capabilities [30]. Zhong et al. [31] proposed the Polarized Multi-scale Feature Self-attention (PMFS) block, which captures global contextual features, thus enabling the effective handling of

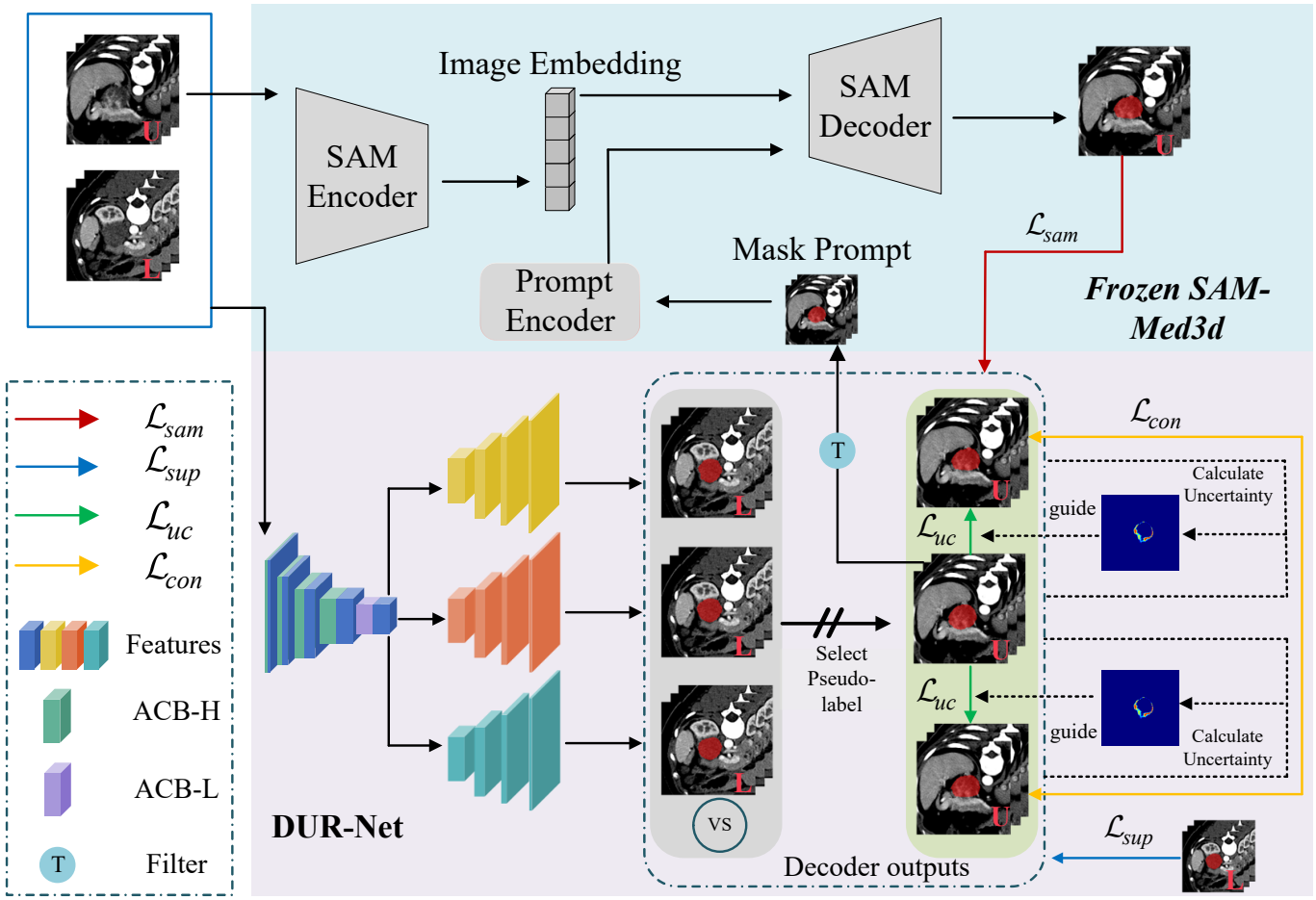


Fig. 2. Overview of the proposed DUR-Net+ framework, which consists of a frozen SAM-Med3D branch and a semi supervised model DUR-Net. Specifically, both labeled and unlabeled data are input into DUR-Net simultaneously. SAM-Med3D is a universal model that only processes unlabeled data. In the prediction of multiple decoders, the optimal decoder is selected by comparing the loss values of labeled data, and then the pseudo labels generated by the optimal decoder are self-rectified by the model. In this process, uncertainty estimation is introduced to emphasize the reliable parts of the pseudo labels. This is the dynamic uncertainty rectification process in DUR-Net. In addition, to compensate for the insufficient self-rectification ability of the DUR-Net, additional supervision information is generated using the prior knowledge of SAM-Med3D to guide DUR-Net. In order to improve the attention to the target tumor area, we also designed attention convolution blocks (ACB, including ACB-H and ACB-L) in the encoding stage of DUR-Net for extracting local and global features. The details related to ACB are shown in Fig. 3 and Fig. 4.

challenges such as low contrast, irregular lesion shape, blurred boundaries, and metal artefacts in images. Azad et al. [32] employed deformable convolution in conjunction with large kernel attention in order to enhance the capacity to capture objects with irregular shape and size features. The DA-Block [33] integrates position-based and channel-based components, enabling the learning from disparate perspectives and thus facilitating more accurate and detailed feature extraction.

In contrast to the aforementioned approaches that employ self-attentive blocks, we propose a more efficient methodology for capturing location information, with the aim of enhancing the feature representation of the network. In consideration of the variability in feature resolution that occurs during the encoding phase, we have devised two distinct Attentional Convolution Blocks, which permit the flexible extraction of features at varying scales.

III. METHODOLOGY

In semi-supervised segmentation, we assume n labeled images denoted as $D_l = \{(x_1^l, y_1), \cdots, (x_n^l, y_n)\}$, where y_i is the label for x_i^l. Also, there are m unlabeled images denoted as $D_u = \{x_1^u, \cdots, x_m^u\}$, where $n \ll m$. The objective of semi-supervised segmentation is train a robust segmentation model with the joint exploration of the labeled and unlabeled data.

An overview of the proposed framework named DUR-Net+ is depicted in Fig. 2 , which consists of a semi-supervised segmentation model, named DUR-Net, and the SAM-Med3D branch. DUR-Net includes a shared encoder and multiple independent decoders. The shared encoder f_{θ_e} and each decoder form a sub-model, can be defined as follows:

$$f_{\theta^i_{sub}} = f_{\theta_e} \bigodot f_{\theta^i_d}, \ i \in 1, \ldots, n \tag{1}$$

where each sub-model is a standard encoder-decoder architecture like V-Net, as shown in Fig. 3. Considering the efficiency, n is set to 3 in this paper. The three sub-models differ only in their upsampling methods, which are set as transposed convolution, linear interpolation, and nearest-neighbor interpolation, respectively. This design enhances the diversity within the m-

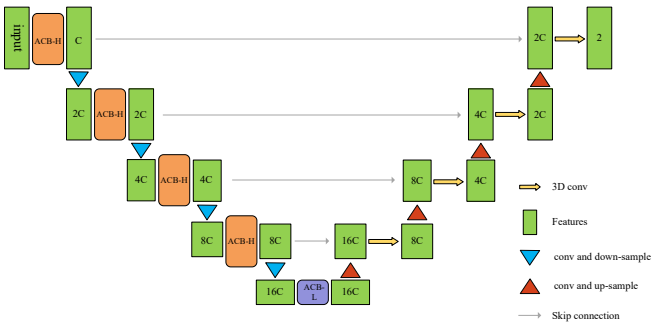


Fig. 3. Illustration of baseline architecture adapted from V-Net [10] with ACB module.

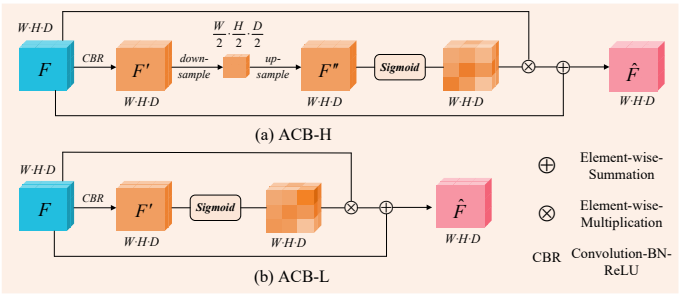


Fig. 4. The Attentional Convolution Block has two variants: ACB-H for processing high-resolution features and ACB-L for processing low-resolution features.

odel. Apart from this, the network layer structures of all sub-models remain consistent.

To mitigate the interference caused by blurred boundaries, irregular shapes, and low contrast, we designed an Attentional Convolution Block (ACB) in the encoding phase of DUR-Net. This block enhances the learning of global and local features through targeted processing of features at different resolutions, improves the model's ability to focus on target regions. Furthermore, to fully utilize unlabeled data, we proposed a dynamic uncertainty rectification mechanism in DUR-Net. Specifically, by real-time comparison of the outputs of three independent decoders, the decoder with the best performance is selected to generate pseudo-labels for model self-rectification. Then, an uncertainty estimation mechanism is introduced to emphasize highly reliable information in pseudo-labels, thereby reducing noise interference.

In addition to model self-rectification, we further introduced additional supervisory information to strengthen the rectification effect, which is achieved by integrating prior knowledge from SAM-Med3D. The pre-trained data of SAM-Med3D covers various abdominal organ structures such as kidneys and liver, and its 3D feature extraction capability can capture anatomical information around pheochromocytomas, providing the model with additional spatial structure supervisory signals. This mechanism can effectively compensate for potential semantic ambiguity of the dynamic rectification mechanism in complex boundary scenarios, ultimately improving the segmentation performance of the model. In response to the deficiency of manually providing prompt information for SAM-Med3D, we cleverly utilize the pseudo labels generated by DUR-Net as mask prompts, which can be achieved without introducing additional calculations.

A. Attention Convolution Block (ACB)

In the pheochromocytoma segmentation, the presence of blurred boundaries, irregular shapes, and low contrast can lead to inaccurate judgement of the boundary between the tumor and surrounding organs. We need to introduce a mechanism to focus on the region of interest, so as to ignore the influence of irrelevant areas. In this paper, we design an Attention Convolution Block (ACB) that enables the framework to cope with variations in tumour size, shape and location.

The ACB appears at the beginning of each stage of the encoder and receives the output (or input image) of the previous stage as input. As shown in Fig.4, there are two variants of ACB, ACB-H applied at high resolution and ACB-L applied at low resolution. Let $F \in \mathcal{R}^{C_i \times W \times H \times D}$ and $\hat{F} \in \mathcal{R}^{C_o \times W \times H \times D}$ represent the input and output feature maps of ACB respectively, $W \times H \times D$ denotes the spatial dimensions. The ACB-H initially processes feature map F through convolutional layers with kernel sizes of 1 and 3, respectively. Each convolution is followed by a batch normalization layer and a ReLU activation layer. This sequence allows the model to learn tumor features across varying receptive fields, enhancing local feature representation. Next, a downsampling layer reduces the resolution of the feature maps, followed by pointwise convolution to map interactions between channels. This step not only strengthens the correlation between channels but also improves the model's ability to capture global contextual information and establish dependencies between distant features. As a result, the model is better equipped to detect complex tumors at a larger scale. Afterward, the feature resolution is restored using upsampling to generate F'', which is passed through a Sigmoid activation function to produce the attention coefficient α. In this work, downsampling is performed using an average pooling layer with a stride and kernel size of 2, while upsampling is achieved via trilinear interpolation. The resulting attention coefficients are then multiplied with the original feature map F to increase the network's attention to the region of interest. After assigning different attentional weights to all voxels in F, they are summed with F to obtain $\hat{F}$, as expressed by the formula:

$$\hat{F} = (1 + \alpha)F. \tag{2}$$

However, at the bottleneck of the segmentation network, the feature map resolution becomes very low. Continuing to utilize ACB-H may lead to the loss of significant high-level semantic features. To address this, we propose ACB-L, a variant of ACB-H. In ACB-L, the downsampling and upsampling layers are removed, while the remaining operations remain consistent with those of ACB-H.

B. Dynamic Uncertainty Rectification

Generating pseudo labels for unannotated images has proven to be an effective strategy in semi-supervised medical image segmentation [20]. The critical challenge lies in improving the quality of these pseudo labels, as using low-quality ones for rectification can severely degrade the model's performance

[34]. A commonly used approach is to select the best performer among multiple options as the pseudo label, filtering out outputs with sub-optimal performance. For instance, Wang et al. [35] evaluated the performance of sub-networks using labeled data, and the most effective sub-network was subsequently used to generate pseudo labels. Wu et al. [36] compared the prediction results of peer networks and selected pixels with the highest probability of being in the foreground as pseudo labels. Although these strategies improve the quality of pseudo labels to some extent, the lack of ground truth makes pseudo labels still contain noise.

To address this, we propose a dynamic uncertainty rectification mechanism aimed at further optimizing the rectification process. The mechanism consists of two parts: first, selecting the optimal feature representation from the various decoder outputs as the pseudo label; and second, incorporating uncertainty estimation to mitigate the impact of noise within the pseudo label.

Algorithm 1 Dynamic Selection of Pseudo-label.

Require:
The set of labeled volumes for a batch:$\{x^l, y^l\}$
The set of unlabeled samples for a batch: $\{x^u\}$
The Decoder : $\mathcal{D}_1(\cdot)$, $\mathcal{D}_2(\cdot)$, $\mathcal{D}_3(\cdot)$
The Dice loss function: $\mathcal{L}_{dice}(\cdot)$
Ensure:
The pseudo labels of unlabeled samples for a batch, y_p^u

$x \leftarrow \{x^l, x^u\}$
$dice_1 \leftarrow 0;\ dice_2 \leftarrow 0;\ dice_3 \leftarrow 0$
$\hat{y}_1^l,\ \hat{y}_1^u \leftarrow \text{softmax}(\mathcal{D}_1(x))$
$\hat{y}_2^l,\ \hat{y}_2^u \leftarrow \text{softmax}(\mathcal{D}_2(x))$
$\hat{y}_3^l,\ \hat{y}_3^u \leftarrow \text{softmax}(\mathcal{D}_3(x))$
for each $y_i^l \in y^l$ and $\hat{y}_{1,i}^l \in \hat{y}_1^l$ and $\hat{y}_{2,i}^l \in \hat{y}_2^l$ and $\hat{y}_{3,i}^l \in \hat{y}_3^l$ **do**
$dice_1 \leftarrow dice_1 + \mathcal{L}_{dice}\left(\hat{y}_{1,i}^l, y_i^l\right)$
$dice_2 \leftarrow dice_2 + \mathcal{L}_{dice}\left(\hat{y}_{2,i}^l, y_i^l\right)$
$dice_3 \leftarrow dice_3 + \mathcal{L}_{dice}\left(\hat{y}_{3,i}^l, y_i^l\right)$
end for
if $dice_1 < dice_2$ and $dice_1 < dice_3$
$y_p^u \leftarrow \hat{y}_1^u$
elif $dice_2 < dice_1$ and $dice_2 < dice_3$
$y_p^u \leftarrow \hat{y}_2^u$
elif $dice_3 < dice_1$ and $dice_3 < dice_2$
$y_p^u \leftarrow \hat{y}_3^u$
end if
return y_p^u

1) Dynamic selection of pseudo labels : Unlike approaches that select pseudo labels across multiple sub-networks, our method selects pseudo labels across multiple independent decoder outputs that share the same encoder. This design offers three key advantages. First, it reduces characterization discrepancies that may arise from different network parameters at the encoder stage. Second, it introduces constraints across different learners, preventing them from converging in opposite directions. Third, the multi-decoder setting provides diverse perspectives and different levels of information, helping the model better understand intricate tumor characteristics and reduce biases that may arise from relying on a single branch.

Algorithm 1 describes the dynamic selection process for generating pseudo labels. Following [35], we utilize Dice loss to assess the performance of each decoder in real-time, selecting the best-performing branch as the pseudo label generator. Note that the generator consists of a shared encoder and a independent decoder. Dice loss is computed for each batch of labeled data, facilitating the evaluation of each branch's performance in the segmentation task. We argue that the best-performing branch is more reliable in processing unlabeled data, making its predictions suitable as pseudo labels.

Throughout the training process, pseudo labels are not generated by fixed branches; instead, each branch has the opportunity to serve as a pseudo label generator. This design enables each branch to leverage its unique strengths at different training stages, thereby capturing diverse feature representations. When one branch excels in generating pseudo labels, the others can be trained using these pseudo labels to further improve their performance.

2) Uncertainty guided rectification : For the acquired pseudo labels, we can use them to rectify the suboptimal decoder, consistency regularization loss is used to implement this process:

$$\mathcal{L}_{cr} = \frac{1}{K}\sum_{k=1}^{K} \|P_l - \tilde{P}^k\|_2^2, \tag{3}$$

where P_l represents the pseudo labels, $\tilde{P}^k$ represents the output of the kth sub-optimal decoder, K is the number of sub-optimal decoders and its value is set to 2. However, directly using pseudo labels to participate in model rectification is not optimal because without the guidance of ground truth labels, pseudo labels may be unreliable and noisy, which can affect model training. We propose to introduce uncertainty estimation to emphasise reliable predictions. Inspired by the [22] , KL divergence between the optimal decoder and sub-optimal decoders is used as a measure of uncertainty:

$$D_i^k = \sum_{i=0}^{n_v} \tilde{P}_i^k \cdot \log \frac{\tilde{P}_i^k}{P_i}, \tag{4}$$

where D_i^k denotes the KL divergence between the outputs of the kth sub-optimal decoder and the optimal decoder, n_v indicates the number of voxels, $\tilde{P}_i^k$ denotes the prediction of the ith voxel in the output of the kth suboptimal decoder, and P_i denotes the prediction of the ith voxel in the pseudo label. For a given voxel i, a larger value D_i^k indicates the prediction for that pixel is far from the pseudo label, i.e., with high uncertainty. Then, we use $e^{-D_i^k}$ as the uncertainty factor to adaptively adjust weight in the consistency regularization loss, the uncertainty loss can eventually be expressed as follows:

$$\mathcal{L}_{uc} = \frac{1}{K}\sum_{k=1}^{K} e^{-D_i^k} \cdot \left\|P_i - \tilde{P}_i^k\right\|_2^2 + \frac{1}{K}\sum_{k=1}^{K} D_i^k. \tag{5}$$

By introducing uncertainty estimation, voxels with lower uncertainty are given a higher weight in the uncertainty loss,

while voxels with higher uncertainty are given a lower weight. With this rectification, the model can learns more reliable information and reduces the effect of noise. Following [28], we incorporate the uncertainty term $\sum_{k=1}^{K} D_i^k$ into the constraints, which can enhance the model's robustness.

Furthermore, for the two suboptimal outputs, they originate from different decoders, which generates perturbations at the decoder level. By maintaining consistency between them, the framework's resilience to perturbations can be further improved. We implement this constraint through the cosine distance:

$$\mathcal{L}_{con} = 1 - \cos(\tilde{P}_1, \tilde{P}_2), \tag{6}$$

$$\cos(\tilde{P}_1 \cdot \tilde{P}_2) = \frac{\tilde{P}_1 \cdot \tilde{P}_2}{\left|\left|\tilde{P}_1\right|\right|_2 \cdot \left|\left|\tilde{P}_2\right|\right|_2}. \tag{7}$$

Note that since the optimal decoder is not fixed, the suboptimal decoder also has the same dynamic nature. This nature enables flexible adjustment of the similarity limits between the decoders, which further promotes synergy between the decoders and helps the model learn more complementary information.

C. Introducing prior knowledge of SAM-Med3D

The prior knowledge of the visual foundation model plays a unique role in enhancing or supplementing semantic information [37], [38]. For SAM-Med3D [8], it is able to recognize a wide range of semantic representations due to the fact that SAM-Med3D had been extensively trained on different medical data and its semantic annotations cover a wide range of anatomical structures. This rich prior knowledge enables SAM-Med3D to provide more comprehensive information when processing specific medical images, thus improving the model's ability to understand complex structures. Therefore, we use SAM-Med3D as an additional supervised branch to process unlabeled data using its prior knowledge. Specifically, we pass the unlabeled pheochromocytoma data through SAM-Med3D to get a valuable insight and use that insight as a supervisory signal to rectify the framework:

$$\mathcal{L}_{sam} = \frac{1}{K} \sum_{k=1}^{K} \sum_{j=1}^{m} Dice(\tilde{P}_j^k, P_j^s). \tag{8}$$

where $\tilde{P}_j^k$ represents the prediction of the suboptimal decoder, P_j^s is the SAM-Med3D inference result. and m is the number of unlabeled data.

Like models [6], [39] that require manual prompts, SAM-Med3D also requires expert prompts. This reliance on manually generated prompts clearly hinders its clinical application. In this paper, we use pseudo labels generated by the semi-supervised segmentation model to generate prompts. Since coarse segmentation contains noise, it needs to be filtered to retain reliable prompts. Specifically, we select voxels with larger foreground probability values in the pseudo labels; higher foreground probabilities indicate that the corresponding predictions are more reliable and play a greater role in pheochromocytoma segmentation.

$$\hat{P}_i = \begin{cases} P_i, & \text{if } P_i \geq T \\ 0\,, & \text{otherwise} \end{cases}. \tag{9}$$

where T is the threshold to determine the high-quality mask prompt, P_i denotes the prediction of the pseudo-label on the ith voxel. $\widehat{P}_i$ is the mask prompt information required by SAM-Med3D. In this way, low-confidence voxels prone to misclassification were filtered out, eliminating the need for manual annotation.

IV. Overall loss function

For training with labeled samples, we adopt a combination of Dice loss [35] and Focal loss [40]:

$$\mathcal{L}_{sup} = \alpha \mathcal{L}_{dice} + (1 - \alpha) \mathcal{L}_{focal} \tag{10}$$

$$\mathcal{L}_{focal} = (1 - P)^{\gamma} \mathcal{L}_{ce} \tag{11}$$

where α is the coefficient to balance these two loss terms. P is the prediction of decoder. The scaling factor $(1 - P)^{\gamma}$ reduces the contribution of easy pixels and focuses training on hard ones. Whenever γ is set to 0 the Focal loss will be equal to the Cross Entropy loss.

The total training loss function of the network is computed as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda_1 \mathcal{L}_{uc} + \lambda_2 \mathcal{L}_{sam} + \mathcal{L}_{con} \tag{12}$$

where λ_1 and λ_2 are used to balance the $\mathcal{L}_{uc}$ and $\mathcal{L}_s$. It should be noted that $\mathcal{L}_{uc}$ and $\mathcal{L}_{con}$ were applied to both labeled and unlabeled datasets.

V. Experiment and result analysis

A. Dataset and Preprocessing

The PUMCH-PHEO dataset was collected by Peking Union Medical College Hospital and includes 100 abdominal CT enhanced scan samples, all of whom are patients with abdominal pheochromocytoma. The maximum diameter range of the tumor is 1-16 cm, with 48 cases in the left abdomen and 52 cases in the right abdomen. The annotation work was independently completed by two radiologists with more than 10 years of experience. The areas of disagreement were arbitrated by the chief physician and finally reviewed by senior experts. The average Dice coefficient between groups was 0.912. During the preprocessing process, the samples were isotropically resampled to 1 × 1 × 1 mm³, with window width and level adjusted to 255 Hu/127.5 Hu. Based on label positions, they were cropped into 128 × 128 × 128 voxels (10-20 voxels for edge axis/coronal plane extension, 5-10 voxels for sagittal plane extension), and zero mean unit variance normalization was performed.

The JCH-PHEO dataset is sourced from Jiangmen Central Hospital and includes 128 abdominal CT samples with a maximum tumor diameter range of 1.5-14 cm. The annotation was completed by three radiologists, and the divergent areas were arbitrated by the chief urologist. The initial inter group Dice coefficient was 0.895, which was increased to 0.932 after arbitration. The preprocessing process is consistent with PUMCH-PHEO dataset, including resolution resampling, window width and level adjustment, region of interest cropping, and normalization.

B. Experimental details and evaluation metrics

For all experiments, we adopted the SGD optimizer with a learning rate 0.01 and a weight decay factor 0.0001 for training. The batch size was set to 4, containing two labeled data and two unlabeled data. The weights α was set to 0.5, the loss coefficients λ_1 and λ_2 are time-dependent Gaussian warming function [9], $\lambda(t) = w_{\max} \cdot e^{(-5(1-\frac{t}{t_{\max}})^2)}$, $w_{\max}$ was set to 0.1. We applied the random cropping, rotation and flipping as data augmentation and use the V-Net [10] architecture as backbone on all datasets. The total training iteration $t_{\max} = 10k$. During the testing phase, only the first decoder output was used as the final prediction. All experiments were conducted in the same environment with hardware comprising an NVIDIA Quadro GV100 and software configured with PyTorch 2.1.0 + CUDA 12.0. The random seed was set to 1337. For the evaluation metrics, following previous work [35], [36], we use the Dice, Jaccard, Average Surface Distance (ASD), and 95% Hausdorff Distance (95HD).

C. Results

We compare the proposed framework with the state-of-the-art (SOTA) semi-supervised segmentation methods. These methods include MT [23], UA-MT [41], EM [26], DTC [24], URPC [28], MC-Net+ [9], ACMT [42]. Note that all groups use the same training strategies and backbone model for fair comparisons.

1) Performance on PUMCH-PHEO dataset: We present the quantitative comparison results under three label ratios in Table I. Compared to the baseline, the proposed DUR-Net+ achieves impressive improvement in all four metrics. Specifically, when trained with only 10% labeled data, Dice increases from 32.64% to 83.36%, Jaccard boosts from 26.32% to 75.05%, 95HD drops from 19.14% to 6.82%, ASD drops from 7.35% to 2.49%. A similar situation is observed when the framework is trained with only 5% or 20% of the labeled data. The results demonstrate that the proposed framework can effectively utilize unlabeled data to improve segmentation performance. In comparison to the SOTA models, the proposed DUR-Net+ also outperforms other comparison methods in three semi-supervised settings, highlighting the superiority of our method. We attribute this to three aspects: first, we evaluate the pseudo labels generated by dynamic selection so that voxels with higher reliability have greater weight in rectification, thereby suppressing noise. Second, by using pseudo labels as prompts, SAM-Med3D provides rich supervised information for semi-supervised segmentation models, enhancing the exploitation of unlabeled data. Third, the feature learning in ACB allows the framework to automatically focus on regions of interest and reduce the influence of irrelevant regions.

When SAM-Med3D was removed, DUR-Net also achieved comparable results. In all settings, the proposed method yields the best segmentation Dice, the best Jaccard, and the best or second best HD95 or ASD performance. In contrast to MC-Net+ [9], which converts all decoder outputs into pseudo labels for rectification, the proposed method delivers more than 8% improvement in Dice and Jaccard metrics. Upon adding SAM-Med3D, we found that regardless of the label ratio, the most notable changes occurred in the 95HD and ASD metrics. For example, both 95HD and ASD have been reduced by more than 68% with 20% labeled data. This indicates the strengthened capability of the framework to delineate tumor boundaries. We analyze the reason for this: SAM-Med3D prior knowledge is obtained by training on large-scale medical datasets (including abdominal multi-organ data) with semantic annotations covering a wide range of anatomical structures. This gives DUR-Net+ an distinct advantage in distinguishing pheochromocytoma and abdominal multi-organs. In contrast, just learning the pheochromocytoma features does not help the model to improve the segmentation of the boundary. As seen in Table I, existing semi-supervised methods exhibited consistent poor performance in both 95HD and ASD metrics, suggesting their predictions suffer from serious over-segmentation or under-segmentation issues. We think that the close geometric relationship between pheochromocytoma and surrounding abdominal organs, causing weak boundary situations that affect boundary judgment.

Fig.5 shows the visualization results of different methods trained with 20% labeled data. In the first row of samples, methods such as EM and UAMT failed to identify the target region of pheochromocytoma completely due to the effect of low contrast, and even URPC, which aggregates multi-scale features, showed more serious over-segmentation.The MT and ACMT methods, although showing better boundary segmentation, suffer from the problem of false positives. The DUR-Net+ method proposed in this paper, on the other hand, effectively copes with the challenge of low contrast and achieves complete segmentation of the target region. Compared with the first row of samples, the degree of low contrast in the second row of samples was reduced, and thus most methods achieved more reliable segmentation results, with only the DTC method underperforming. In the third row of samples, the close proximity of the pheochromocytoma to the surrounding organs makes it difficult for most methods to accurately identify and separate the target region during segmentation. With the guidance of SAM-Med3D, DUR-Net+ was able to effectively reduce the occurrence of false alarms, while reaching a complete segmentation of the target region and significantly alleviating the weak boundary conflict problem. Overall, compared with other semi-supervised segmentation methods, DUR-Net+ is able to simultaneously overcome the challenges posed by low contrast as well as interference from surrounding organs, enhancing the segmentation of pheochromocytoma.

2) Performance on JCH-PHEO dataset: Our method is further evaluated on another central dataset with 10 and 20 labeled samples for semi-supervised training settings. As shown in Table 2, the DUR-Net+ method proposed in this paper still achieves significant improvements in all four metrics compared to the current frontier models. Of interest, the ACMT method still performs the best among the existent methods and even outperforms DUR-Net in two coefficients, Dice and Jaccard coefficients, at 10% setting, suggesting that the focus on ambiguously challenging regions promotes pheochromocytoma segmentation. In addition, compared to

TABLE I

QUANTITATIVE COMPARISON OF THE SOTA SEMI-SUPERVISED METHOD WITH OUR METHOD ON THE PUMCH-PHEO DATASET.* DENOTES ASTATISTICAL SIGNIFICANCE TEST BETWEEN OUR METHOD AND THE COMPARATIVE METHOD ($P < 0.05$). RED-COLORED AND BLUE-COLORED VALUES CORRESPOND TO THE BEST AND 2ND BEST PERFORMING MODEL, RESPECTIVELY.

Method	scans used		Metrics				Cost
	Labeled	Unlabeled	Dice[%]↑	Jaccard[%]↑	95HD[voxel]↓	ASD[voxel]↓	Params.(M)
V-Net	4(5%)	0	30.85*	24.80*	27.50*	11.59*	9.45
V-Net	8(10%)	0	32.64*	26.32*	19.14*	7.35*	9.45
V-Net	16(20%)	0	60.94*	50.69*	14.18*	5.55*	9.45
V-Net	80(All)	0	89.73	82.30	4.55	1.53	9.45
EM [26]	4(5%)	76	51.74*	41.01*	27.53*	14.40*	9.45
MT [23]	4(5%)	76	35.71*	26.62*	40.09	22.23*	9.45
UAMT [41]	4(5%)	76	44.67*	34.41*	35.39*	14.13	9.45
DTC [24]	4(5%)	76	32.26*	23.17*	40.80*	18.51*	30.45
URPC [28]	4(5%)	76	40.72*	28.97*	41.04	20.15*	9.45
MC-Net+[9]	4(5%)	76	33.17*	25.46*	32.22*	18.68	15.25
ACMT [42]	4(5%)	76	45.45*	35.51*	37.96*	19.65*	5.89
DUR-Net	4(5%)	76	55.50	44.09	22.64	9.92	18.37
DUR-Net+	4(5%)	76	74.77	64.36	6.83	2.80	18.37
EM [26]	8(10%)	72	55.15*	43.25*	23.57*	11.21*	9.45
MT [23]	8(10%)	72	52.08*	40.52*	31.20*	12.70	9.45
UAMT [41]	8(10%)	72	51.60*	41.86*	25.94*	12.01*	9.45
DTC [24]	8(10%)	72	49.07*	37.48*	16.75*	6.34*	30.45
URPC [28]	8(10%)	72	50.06*	36.56*	36.07*	15.76	9.45
MC-Net+[9]	8(10%)	72	64.30*	54.66*	26.39*	10.42*	15.25
ACMT [42]	8(10%)	72	65.94	54.80	25.38*	5.92*	5.89
DUR-Net	8(10%)	72	73.63	63.30	14.81	6.00	18.37
DUR-Net+	8(10%)	72	83.36	75.05	6.82	2.49	18.37
EM [26]	16(20%)	64	75.80*	64.84*	20.64*	7.20	9.45
MT [23]	16(20%)	64	76.43*	64.70*	15.39*	6.02*	9.45
UAMT [41]	16(20%)	64	77.17*	66.37	15.09*	6.86*	9.45
DTC [24]	16(20%)	64	75.33*	63.92*	14.39*	6.24*	30.45
URPC [28]	16(20%)	64	74.95	65.81*	22.07*	9.17*	9.45
MC-Net+[9]	16(20%)	64	76.14*	67.03	21.05*	8.34*	15.25
ACMT [42]	16(20%)	64	80.05*	70.04*	11.67*	5.34*	5.89
DUR-Net	16(20%)	64	85.92	79.07	12.29	4.47	18.37
DUR-Net+	16(20%)	64	86.54	79.93	4.97	1.27	18.37

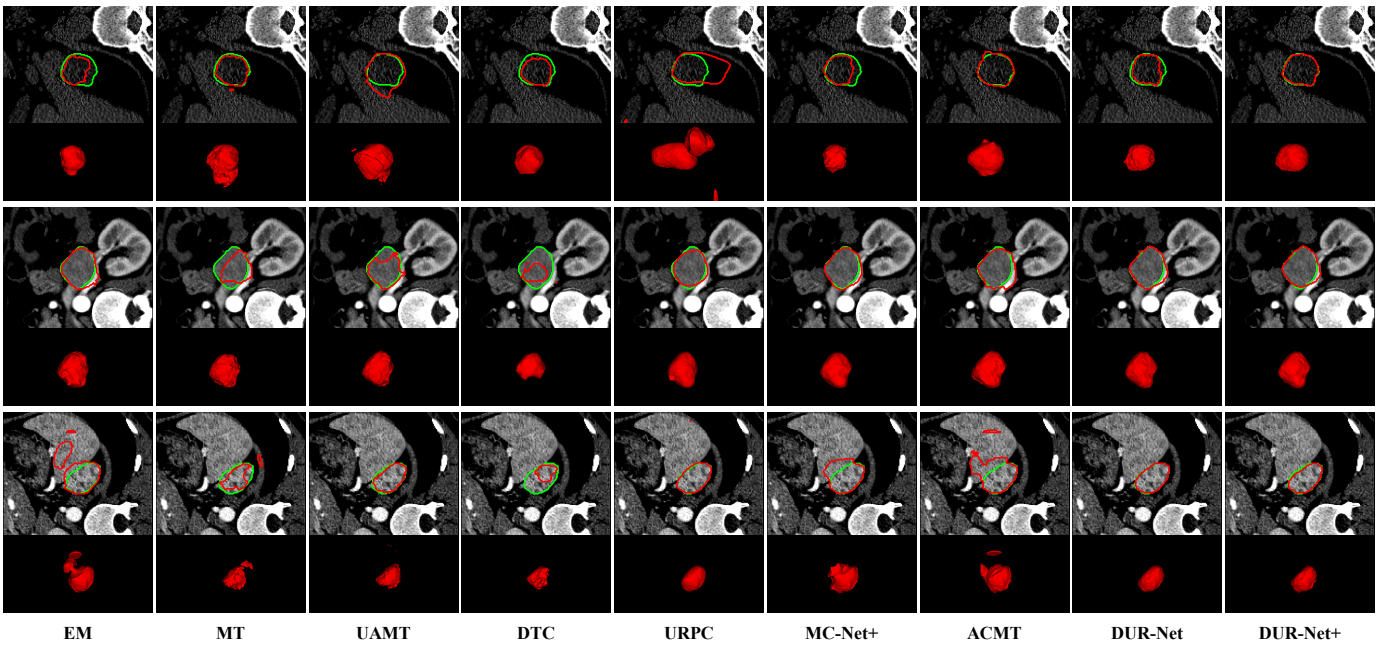


Fig. 5. Qualitative comparisons on the PUMCH-PHEO dataset with 20% labeled cases. The areas outlined in red and green represent the predicted results and ground truth (GT), respectively.

DUR-Net , DUR-Net+ shows statistically significant improvement in all evaluation metrics, which fully demonstrates the effectiveness of SAM-Med3D a priori knowledge in enhancing the segmentation capability of the model. In particular, the enhancement of model performance using supervised signals from SAM-Med3D at 10% annotation ratio is significantly higher than that at 20% annotation ratio, a phenomenon that is mutually corroborated with the patterns presented in Table 1, revealing the outstanding advantages of this module in low annotation data scenarios. In addition, we present some of the qualitative comparison results in Fig.6. It is found that despite the fact that pheochromocytomas present fuzzy boundaries and diverse morphologies in different populations, our proposed DUR-Net+ method can accurately outline their boundaries and generate plausible segmentation results.

D. Ablation analysis

1) Experiments on the component: As shown on Table III, we conducted ablation experiments on the PUMCH-PHEO dataset to validate the performance of each component. Based on the shared encoder and three independent decoders, we added the components including Attention Convolution Block (ACB), Dynamic Uncertainty Rectification (DUR), and SAM-Med3D. It can be noticed that each component helps to improve the segmentation performance. The Dice score improves by 8.82% with the addition of ACB, which indicates that the region of interest is getting attention through ACB. In the absence of ACB, the utilisation of DUR to select pseudo-labels and assign elevated weights to plausible voxels resulted in an improvement of nearly 9% in the Dice metric. However, the ACB and DUR play a role more in terms of target area. SAM-Med3D not only enhances the segmentation of the target region, but also improves the discrimination of the tumor boundary. After adding SAM-Med3D, the Dice increased by 13.87%, the Jaccard by 14.36%, the 95HD dropped by 5.75%, and the ASD dropped by 1.58%. These findings underscore the efficacy of SAM-Med3D's prior knowledge in enhancing the framework's holistic performance. Moreover, the integration of SAM-Med3D into either the ACB or DUR modules yields consistent improvements across all four evaluation metrics, thereby substantiating the rectified role of SAM-Med3D. Notably, experimental results further reveal that synergistic utilization of all components culminates in optimal performance.

To further demonstrate the effectiveness of ACB, DUR and SAM-Med3D, we generate the feature maps of unlabeled samples with different morphologies and backgrounds in Fig.7. It can be seen that ACB brings attention to the region of interest. DUR highlights the semantic representation and reduces noise by learning reliable information. With the introduction of SAM-Med3D, the framework's understanding of the overall structural relationships is enhanced, enabling a clear distinction between the tumor and surrounding tissue. In addition, by integrating all components, complementary effects can be achieved to further improve the focus on the tumor region and the recognition accuracy.

2) Compared with prompt-based inference of Sam-Med3D: As shown in Table IV, the segmentation performance of DUR-Net improves with the increase in labeled data, while SAM-Med3D enhances its Dice coefficient from 62.50% to 78.23% by increasing the number of manual prompts. This indicates that although SAM-Med3D enables zero-shot segmentation without labeled training data, achieving acceptable performance requires generating prompts for each test image, leading to substantial manual effort. In contrast, DUR-Net+ significantly outperforms individual models at different labeling ratios by integrating the advantages of DUR-Net and SAM-Med3D, fully verifying the performance enhancement effect of collaborative learning.

3) Comparison with other attention modules: The feasibility of ACB is investigated through a comparative analysis with Squeeze and Excite (SE) [43], Large Kernel Attention (LKA) [44], Global Attention Mechanism (GAM) [45], and Convolutional Block Attention Module (CBAM) [46]. For fair comparison, all attention modules were inserted into the same position in the encoder. As shown in Table V, the LKA with large kernel convolutional attention achieves the best performance on the Dice and Jaccard metrics among the existing attention modules. The CBAM, with the cascading optimization of channel and spatial attention, has stronger discriminative power on the spatial position and feature dimension of boundary pixels, and therefore performs the best on 95HD and ASD. However, ACB is significantly better than CBAM in Dice and Jaccard coefficients (up 6.8% and 7.6% respectively), mainly because its feature fusion strategy can more fully capture the semantic information of the tumor body region. It is worth noting that although ACB is inferior to CBAM in boundary indicators, it still maintains the second best performance, with a difference of only 0.16 for 95HD and 0.26 for ASD. In contrast, CBAM's advantage in Dice and Jaccard coefficients is not significant. Overall, ACB produces the best performance.

4) Decoders inter-rectification: In the dynamic uncertainty rectification mechanism, pseudo labels are selected and all decoders are encouraged to learn from each other during training, which in turn improves the overall performance of the framework. Figure.8 illustrates the prediction outcomes of three autonomous decoders across distinct iterations. Initially, notable discrepancies exist between the decoders. And as the iterations advance, they rectify each other and ultimately converge towards a consensus.

5) The effects of $\mathcal{L}_{uc}$ and $\mathcal{L}_{con}$: Two types of regularization are introduced in dynamic uncertainty rectified, which include uncertainty rectification between pseudo labels and suboptimal predictions, and consistency regularization between suboptimal predictions. The quantitative results in Table VI show that both rectifications improve the performance. Uncertainty rectification produces a larger contribution than consistency regularisation. The combination of these two operations can gives the best results.

6) Comparison with different reliability measures: We evaluated the effectiveness of our dynamic prompt generation method by conducting comparative experiments with methods that only use the output of a single decoder as prompt. To

TABLE II

QUANTITATIVE COMPARISON OF THE SOTA SEMI-SUPERVISED METHOD WITH OUR METHOD ON THE JCH-PHEO DATASET.* DENOTES ASTATISTICAL SIGNIFICANCE TEST BETWEEN OUR METHOD AND THE COMPARATIVE METHOD ($P < 0.05$). RED-COLORED AND BLUE-COLORED VALUES CORRESPOND TO THE BEST AND 2ND BEST PERFORMING MODEL, RESPECTIVELY.

Method	scans used		Metrics				Cost
	Labeled	Unlabeled	Dice[%]↑	Jaccard[%]↑	95HD[voxel]↓	ASD[voxel]↓	Params.(M)
V-Net	10(10%)	0	23.29*	16.72*	30.21*	13.26*	9.45
V-Net	20(20%)	0	47.83*	40.36*	15.32*	4.30*	9.45
V-Net	100(All)	0	78.52	70.92	11.84	3.13	9.45
EM [26]	10(10%)	90	35.02*	27.56*	24.06*	6.36*	9.45
MT [23]	10(10%)	90	37.53*	27.37*	25.70	8.87*	9.45
UAMT [41]	10(10%)	90	40.87*	31.36*	25.23*	7.13*	9.45
DTC [24]	10(10%)	90	46.90*	34.88	20.76*	7.44*	9.45
URPC [28]	10(10%)	90	42.98	33.37*	28.34*	12.46	5.89
MC-Net+ [9]	10(10%)	90	49.19*	37.62*	19.42	6.43*	15.25
ACMT [42]	10(10%)	90	57.99*	47.55	22.36*	11.69*	30.45
DUR-Net	10(10%)	90	57.17	45.93	18.48	6.52	18.37
DUR-Net+	10(10%)	90	76.92	69.53	9.51	4.09	18.37
EM [26]	20(20%)	80	51.94*	41.73*	16.78*	5.72	9.45
MT [23]	20(20%)	80	55.71*	43.62*	18.30	4.54*	9.45
UAMT [41]	20(20%)	80	64.89*	54.61*	15.86	5.97*	9.45
DTC [24]	20(20%)	80	54.42*	41.71*	21.92*	6.12*	9.45
URPC [28]	20(20%)	80	60.28*	51.37*	26.30*	10.30*	5.89
MC-Net+ [9]	20(20%)	80	62.36*	49.85*	17.36	4.91*	15.25
ACMT [42]	20(20%)	80	71.31*	60.98*	10.34*	3.53*	30.45
DUR-Net	20(20%)	80	72.32	63.57	13.26	4.20	18.37
DUR-Net+	20(20%)	80	79.77	71.03	8.50	3.03	18.37

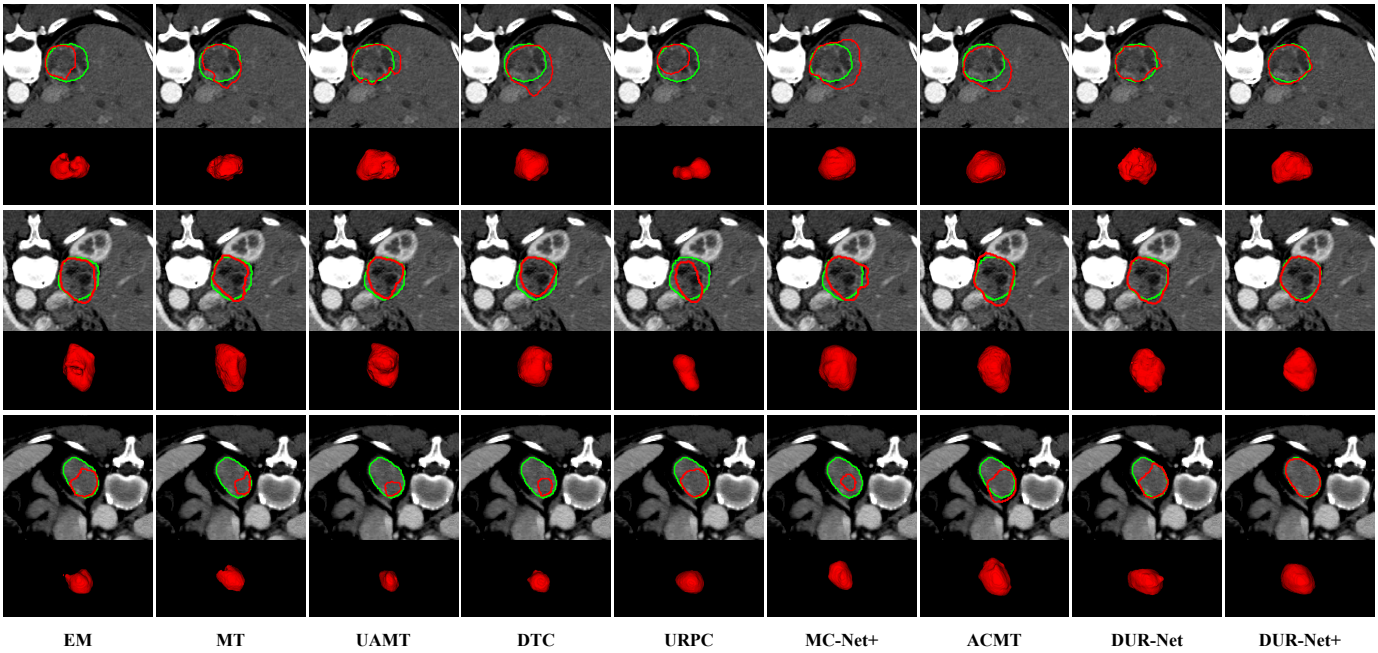


Fig. 6. Qualitative comparisons on the JCH-PHEO dataset with 10% labeled cases. The areas outlined in red and green represent the predicted results and ground truth (GT), respectively.

ensure fairness, a threshold of 0.85 was applied for filtering in all experiments. As shown by the results in Table VII, the dynamically generated prompt method demonstrates excellent overall performance, ranking first in the Dice, Jaccard, and 95HD metrics. This suggests that the dynamic adjustment of prompt enables the framework to capture a more comprehensive feature representation, reducing information redundancy and bias from fixed prompts, and thereby improving segmentation accuracy on complex tumors.

7) The Effects of hyperparameter T: It is important to note that the threshold T plays a critical role in obtaining high-quality mask prompts. A lower threshold may introduce noise into the mask prompts, resulting in false signals. Conversely, a higher threshold can lead to the loss of useful information and fail to provide adequate prompts. We conducted ablation experiments on the PUMCH-PHEO dataset to determine the optimal confidence threshold T. As illustrated in Fig.9, both excessively high and low thresholds yield sub-optimal perfor-

TABLE III
EFFECTIVENESS OF THE COMPONENTS ON THE PUMCH-PHEO DATASET

Scans Used		Module			Metrics			
labeled	unlabeled	ACB	DUR	SAM-Med3D	Dice(%)↑	Jaccard(%)↑	95HD(voxel)↓	ASD(voxel)↓
8(10%)	0				63.52	53.27	15.81	5.18
	72(90%)	✓			72.34	61.94	19.51	7.15
	72(90%)		✓		72.48	61.25	17.85	7.63
	72(90%)			✓	77.39	67.63	10.06	3.60
	72(90%)	✓	✓		73.63	63.30	16.81	6.72
	72(90%)	✓		✓	80.45	70.78	9.78	3.47
	72(90%)		✓	✓	80.52	70.97	7.06	2.62
	72(90%)	✓	✓	✓	**83.36**	**75.05**	**6.82**	**2.49**

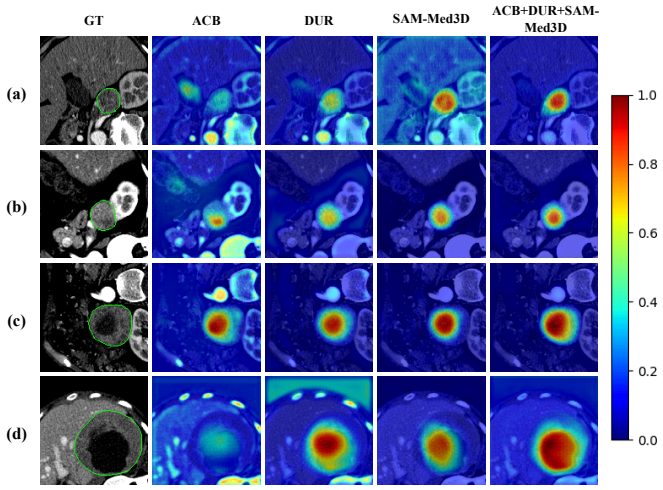


Fig. 7. Visualization of the feature map after adding different modules, the area outlined by the green line indicates pheochromocytoma.

TABLE IV
COMPARED WITH PROMPT-BASED INFERENCE OF SAM-MED3D

Method	Manual Annotation	Manual Prompt	Dice ↑[%]
DUR-Net	4labeled case	-	55.50
	8labeled case	-	73.63
	16labeled case	-	83.36
SAM-Med3D	-	4points	62.50
	-	8points	70.66
	-	16points	78.23
DUR-Net+	4labeled case	-	74.77
	8labeled case	-	83.36
	16labeled case	-	86.54

mance. The framework achieves its best performance when T is set to 0.85.

E. Limitations Discussion

In this paper, we integrated SAM-Med3D prior knowledge to facilitate semi-supervised pheochromocytoma segmentation, yielding significant performance gains. However, results on the PUMCH-PHEO dataset revealed a gradual attenuation of SAM-Med3D's efficacy with increasing labeling ratios. This is evident from marginal metric improvements when the labeling rate reaches 20%. The phenomenon stems from SAM-Med3D's limited knowledge of the target domain, as it was deployed for direct inference without thorough domain-specific training. While its robust generalization and feature extraction capabilities establish a high performance baseline, as demonstrated by substantial improvements with only 5% labeled data, the absence of pheochromocytoma-specific semantic understanding constrains its full potential. Thus,we

TABLE V
ABLATION EXPERIMENTS WITH DIFFERENT ATTENTION MODULES,RED-COLORED AND BLUE-COLORED VALUES CORRESPOND TO THE BEST AND 2ND BEST PERFORMING MODEL, RESPECTIVELY.

Module	Params(M)	Metrics			
		Dice[%]↑	Jaccard[%]↑	95HD[voxel]↓	ASD[voxel]↓
SE [43]	15.257	77.36	66.88	14.10	6.44
LKA [44]	15.744	79.49	69.67	8.15	2.75
GAM [45]	30.265	74.40	61.96	10.35	3.40
CBAM [46]	15.262	76.53	67.45	6.66	2.23
ACB (Ours)	18.370	83.36	75.05	6.82	2.49

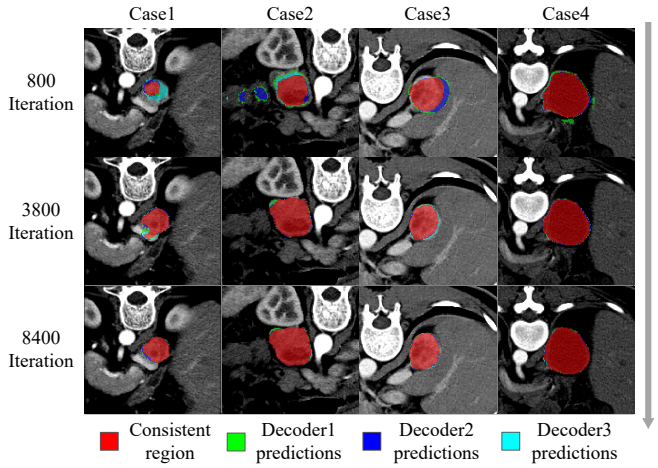


Fig. 8. Visualization of the prediction results of three decoders under different iteration conditions.

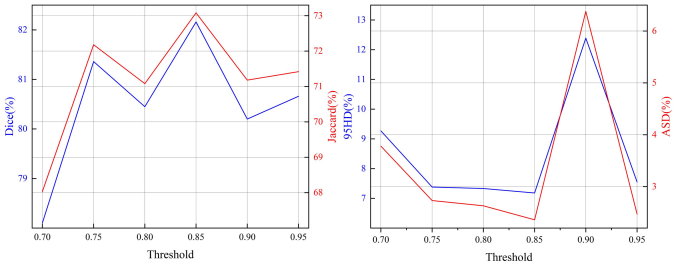


Fig. 9. Parameter analysis of confidence threshold on pheochromocytoma semantic segmentation.

TABLE VI
ABLATION STUDIES ON DIFFERENT LOSSES

$\mathcal{L}_{uc}$	$\mathcal{L}_{con}$	Dice(%)↑	Jaccard(%)↑	95HD(voxel)↓	ASD(voxel)↓
		80.45	70.78	9.78	3.47
✓		81.82	72.69	7.45	2.83
	✓	81.43	71.26	9.48	3.91
✓	✓	**83.36**	**75.05**	**6.82**	**2.49**

TABLE VII
PERFORMANCE OF DIFFERENT PROMPT METHOD

Method	Dice(%)↑	Jaccard (%)↑	95HD(voxel)↓	ASD (voxel)↓
Decoder1	79.83	70.20	6.90	**2.44**
Decoder2	81.57	72.55	7.54	2.47
Decoder3	80.49	70.69	10.05	3.35
Ours	**83.36**	**75.05**	**6.82**	2.49

believe that fine-tuning SAM-Med3D to acquire task-specific knowledge, thereby enhancing its supervisory role, constitutes a crucial direction for future research.

VI. Conclusion

This paper presents a semi-supervised framework for pheochromocytoma segmentation. Specifically, we first constructed the semi-supervised model DUR-Net, whose core designs are reflected in two aspects: Attentional Convolution Blocks (ACBs) are embedded in the encoding stage to enhance the learning ability of complex tumor features through differential processing of features with varying resolutions, thereby strengthening the model's focus on regions of interest; meanwhile, a dynamic uncertainty rectification mechanism is proposed to select pseudo-labels through dynamic comparison of multi-decoder outputs, enabling self-rectification of the model. On this basis, the framework incorporates prior knowledge from SAM-Med3D as an external supervisory signal, leveraging its rich semantic understanding of abdominal anatomical structures to form a synergy with the internal learning mechanism of DUR-Net, further optimizing segmentation accuracy. Experimental results on pheochromocytoma datasets from two different centers validate the effectiveness of the proposed semi-supervised segmentation framework, providing a feasible solution for accurate tumor segmentation in scenarios with limited annotations.

References

[1] Yu Shi Zhang et al. "Surgical Treatment of Pheochromocytoma and Retroperitoneal Paraganglioma". In: *Surgical Management of Pheochromocytoma and Retroperitoneal Paraganglioma*. Springer, 2024, pp. 23–27.

[2] San Tang et al. "Adaptive cosegmentation of pheochromocytomas in CECT images using localized level set models". In: *IEEE Journal of Biomedical and Health Informatics* 20.2 (2015), pp. 549–562.

[3] Dong Wang et al. "PheoSeg: A 3D transfer learning framework for accurate abdominal CT pheochromocytoma segmentation and surgical grade prediction". In: *Knowledge-Based Systems* 301 (2024), p. 112202.

[4] David C Oluigbo et al. "Weakly supervised detection of pheochromocytomas and paragangliomas in CT". In: *Medical Imaging 2024: Computer-Aided Diagnosis*. Vol. 12927. SPIE. 2024, pp. 176–180.

[5] Bikash Santra et al. "Anatomical Location-Guided Deep Learning-Based Genetic Cluster Identification of Pheochromocytomas and Paragangliomas from CT Images". In: *International Workshop on Applications of Medical AI*. Springer. 2023, pp. 62–71.

[6] Alexander Kirillov et al. "Segment anything". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4015–4026.

[7] Ho Hin Lee et al. "Foundation models for biomedical image segmentation: A survey". In: *arXiv preprint arXiv:2401.07654* (2024).

[8] Haoyu Wang et al. "SAM-Med3D: Towards General-purpose Segmentation Models for Volumetric Medical Images". In: *arXiv preprint arXiv:2310.15161* (2024).

[9] Yicheng Wu et al. "Mutual consistency learning for semi-supervised medical image segmentation". In: *Medical Image Analysis* 81 (2022), p. 102530.

[10] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation". In: *2016 fourth international conference on 3D vision (3DV)*. Ieee. 2016, pp. 565–571.

[11] Danni Yang et al. "Sam as the guide: mastering pseudo-label refinement in semi-supervised referring expression segmentation". In: *arXiv preprint arXiv:2406.01451* (2024).

[12] Nanqing Liu et al. "PointSAM: Pointly-Supervised Segment Anything Model for Remote Sensing Images". In: *IEEE Transactions on Geoscience and Remote Sensing* (2025).

[13] Guoning Zhang et al. "IPLC: iterative pseudo label correction guided by SAM for source-free domain adaptation in medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 351–360.

[14] Lara Siblini et al. "Optimal prompting in sam for few-shot and weakly supervised medical image segmentation". In: *International Workshop on Foundation Models for General Medical AI*. Springer. 2024, pp. 103–112.

[15] Jun Ma et al. "Segment anything in medical images". In: *Nature Communications* 15.1 (2024), p. 654.

[16] Nhat-Tan Bui et al. "Sam3d: Segment anything model in volumetric medical images". In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2024, pp. 1–4.

[17] Chunhui Zhang et al. "A comprehensive survey on segment anything model for vision and beyond". In: *arXiv preprint arXiv:2305.08196* (2023).

[18] Chengyin Li et al. "AutoProSAM: Automated Prompting SAM for 3D Multi-Organ Segmentation". In: *arXiv preprint arXiv:2308.14936* (2023).

[19] Zhen Chen et al. "UN-SAM: Universal Prompt-Free Segmentation for Generalized Nuclei Images". In: *arXiv preprint arXiv:2402.16663* (2024).

[20] Dong-Hyun Lee et al. "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks". In: *Workshop on challenges in representation learning, ICML*. Vol. 3. 2. Atlanta. 2013, p. 896.

[21] Kihyuk Sohn et al. "Fixmatch: Simplifying semi-supervised learning with consistency and confidence". In: *Advances in neural information processing systems* 33 (2020), pp. 596–608.

[22] Xiangyu Zhao et al. "Rcps: Rectified contrastive pseudo supervision for semi-supervised medical image segmentation". In: *IEEE Journal of Biomedical and Health Informatics* (2023).

[23] Antti Tarvainen and Harri Valpola. "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results". In: *Advances in neural information processing systems* 30 (2017).

[24] Xiangde Luo et al. "Semi-supervised medical image segmentation through dual-task consistency". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 10. 2021, pp. 8801–8809.

[25] Chenchu Xu et al. "Deep Generative Adversarial Reinforcement Learning for Semi-Supervised Segmentation of Low-Contrast and Small Objects in Medical Images". In: *IEEE Transactions on Medical Imaging* (2024).

[26] Tuan-Hung Vu et al. "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 2517–2526.

[27] Ling Huang et al. "A review of uncertainty quantification in medical image analysis: probabilistic and non-probabilistic methods". In: *Medical Image Analysis* (2024), p. 103223.

[28] Xiangde Luo et al. "Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency". In: *Medical Image Analysis* 80 (2022), p. 102517.

[29] Yuang Shi et al. "Uncertainty-weighted and relation-driven consistency training for semi-supervised head-and-neck tumor segmentation". In: *Knowledge-Based Systems* 272 (2023), p. 110598.

[30] Yutong Xie et al. "Attention mechanisms in medical image segmentation: A survey". In: *arXiv preprint arXiv:2305.17937* (2023).

[31] Jiahui Zhong et al. "PMFSNet: Polarized Multi-scale Feature Self-attention Network For Lightweight Medical Image Segmentation". In: *arXiv preprint arXiv:2401.07579* (2024).

[32] Reza Azad et al. "Beyond self-attention: Deformable large kernel attention for medical image segmentation". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, pp. 1287–1297.

[33] Guanqun Sun et al. "DA-TransUNet: integrating spatial and channel dual attention with transformer U-net for medical image segmentation". In: *Frontiers in Bioengineering and Biotechnology* 12 (2024), p. 1398237.

[34] Rushi Jiao et al. "Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation". In: *Computers in Biology and Medicine* (2023), p. 107840.

[35] Yongchao Wang et al. "Mcf: Mutual correction framework for semi-supervised medical image segmentation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 15651–15660.

[36] Huimin Wu et al. "Compete to win: Enhancing pseudo labels for barely-supervised medical image segmentation". In: *IEEE Transactions on Medical Imaging* 42.11 (2023), pp. 3244–3255.

[37] Quan Zhang et al. "Distilling Semantic Priors from SAM to Efficient Image Restoration Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 25409–25419.

[38] Zipeng Qi et al. "Multi-View Remote Sensing Image Segmentation with Sam Priors". In: *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2024, pp. 8446–8449.

[39] Yuhao Huang et al. "Segment anything model for medical images?" In: *Medical Image Analysis* 92 (2024), p. 103061.

[40] T-YLPG Ross and GKHP Dollár. "Focal loss for dense object detection". In: *proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2980–2988.

[41] Lequan Yu et al. "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation". In: *Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, proceedings, part II 22*. Springer. 2019, pp. 605–613.

[42] Zhe Xu et al. "Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation". In: *Medical Image Analysis* 88 (2023), p. 102880.

[43] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-excitation networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7132–7141.

[44] Liam Chalcroft et al. "Large-kernel Attention for Efficient and Robust Brain Lesion Segmentation". In: *arXiv preprint arXiv:2308.07251* (2023).

[45] Yichao Liu, Zongru Shao, and Nico Hoffmann. "Global attention mechanism: Retain information to enhance channel-spatial interactions". In: *arXiv preprint arXiv:2112.05561* (2021).

[46] Sanghyun Woo et al. "Cbam: Convolutional block attention module". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19.