Anatomical Knowledge Mining and Matching for Semi-supervised Medical Multi-structure Detection

Bin Pu[1], Liwen Wang[2], Jiewen Yang[3], Xingbo Dong[2], Benteng Ma[3], Zhuangzhuang Chen[3], Lei Zhao[1], Shengli Li[4], Kenli Li[1†]

[1]Hunan University, Changsha, China
[2]Anhui University, Hefei, China
[3]The Hongkong University of Science and Technology, HKSAR, China
[4]Shenzhen Maternity and Child Healthcare Hospital, Southern Medical University, Shenzhen, China
{pubin, lkl}@hnu.edu.cn, liwenwang919@gmail.com, jyangcu@connect.ust.hk, xingbo.dong@ahu.edu.cn

Abstract

In medical image analysis, detecting multiple structures is crucial for evaluations and diagnosis but is often limited by the lack of high-quality annotations. Semi-supervised object detection emerges as a potent methodology to enhance model performance and generalization by leveraging a vast pool of unlabeled data alongside a minimal set of labeled data. A striking observation is that both unlabelled and labeled medical images contain a priori anatomical knowledge from human screening. In this work, we introduce a novel semi-supervised approach named Semi-akmm for mining and matching anatomical knowledge in ultrasound images. We develop an Adaptive Prior Knowledge Transfer (APKT) module to mine and explore the distribution and knowledge of potential proposal boxes by proposal proportion constraint. Furthermore, within a teacher-student learning framework, we put forward an Anatomical Structure Matching (ASM) module to facilitate co-learning consistent topological prior knowledge between the student and teacher models. To our knowledge, this marks the inception of an efficient semi-supervised medical multi-structure detection model. Our experiments across five publicly available ultrasound datasets demonstrate that Semi-akmm sets a new benchmark in performance with solid results that outperform existing methods.

Introduction

Medical multi-structure detection plays an essential role in assisted diagnostics, e.g., training the skills of novice inexperienced novice physicians (Yang et al. 2020), standardized screening (Loh et al. 2021; Pu et al. 2021), and diagnosis of structural-deficiency-based diseases (Elakkiya et al. 2021) (e.g., single-ventricle diseases due to the absence of one ventricular structure). However, developing robust models that can accurately recognize and analyze multiple structures in medical images requires a large amount of box-labeled data. The large-scale data annotation process is time-consuming, expensive, and unaffordable in real applications, so obtaining labels can be a bottleneck. Especially for medical image analysis, labeling these structures usually requires trained experts with prior biomedical knowledge (Zhou et al. 2021a), which may not be available. However, unlabeled medical data is easier to acquire.

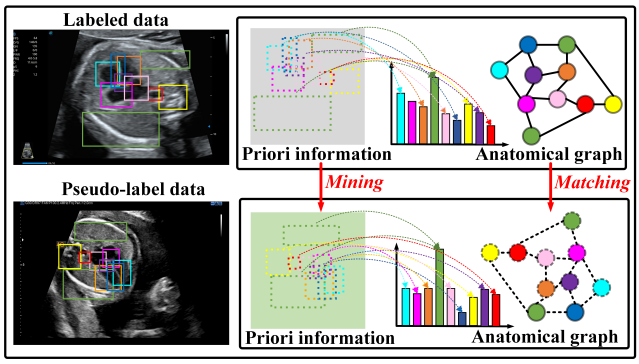

Figure 1: Motivation of our Semi-akmm. Two types of priori medical knowledge are distributional proposal knowledge, such as proportions and areas, and topological graph knowledge between multiple structures.

Semi-supervised learning is a settlement route where more robust detection performance can be obtained using unlabeled data and a small amount of labeled data trained together (Zhou and Zhou 2021). These approaches (Xu et al. 2021; Li et al. 2022; Liu et al. 2020; Zhang et al. 2023; Wang et al. 2023b; Hua et al. 2023; Li et al. 2023) have shown their strong robustness and effectiveness in natural image detection scenarios. For example, (Chen et al. 2022a) developed an effective label-matching framework to solve the issue of semi-supervised object detection task that lead to severe confirmation bias during self-training. These methods are universal semi-supervised detection methods for natural scenarios, and their direct application to medical semi-supervised multi-structure detection may lead to performance degradation or be inapplicable.

In medical image analysis, although many semi-supervised learning methods have been proposed for a wide range of applications, such as abnormality classification (Lecouat et al. 2018), 2D image segmentation (Zhang et al. 2017; Luo et al. 2021; Wu et al. 2022), and 3D volume segmentation (Yu et al. 2019; Chen et al. 2023). However, multi-structure detection applications have not been widely

†Corresponding author.

explored and investigated. Currently, there are no established systematic methods to address this research gap.

Unlike natural detection scenes, medical images (e.g., ultrasound images) typically contain information from human body screening that is enriched with a priori knowledge, such as complex anatomical relationships of multiple structures. For example, as shown in Fig. 1, we observe the existence of two types of prior knowledge in multiple organs of the human body. Firstly, the distribution of the proportions and areas of these structure proposals has a priori information. For example, the size of the human anatomy and the number of structures in the overall data are stable a priori information. Secondly, there is undoubtedly topological information between the multiple structures (e.g., the left ventricle, the left atrium, the right ventricle, and the right atrium). The anatomical, structural, and morphological information contained in medical images can be considered as strong prior knowledge. Based on this motivation, we designed a novel semi-supervised multi-structure detection method called ***Semi-akmm***, which integrates the multiple priori knowledge of medical images.

Drawing on the widely utilized mean-teacher learning framework (Tarvainen and Valpola 2017), we implement Semi-akmm in terms of both ***mining*** and ***matching*** prior anatomical structure knowledge. ***(1) Mining.*** The area of anatomical structures and the probability of their occurrence in the training dataset are varied (e.g., The distribution of proposals for small structures differs from that of large structures), and we mine this priori information and devise an Adaptive Prior Knowledge Transfer (APKT) module to constrain the student and teacher by the area and proportion of the structures. ***(2) Matching.*** Based on the observed relatively the relationships between the positions of multiple anatomical structures (see Fig. 1), anatomical relationships were constructed as graph-based knowledge and an Anatomical Structure Matching (ASM) module was devised to align teacher and student. The contributions can be outlined as:

- We present the first benchmark for the semi-supervised medical multi-structure detection task.
- We propose an Adaptive Prior Knowledge Transfer (APKT) module to mine the prior distribution information of the structure proposal.
- We propose an Anatomical Structure Matching (ASM) module to align the prior anatomical knowledge of the teacher (unlabeled data) and student (labeled data).
- Extensive experiments are performed on five public datasets, and quantitative experimental results and visualization analysis show that our Semi-akmm achieves state-of-the-art performance.

Related Work

Semi-supervised Object Detection in Natural Scenario

Semi-supervised object detection (SSOD) can be roughly categorized into consistency regularization (CR) based and pseudo-label (PL) based methods. CR-based methods aim to harmonize the detector's predictions by employing various augmentations. CSD (Jeong et al. 2019), as a typical SSD approach, compels the object detector to deliver uniform predictions for both the original image and its flipped horizontal counterpart. PL-based methods aim to produce high-quality and accurate pseudo-class labels stably. For the pursuit of superior performance, more researchers have adopted popular PL-based methods and developed a series of solutions (Sohn et al. 2020; Liu et al. 2020; Tang et al. 2021; Xu et al. 2021; Zhou et al. 2021b; Wang et al. 2023b; Hua et al. 2023; Ge et al. 2023; Liu et al. 2023b; Wang et al. 2023a; Nie et al. 2023; Li et al. 2023; Liu et al. 2023a; Zhang, Sun, and Wei 2023; Kar et al. 2023). (Sohn et al. 2020) introduced a novel approach to SSOD by proposing a teacher-student framework based on a basic multi-stage pipeline where the teacher model uses weakly augmented images to generate pseudo labels, which are then used to train the student model using strong augmentations to match the respective pseudo labels. To simplify the multi-stage process, several studies (Tang et al. 2021; Liu et al. 2020; Xu et al. 2021; Zhou et al. 2021b; Liu et al. 2023a; Chen et al. 2022b; Zhou et al. 2022) based on an end-to-end scheme have been designed to progressively update teachers through students' EMA and forecast pseudo-labels online for further improving the detection performance. However, previous studies have not properly explored the semi-supervised detection of multiple structures in medical images. Unlike the natural scene approach, the proposed semi-supervised approach focuses on mining and matching the prior topological knowledge.

Semi-supervised Learning in Medical Scenario

Due to the difficulty of medical image annotation, semi-supervised learning is widely used in medical image analysis, such as image detection (Zhou et al. 2021a), classification (Bai et al. 2023; Chaitanya et al. 2023; Zhang et al. 2022; Zeng et al. 2023; Gao et al. 2023), and segmentation (Shi et al. 2021; Luo et al. 2022; Basak and Yin 2023; Bai et al. 2023; You et al. 2024). In classification, (Zeng et al. 2023) designed a pseudo-loss estimation and feature adversarial training semi-supervised framework via loss distribution modeling and adversarial training. In segmentation, (Bai et al. 2023) proposed alleviating the empirical mismatch problems between labeled and unlabeled data distribution by copy-pasting labeled and unlabeled data bi-directionally based on a simple Mean Teacher framework. However, only very few studies have explored semi-supervised object detection in medical images. (Zhou et al. 2021a) presented a new adaptive consistency cost function and a heterogeneous perturbation strategy for the effective supervision of unlabeled medical data as a pioneering study. However, (Zhou et al. 2021a) is mainly designed for single object detection tasks (e.g., lesion detection or nuclei detection), and multi-structure semi-supervised tasks in medical images are completely unexplored. This paper incorporates a priori anatomical knowledge of multi-structures as a first benchmark for semi-supervised multi-structure detection.

Method

Fig. 2 shows an overview of our method consisting of **Anatomical Structure Matching (ASM)** and **Adaptive**

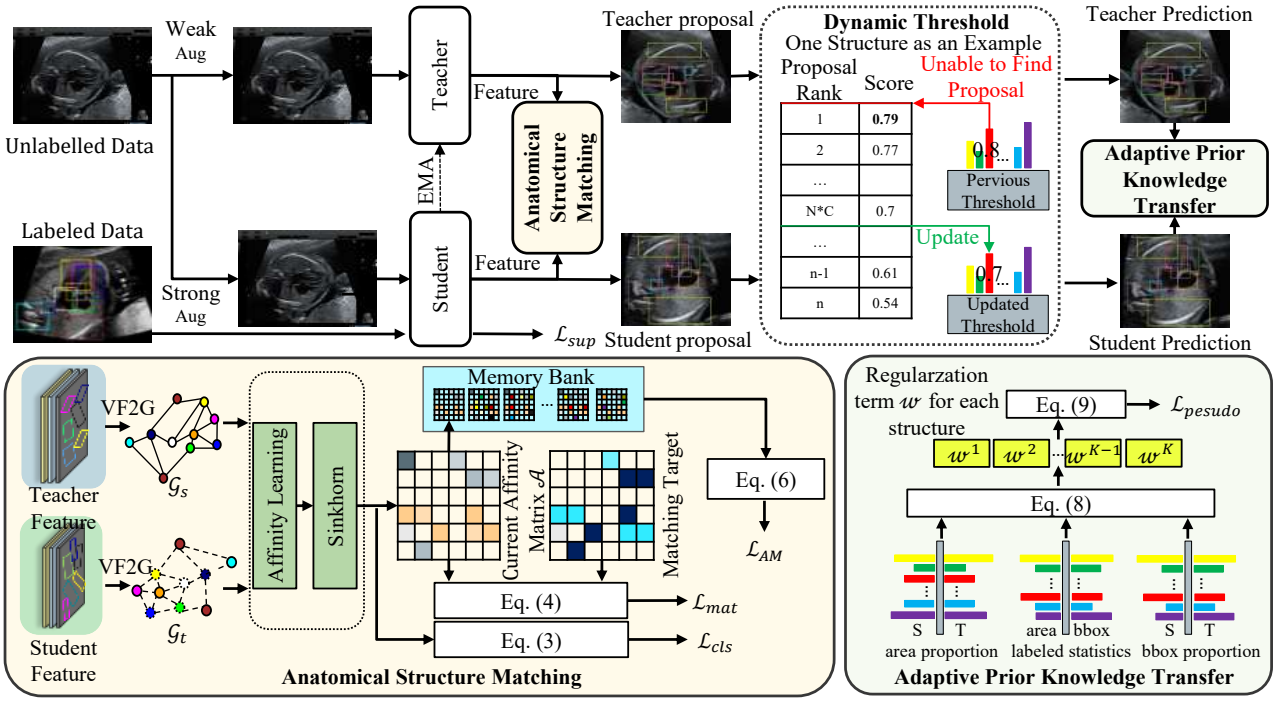

Figure 2: Overview of Semi-akmm. The unlabeled data will be entered into the teacher and student module using different argumentation strategies. The labeled data that provides strong prior medical knowledge goes to the student network for training. The features from both networks will be entered into ASM for the topological representation alignment. APKT dynamically assigns thresholds and regularization terms for bounding box regression and importance measurement of structures. This module allows us to pay more attention to those complex and hard-to-detect structures.

Prior Knowledge Transfer (APKT), which is based on the mean-teacher framewrok (Tarvainen and Valpola 2017) with a ResNet50 object detection backbone (He et al. 2016).

Anatomical Structure Matching

Anatomical Structure Formulation The ASM module aims to establish the structural representation relationship between labeled and unlabeled data, facilitating representation alignment. In the first step, we acquire the feature map $\hat{\mathcal{F}}^{s/t}$ of unlabelled data from the teacher and student network, and the feature map $\mathcal{F}^s$ of labelled data from the student network. Each of them consists of features from shallow to deep blocks of the encoder in both student and teacher networks. We introduce the Feature Pyramid Networks (FPN) (Lin et al. 2017) to ensure that the dimension of feature maps from different blocks remains the same. For the second step, we equidistantly sample feature point of $\hat{\mathcal{F}}^{s/t}$ and $\mathcal{F}^t$ according to the pseudo and ground truth bounding box $\hat{Y}_{box}$ and Y_{box}, respectively. In total, no more than M number of points will be sampled from the deep to shallow feature. We then reformulate these points as visual node $\hat{\mathcal{V}}^{s/t} \in \mathbb{R}^{\hat{M}^{s/t} \times d}$ and $\mathcal{V}^s \in \mathbb{R}^{M^s \times d}$.

Those visual nodes are decoupled from the feature map as the whole feature map has a large amount of redundant information and may not be suitable for us to explore the key detected structures. For example, as shown in Fig. 1, approximately only 1/4 of the fetus ultrasound image contains interesting areas, suggesting that 3/4 of the background information is unrelated to our detection task. Furthermore, to construct the structural information, the graphical representation is more suitable for our task due to the following advantages: 1) graphs offer a more flexible measure for representing irregular structures. 2) graphs can articulate the relationship and connectivity between different structures without being affected by their angle and position in images.

Based on the above discussion, we first compute the pairwise distance of each visual node to establish their connectivity as edges, formulated as $\mathcal{E} = \mathcal{V} \times \mathcal{V}^{\intercal} \in \mathbb{R}^{M \times M}$. With the visual nodes $\mathcal{V}$ and their connectivity $\mathcal{E}$, we can construct the graphical representation between each detected structure using a graph convolutional neural network (GCNN) with the following formulation:

$$\mathcal{G} = \text{GCNN}(\{\mathcal{V}, \mathcal{E}\}) \in \mathbb{R}^{M \times d}. \quad (1)$$

For different inputs, we obtain their graphical representations $\hat{\mathcal{G}}^{s/t} \in \mathbb{R}^{M \times d}$ and $\mathcal{G}^s \in \mathbb{R}^{M \times d}$ using Eq. (1) for unlabeled and labeled data.

Anatomical Structure Optimization In the mean-teacher architecture, the teacher network does not participate in training and receives the unlabeled image with weak augmentation as input. This allows the teacher to generate more precise pseudo-labels for the student network as a reference. On the other hand, the student network, which receives

the labeled image as input, can reveal ground truth structures. In the matching stage, the graph $\hat{\mathcal{G}}^s$ has two matching targets: $\hat{\mathcal{G}}^t$ ($\hat{\mathcal{G}}^s \rightarrow \hat{\mathcal{G}}^t$) and $\mathcal{G}^s$ ($\hat{\mathcal{G}}^s \rightarrow \mathcal{G}^s$). To achieve a more robust graphical representation, we first introduce the attention mechanism to fuse information across different graphs. In the following statement, we use the graph pairs $\mathcal{G}^a \in \mathbb{R}^{M^a \times d}$ and $\mathcal{G}^b \in \mathbb{R}^{M^b \times d}$ to describe the general process of ASM between graphs, and we have the following:

$$[\dot{\mathcal{G}}^a, \dot{\mathcal{G}}^b] = \xi(\mathcal{G}^a, \mathcal{G}^b) = (\mathcal{G}^{a/b}\mathcal{W}^q)(\mathcal{G}\mathcal{W}^k)^{\mathsf{T}}(\mathcal{G}\mathcal{W}^v), \quad (2)$$

where the graph pair $\dot{\mathcal{G}}^a \in \mathbb{R}^{M^a \times d}, \dot{\mathcal{G}}^b \in \mathbb{R}^{M^b \times d}$, $\xi(\cdot,\cdot)$ denotes the cross-attention operation, $\mathcal{G} = \text{Concat}(\mathcal{G}^a, \mathcal{G}^b) \in \mathbb{R}^{(M^a+M^b)\times d}$, and $W^q, W^k, W^v \in \mathbb{R}^{d\times d}$ are the learnable weights of projection layers that project the features to latent spaces, namely the query, key, and value. The graph pair $[\dot{\mathcal{G}}^a, \dot{\mathcal{G}}^b]$ is supervised by the cross-entropy loss (denoted as $\text{CE}(\cdot)$) to ensure that each attentive node matches the correct classes, defined as:

$$\mathcal{L}_{\text{cls}}([\dot{\mathcal{G}}^a, \dot{\mathcal{G}}^b]) = \text{CE}(\dot{\mathcal{G}}^a, Y^a_{cls}) + \text{CE}(\dot{\mathcal{G}}^b, Y^b_{cls}). \quad (3)$$

Similar to (Li, Liu, and Yuan 2022), we build the affinity matrix $\mathcal{A}^{\dot{\mathcal{G}}^a \rightarrow \dot{\mathcal{G}}^b} \in \mathbb{R}^{M^a \times M^b}$ between graph pair $\dot{\mathcal{G}}^a$ and $\dot{\mathcal{G}}^b$ via the affinity net $\phi([\dot{\mathcal{G}}^a, \dot{\mathcal{G}}^b])$ (see Fig. 2) to represent the affinity relationship of nodes across graphs $\mathcal{G}^a, \mathcal{G}^b$. For each affinity matrix, we have the matching optimization:

$$\begin{aligned}\mathcal{L}_{\text{mat}}(\phi([\dot{\mathcal{G}}^a, \dot{\mathcal{G}}^b])) = &\sum\nolimits_{i,j} \mathbb{I}(i^a = j^b) \cdot \sigma(\mathcal{A}^{\dot{\mathcal{G}}^a \rightarrow \dot{\mathcal{G}}^b}_{i,j}) \\ &- \sum\nolimits_{i,j} \mathbb{I}(i^a \neq j^b) \cdot \sigma(\mathcal{A}^{\dot{\mathcal{G}}^a \rightarrow \dot{\mathcal{G}}^b}_{i,j}),\end{aligned} \quad (4)$$

where $\mathbb{I}(\cdot)$ is the indicator function and $\sigma(\cdot,\cdot)$ represents the Sinkhorn iteration (Cuturi 2013), i, j denote the i-th row and j-th colume of the affinity matrix $\mathcal{A}^{\dot{\mathcal{G}}^a \rightarrow \dot{\mathcal{G}}^b}$.

Cross-sample Anatomical Structure Consistency The graph matching via Eq. (4) can construct a consistent structural representation between two different samples. However, this operation is not able to avoid the noise brought by the imprecise prediction from the pseudo label and the bias brought by only a few labeled data. This section also introduces the cross-sample anatomical structure consistency method to eliminate the disadvantages discussed above.

We use the memory bank to store the affinity matrix of different pairs across samples from the teacher and student networks. Then, for each current pair from these two domains, we will calculate the similarity between the current affinity matrix $\mathcal{A}$ and the ancestral matrix $\overline{\mathcal{A}}$ stored in the memory bank $\mathcal{M} = \{\overline{\mathcal{A}_1}, ..., \overline{\mathcal{A}_N}\}$. Due to the affinity matrix having different sizes with different input samples, we thus first conduct the regularization of each matrix in the $\mathcal{M}$ as the same size as the current affinity matrix. The similarity loss can be formulated as:

$$\mathcal{L}_{\text{sim}}(\mathcal{A}, \mathcal{M}) = \sum\nolimits_{n=1}^{N} \left\| \mathcal{A} - \tau[\overline{\mathcal{A}_n}] \right\|_1, \quad (5)$$

where $\tau[\cdot]$ resize $\overline{\mathcal{A}_n}$ to the same shape as $\mathcal{A}$. The overall matching loss $\mathcal{L}_{\text{AM}}$ of graph pair $\mathcal{G}^a, \mathcal{G}^b$ can be written as:

$$\begin{aligned}\mathcal{L}_{\text{AM}}(\mathcal{G}^a, \mathcal{G}^b) = &\lambda_1 \cdot \mathcal{L}_{\text{cls}}(\xi(\mathcal{G}^a, \mathcal{G}^b)) + \lambda_2 \cdot \mathcal{L}_{\text{mat}}(\phi(\xi(\mathcal{G}^a, \mathcal{G}^b))) \\ &+ \lambda_3 \cdot \mathcal{L}_{\text{sim}}(\phi(\xi(\mathcal{G}^a, \mathcal{G}^b)), \mathcal{M}),\end{aligned} \quad (6)$$

where λ_1, λ_2 and λ_3 are the scaling factor of the matching loss $\mathcal{L}_{\text{AM}}$ and set as 1, 1 and 0.5 in our experiment. In the final, the overall loss of the ASM can be formulated as:

$$\mathcal{L}_{\text{ASM}} = \mathcal{L}_{\text{AM}}(\hat{\mathcal{G}}^s, \mathcal{G}^s) + \mathcal{L}_{\text{AM}}(\hat{\mathcal{G}}^s, \hat{\mathcal{G}}^t). \quad (7)$$

Adaptive Prior Knowledge Transfer

In Fig. 2, we introduce the APKT module to maximize the informative value of the teacher network's intermediate output. This module aligns dynamically assigned thresholds with various detected structures and the proposal proportion. As depicted in Fig. 1, ultrasound images often exhibit similar morphology across different structures, including shape, size, pattern, and features. We hypothesize that a neural network's detection performance is influenced by the difficulty of detecting these structures. Table 2 consistently shows that DAO's mAP surpasses that of SP, attributed to DAO's more prominent pattern and larger size. To address these variations, the APKT module introduces the Proposal Proportion Loss and Dynamic Detection Threshold mechanisms. The Proposal Proportion Loss leverages general prior knowledge from medical images to account for detection difficulty, while the Dynamic Detection Threshold enhances the network's ability to detect small and complex structures.

Proposal Proportion Loss For training the student network, the predicted proposals of different structures in unlabelled data are denoted as $\{\hat{P}^k\}_{k=1}^K$, where K is the total number of structures that need to be detected, and p^k denotes the all detected proposals in k-th structure. Similarly, the labelled data can be formulated as $\{P^k\}_{k=1}^K$. For each proposal in $\hat{P}^k$ and P^k, we have the corresponding predicted/ground-truth bounding box and classification label $\{\hat{Y}^k_{i,box}, \hat{Y}^k_{i,cls}\}_{i=1}^{\hat{P}^k}$ and $\{Y^k_{i,box}, Y^k_{i,cls}\}_{i=1}^{P^k}$, respectively. During the training in mean-teacher architecture, some complex or small objects may cause difficulty in detection and thus degrade the quality of pseudo labels. During the training of unlabelled data, we consider that more attention needs to be paid to those structures. Hence, in order to explore the difficulty of detection in different structures, we utilize the proposals generated by labeling data that are able to describe the detection ability of the student network.

As shown in Fig. 2, given the predicted proposals from unlabelled data $\{\hat{P}^k\}_{k=1}^K$ and ground truth proposal $\{P^k\}_{k=1}^K$ from labelled data, we compute the regularization term $\hat{w}$ of each structure as the following:

$$\hat{w}^k = \frac{\sum_{i=1}^{\hat{P}^k} \text{area}(\hat{Y}^k_{i,box}) \cdot P^k}{\sum_{j=1}^{P^k} \text{area}(Y^k_{j,box}) \cdot \hat{P}^k} + \frac{\sum_{k'=1}^{K} \mathbb{I}(\hat{Y}^{k'}_{cls} = \hat{Y}^k_{cls})}{\sum_{k'=1}^{K} \mathbb{I}(Y^{k'}_{cls} = Y^k_{cls})}, \quad (8)$$

where $\text{area}(\cdot)$ is the operation to compute the proposal area using the bounding box, and $\mathbb{I}(\cdot)$ is the indicator function that calculates the number of detected proposals that belong to k-th structure across the overall proposals. With the regularization term $\hat{w}$, the supervision loss of the unlabelled data by using the pseudo-label can be formulated as:

$$\mathcal{L}_{\text{pesudo}} = \text{CE}(\hat{Y}^s_{\text{cls}}, \hat{Y}^t_{\text{cls}}) + \sum\nolimits_{k=1}^{K} \hat{w}^k \cdot \text{Reg}(\hat{Y}^{s,k}_{\text{box}}, \hat{Y}^{t,k}_{\text{box}}), \quad (9)$$

where $\mathrm{CE}(\cdot)$ and $\mathrm{Reg}(\cdot)$ denote the cross-entropy loss and bounding box regression loss, respectively.

Dynamic Detection Threshold In ASM, the graphical representation of structures is built by nodes that sample from feature maps according to the ground truth and pseudo label. However, some structures that are difficult to detect are unable to acquire corresponding nodes from the label, which leads to an incomplete graph and hinders the optimization process of the ASM. To overcome this problem, we use the dynamic detection threshold instead of the default threshold for different structures. The low threshold causes easy-to-detect structures to have excessive proposals, while the high threshold may lead to missed detection. To perform the dynamic threshold, we first set a constant C, where $C \in (0, 1)$ defines the portion of all detected proposals that need to be reserved in each iteration, and the reserved proposal with the lowest detect confidence will be set as the threshold. We perform the same operation on each structure and make sure that each structure can be detected.

Results

Datasets and Evaluation

Fetal Cardiac Structure (FCS) (Pu et al. 2024) is a diversified ultrasound dataset collected from our two medical centers, each containing two views of the heart, i.e., three vessels and trachea view (3VT) and four-chamber cardiac view (4C), with a total of four datasets. These datasets are from different medical devices, such as Samsung, Sonoscape, and Philip, with a gestational week range of 20-34 weeks. The 3VT and 4C from *A* medical center are denoted as **3VT-A** and **4C-A**. Similarly, they from *B* medical center are denoted as **3VT-B** and **4C-B**. The total number of 4C-A, 4C-B, 3VT-A, and 3VT-B are 810, 809, 891, and 369, respectively. 4C contains 9 anatomical structures, i.e., Left ventricle (LV), Left atrium (LA), Right ventricle (RV), Descending aorta (DAO), Right atrium (RA), Ventricular septum (VS), Spine (SP), Rib (RIB), and Cross (CRO). 3VT contains Superior vena cava (SVC), Arch of Aorta (AOA), Trachea (T), SP, Pulmonary trunk & ductus arteriosus (PTDA), and DAO.

Early Pregnancy View (EPV) (Lin et al. 2022) is a challenging early pregnancy ultrasound dataset collected from different ultrasound devices, a total number of 1131 images, and its gestational range is 10-14 weeks. EP includes 9 key structures, i.e., thalami (TH), midbrain (MB), palate (PAL), Intracranial Transparent (IT, i.e., 4th ventricle), cisterna magna (CM), nuchal translucency (NT), nasal tip (NST), nasal skin (NS), and nasal bone (NB). During the evaluation phase, we present the mean Average Precision (mAP) within all the test datasets, using an Intersection over Union (IoU) threshold of 0.5.

Implementation Details

For a fair comparison, we use Faster-RCNN with FPN (Girshick 2015) and ResNet-50 backbone (He et al. 2016) as the detector, which is implemented in PyTorch and trained 40k iterations with 1 batchsize on one RTX3090 GPU. For data augmentation, we use random horizontal flipping for weak augmentation. Based on this augmentation, we then add random color jittering, grayscale, gaussian blurring and cutout patches for strong augmentation, which is similar to (Liu et al. 2020). We uniformly resized images to 1024×1024 for all stages, and we trained the model using the stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.01 with the weight decay of 5×10^{-4}. The FCS and EPV datasets were divided into train, valid, and test sets in the ratio of 7:1:2, and all the settings remained the same.

Comparison with State-of-the-arts

Overall Results. As shown in Table 1, with 1%, 5%, and 10% labeled data, respectively, our method outperforms the compared methods in most cases, especially on the 3VT-A, 3VT-B and EPV datasets. These results reveal that our approach consistently achieves the highest or second-highest performance across all experimental labeling configurations on five distinct medical multi-object detection datasets. This pattern of results suggests that our methodology exhibits superior stability in comparison to existing methods, demonstrating our approach's effectiveness and superiority.

Results on 3VT-A. As listed in Tabel 2, Semi-akmm has marked advantages over the current SOTA SSOD methods in almost all experimental settings of 3VT-A. At 1% and 10% data labeling settings, our method outperforms the best method Label Matching (Liu, Ma, and Kira 2022) and Unbiased Teacher (Chen et al. 2022a) by 1.3% and 0.9% mAP, respectively. Compared to the fully supervised method, our Semi-akmm can greatly improve 21.4%, 23.3% and 27.3% mAP in labeling 1%, 5% and 10% settings, respectively.

Results on 3VT-B. As reported in Table 3, we performed experiments following different labeling protocols on 3VT-B. From Table 3, it can be observed that Semi-akmm achieves performance improvement in all data labeling settings. For example, at labeling 10% of the data, our method is superior to Label Matching (Liu, Ma, and Kira 2022) by 1% mAP.

Results on EPV. As shown in Table 3, similar to 3VT-B, our method demonstrates superior performance over existing techniques across various labeled data settings. Importantly, EPV encompasses early fetal pregnancy stages, characterized by immature organ structures that are notably small, obscure, and embedded in noise. Despite these challenges, our approach achieves accurate detection, signifying its effectiveness in extracting and aligning a priori knowledge. This proficiency shows our method's exceptional ability to interpret complex medical imagery accurately, indicating a significant advancement in medical image analysis.

Results on 4C-A and 4C-B. As delineated in Table 2 and Table 1 in Appendix[‡], our approach consistently ranks as the best/second-best across most configurations of semi-supervised labeling, unequivocally validating the efficacy and robustness of our methodology. In contrast, the performance is based on labeling configurations. On 4C-A, the Unbiased Teacher (Chen et al. 2022a) outperforms others with 1% labeled data, while the Consistent Teacher (Wang et al. 2023b) performs better with 5% labeled data.

[‡]Appendices are available at https://github.com/LiwenWang919/Semiakmm.

		4C-A			4C-B			3VT-A			3VT-B			EPV		
	Threshold	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
Supervised	0.9	18.1	28.9	30.4	16.2	21.8	23.0	15.3	25.9	32.9	16.7	20.9	22.2	17.6	19.4	21.0
Unbiased Teacher (Liu, Ma, and Kira 2022) ICLR'21	0.7	**29.4**	53.6	56.5	20.9	**52.2**	55.9	29.8	55.6	60.9	37.7	43.5	47.9	40.9	46.4	48.1
Soft Teacher (Xu et al. 2021) ICCV'21	0.9	24.9	51.9	56.9	20.2	50.7	57.9	30.5	56.4	59.9	36.6	43.8	47.8	38.6	45.4	48.2
STAC (Sohn et al. 2021) Arxiv'21	0.9	28.9	52.7	57.7	20.7	51.5	57.3	28.5	55.8	60.4	36.7	43.0	48.5	37.8	44.8	47.5
Label Matching (Chen et al. 2022a) CVPR'22	adaptive	28.4	51.4	57.4	20.3	50.6	57.0	29.9	56.6	60.7	37.6	43.6	48.1	40.8	46.5	47.5
Consistent Teacher (Wang et al. 2023b) CVPR'23	0.9	28.2	**56.5**	57.2	28.1	50.9	58.2	29.6	56.2	60.5	37.3	43.7	48.3	37.8	45.5	47.6
Semi-DETR (Zhang et al. 2023) CVPR'23	0.8	27.9	54.2	56.7	27.5	50.5	57.2	28.8	**57.8**	59.6	36.6	42.9	47.3	38.2	45.8	46.7
Sparse Semi-DETR (Shehzadi et al. 2024) CVPR'24	0.8	29.1	55.2	58.8	28.4	51.5	58.3	30.4	57.6	60.8	37.7	43.9	48.6	40.9	46.2	48.3
Semi-akmm (Ours)	adaptive	29.2	55.9	**60.6**	**29.1**	51.8	**59.8**	**31.2**	57.7	**61.8**	**38.1**	**44.2**	**49.5**	**41.4**	**46.9**	**48.6**

Table 1: Overall quantitative results on all datasets. The bolded and underlined denote the best and the second-best performance.

Dataset		4C-A										3VT-A						
Method	Labeled rate	LA	RA	LV	RV	VS	CRO	SP	DAO	RIB	mAP	DAO	SP	PTDA	T	SVS	AOA	mAP
Supervised	1%	25.7	11.8	14.4	12.9	11.4	27.6	16.2	28.4	14.6	18.1	15.1	20.5	12.7	21.3	11.5	10.4	15.3
	5%	17.8	34.0	20.6	12.0	8.8	17.3	32.6	26.9	36.7	23.0	28.3	25.2	28.1	28.6	19.8	25.5	25.5
	10%	33.9	47.0	34.3	25.2	16.2	25.2	37.6	27.5	26.8	30.4	30.2	28.3	33.8	36.7	32.5	35.9	32.9
Unbiased Teacher (Liu, Ma, and Kira 2022) ICLR'21	1%	33.8	**19.8**	22.2	24.0	25.7	39.1	**31.3**	44.9	23.5	**29.4**	**32.4**	41.8	24.4	30.1	28.1	22.3	29.8
	5%	56.3	60.3	55.0	53.9	55.8	52.9	48.1	59.1	41.1	53.6	52.3	46.0	58.6	54.7	56.7	**65.0**	55.6
	10%	57.2	60.8	60.2	59.9	61.4	52.7	54.7	59.5	42.3	56.5	60.8	50.3	63.1	**62.5**	**62.3**	66.1	60.9
Soft Teacher (Xu et al. 2021) ICCV'21	1%	31.8	13.6	17.5	17.4	13.4	**43.0**	29.1	43.8	14.2	24.9	23.1	43.2	18.3	45.9	31.8	20.6	30.5
	5%	53.7	55.1	57.4	53.5	56.7	48.0	50.5	53.5	38.8	51.9	54.8	47.0	59.7	54.4	57.5	64.8	56.4
	10%	58.7	60.1	62.4	58.5	61.7	53.0	55.5	58.5	43.8	56.9	59.6	50.0	61.8	61.6	58.9	67.3	59.9
STAC (Sohn et al. 2021) Arxiv'21	1%	32.9	**19.8**	22.1	21.6	23.5	41.1	31.0	43.7	24.5	28.9	30.9	41.3	**24.8**	27.0	26.8	19.9	28.5
	5%	51.9	55.6	52.5	56.3	55.5	**56.0**	58.9	48.5	38.8	52.7	51.9	46.4	58.6	55.5	58.5	63.7	55.8
	10%	56.9	60.6	57.5	61.3	60.5	**61.0**	**63.9**	53.5	43.8	57.7	60.2	50.4	62.6	61.6	61.6	66.0	60.4
Label Matching (Chen et al. 2022a) CVPR'22	1%	37.4	18.5	24.8	24.3	24.7	33.6	17.8	46.4	28.5	28.4	21.7	42.3	19.3	44.1	32.3	19.9	29.9
	5%	54.5	54.8	55.8	54.5	56.3	49.2	50.9	51.9	34.9	51.4	55.9	44.8	58.9	58.1	59.4	62.4	56.6
	10%	60.5	60.8	61.8	60.5	62.3	55.2	56.9	57.9	40.9	57.4	61.4	**50.6**	63.8	60.1	60.5	68.1	60.7
Consistent Teacher (Wang et al. 2023b) CVPR'23	1%	**37.9**	16.9	25.1	23.3	**27.2**	31.4	17.4	48.4	26.1	28.2	20.7	42.9	21.3	40.4	31.0	21.0	29.6
	5%	**56.6**	62.7	61.1	**57.8**	57.3	51.6	**56.3**	60.4	**44.4**	**56.5**	52.7	47.0	58.3	54.4	60.1	64.7	56.2
	10%	58.4	61.4	61.5	55.0	61.2	53.5	58.4	60.7	44.7	57.2	**61.4**	**50.6**	63.2	60.9	59.3	67.4	60.5
Semi-DETR (Zhang et al. 2023) CVPR'23	1%	31.1	18.6	21.1	20.1	21.2	40.1	28.0	43.8	26.1	27.9	22.1	40.1	22.3	35.3	30.1	**22.7**	28.8
	5%	54.2	61.5	57.5	52.9	**57.3**	50.3	55.1	57.2	41.6	54.2	**56.8**	**50.3**	59.8	57.8	57.6	64.2	**57.8**
	10%	57.4	61.2	60.4	58.3	57.5	54.2	55.4	60.4	45.4	56.7	57.9	50.5	62.8	61.2	58.0	67.4	59.6
Sparse Semi-DETR (Shehzadi et al. 2024) CVPR'24	1%	37.1	18.3	26.9	**26.3**	23.4	36.5	15.3	48.9	29.6	29.1	24.0	41.6	18.0	46.4	32.2	20.4	30.4
	5%	54.2	63.1	59.9	55.7	56.1	52.6	50.0	61.7	43.3	55.2	58.7	48.3	59.4	**58.6**	57.8	62.9	57.6
	10%	58.4	60.9	64.8	60.3	61.2	54.0	59.5	62.3	**47.8**	58.8	61.4	48.9	64.9	60.3	61.1	68.0	60.8
Semi-akmm (Ours)	1%	37.3	17.7	**27.3**	25.5	25.4	35.6	14.4	**49.2**	**30.0**	29.2	23.3	**43.9**	19.2	**46.6**	**32.4**	21.6	**31.2**
	5%	56.4	**63.4**	**61.9**	56.7	56.8	55.2	50.9	**62.1**	42.7	55.9	55.7	46.6	**60.3**	57.8	**61.8**	63.9	57.7
	10%	**62.0**	**63.8**	**65.5**	**61.8**	**64.2**	57.7	59.0	**64.1**	47.4	**60.6**	61.3	50.4	**65.6**	62.0	**62.3**	**69.3**	**61.8**

Table 2: Quantitative results on 4C-A and 3VT-A.

Dataset		3VT-B							EPV								
Method	Labeled rate	DAO	SP	PTDA	T	SVS	AOA	mAP	TH	NB	PAL	NS	NST	MB	NT	IT	CM	mAP
Supervised	1%	18.3	16.2	19.8	11.4	15.2	18.5	16.7	20.1	17.5	18.4	15.2	17.4	20.7	15.4	16.2	17.6	17.6
	5%	21.1	18.5	21.6	19.6	18.1	26.2	20.9	22.4	20.1	20.5	17.9	19.4	22.1	16.9	18.4	16.7	19.4
	10%	23.7	20.5	26.3	21.6	19.4	21.7	22.2	24.5	19.4	22.6	18.6	20.1	25.5	19.2	20.4	18.4	21.0
Unbiased Teacher (Liu, Ma, and Kira 2022) ICLR'21	1%	37.5	35.3	44.0	32.2	31.2	**46.2**	37.7	58.9	40.5	50.0	38.6	39.7	53.2	19.8	41.8	25.3	40.9
	5%	29.8	**51.4**	51.8	42.6	39.3	46.1	43.5	**65.2**	44.9	59.9	42.8	43.3	56.6	30.6	46.2	28.4	46.4
	10%	44.7	46.9	55.2	45.4	45.1	50.2	47.9	65.4	45.0	62.4	44.6	44.1	57.1	33.3	48.8	31.8	48.1
Soft Teacher (Xu et al. 2021) ICCV'21	1%	33.7	42.8	31.6	34.9	32.1	44.4	36.6	57.1	36.3	45.6	36.6	39.7	50.0	18.7	38.2	25.2	38.6
	5%	31.9	45.2	51.1	**47.7**	40.4	46.8	43.8	62.8	43.7	58.0	40.5	41.6	55.0	30.6	47.2	29.5	45.4
	10%	**49.1**	42.1	48.8	47.4	**49.4**	49.8	47.8	65.0	46.0	62.6	43.7	45.0	57.9	**34.8**	47.4	30.8	48.2
STAC (Sohn et al. 2021) Arxiv'21	1%	34.9	41.7	37.9	30.6	33.1	42.5	36.7	61.4	33.7	42.6	33.4	37.9	50.4	16.2	40.5	23.8	37.8
	5%	31.9	44.9	50.1	44.4	41.9	44.9	43.0	63.0	42.7	59.4	37.4	39.9	54.7	30.2	46.3	29.2	44.8
	10%	39.6	46.4	56.7	47.3	46.9	54.3	48.5	64.7	44.7	62.5	44.1	**45.4**	56.6	31.7	47.5	30.0	47.5
Label Matching (Chen et al. 2022a) CVPR'22	1%	40.1	38.5	39.0	31.8	30.8	45.0	37.6	58.4	40.4	**50.7**	40.0	39.1	53.0	18.7	42.0	24.9	40.8
	5%	32.3	44.2	51.0	46.5	40.7	46.7	43.6	64.1	**45.3**	**61.3**	43.1	42.8	55.8	31.7	46.3	27.8	46.5
	10%	39.1	45.9	54.6	47.1	46.8	55.5	48.1	65.5	**47.9**	62.3	43.6	42.8	57.4	29.9	48.7	29.1	47.5
Consistent Teacher (Wang et al. 2023b) CVPR'23	1%	24.2	**44.4**	44.2	**36.3**	34.2	40.7	37.3	**61.8**	36.9	49.1	38.6	38.6	52.2	19.9	**42.6**	25.1	40.5
	5%	**43.7**	40.2	41.6	42.5	**48.1**	45.9	43.7	64.0	41.8	60.4	40.3	39.9	55.8	31.7	47.6	28.4	45.5
	10%	**49.1**	42.1	48.3	47.6	40.8	51.9	48.3	65.1	44.6	60.6	41.7	45.2	**58.6**	33.9	47.8	30.8	47.6
Semi-DETR (Zhang et al. 2023) CVPR'23	1%	27.9	34.4	**46.2**	33.3	38.0	40.1	36.6	52.4	35.1	46.9	36.8	**41.1**	49.8	17.6	39.4	24.9	38.2
	5%	31.5	43.9	50.9	40.6	43.0	47.2	42.9	65.0	43.3	58.7	37.4	40.0	57.8	31.2	48.8	30.0	45.8
	10%	45.4	45.0	55.4	44.3	43.2	50.5	47.3	64.9	43.4	60.3	43.2	42.7	56.4	32.2	46.5	**30.4**	46.7
Sparse Semi-DETR (Shehzadi et al. 2024) CVPR'24	1%	29.7	36.9	45.8	34.0	**38.9**	41.1	37.7	51.3	34.6	45.7	37.2	40.7	48.5	18.2	40.0	25.0	40.9
	5%	41.4	43.4	50.6	38.5	44.7	44.9	43.9	65.4	43.7	59.1	37.8	40.4	**58.2**	31.6	**49.2**	30.4	46.2
	10%	41.7	46.5	54.2	46.4	48.7	54.1	48.6	**66.5**	45.0	61.9	44.9	44.3	58.1	30.8	48.1	**32.1**	48.3
Semi-akmm (Ours)	1%	**42.3**	37.4	45.4	29.0	32.0	42.9	**38.1**	58.0	**41.0**	50.2	**40.0**	40.9	**53.2**	**20.7**	42.3	**25.9**	**41.4**
	5%	33.9	45.0	**53.1**	43.0	42.5	**48.0**	**44.2**	64.2	44.9	60.1	**43.1**	**44.4**	57.6	**32.0**	46.4	29.2	**46.9**
	10%	40.8	**47.5**	**56.8**	**48.0**	47.4	**56.5**	**49.5**	66.0	45.3	**63.8**	**45.3**	44.5	57.5	33.4	**49.6**	32.0	**48.6**

Table 3: Quantitative results on 3VT-B and EPV.

Further Empirical Analysis

Effect of ASM and APKT. To further explore the effect of ASM and APKT on our model, as shown in Table 4, we performed ablation experiments on five datasets. From Table 4, each component improves detection accuracy and demonstrates the effectiveness of the proposed ASM and APKT. For example, on 4C-A, ASM improves by 4.2% mAP and APKT enhances by 3.1% mAP. On 3VT-A, ASM improves by 2.5% and APKT enhances by 3.6% mAP. When both

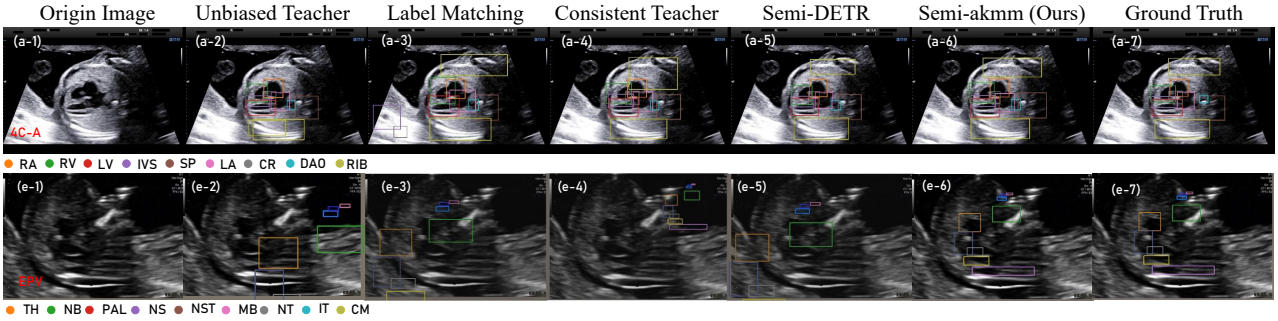

Figure 3: Qualitative result comparison on views 4C-A (top) and EPV (bottom).

Method	4C-A			4C-B			3VT-A			3VT-B			EPV		
	ASM	*APKT*	mAP(%)	*ASM*	*APKT*	mAP(%)	*ASM*	*APKT*	mAP(%)	*ASM*	*APKT*	mAP(%)	*ASM*	*APKT*	mAP(%)
Baseline	-	-	53.1	-	-	52.8	-	-	55.7	-	-	45.6	-	-	43.5
Ours	✓	✗	57.3 (+4.2)	✓	✗	56.4 (+3.6)	✓	✗	58.2 (+2.5)	✓	✗	47.5 (+1.9)	✓	✗	46.4 (+2.9)
	✗	✓	56.2 (+3.1)	✗	✓	57.2 (+4.4)	✗	✓	59.3 (+3.6)	✗	✓	47.9 (+2.3)	✗	✓	45.5 (+2.0)
	✓	✓	**60.6** (+7.5)	✓	✓	**59.8** (+7.0)	✓	✓	**61.8** (+6.1)	✓	✓	**49.5** (+3.9)	✓	✓	**48.6** (+5.1)

Table 4: An ablation study of ASM and APKT. All the reported results in the setting of using 10% labeled data.

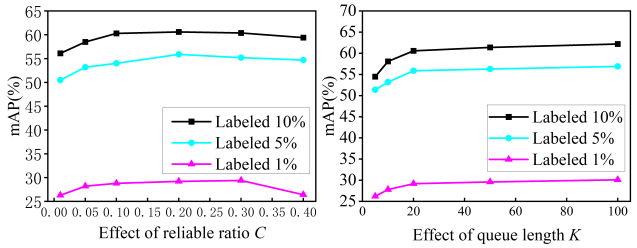

Figure 4: Ablation studies of reliable ratio C and queue length K on 4C-A.

ASM and APKT are available, the model receives a significant boost of 7.5%, 6.1%, and 5.1% mAP on 4C-A, 3VT-A, and EPV, respectively. In summary, each component is effective, and mining and aligning prior knowledge of multiple structures in medical images can improve the robustness of the medical semi-supervised detection task.

Effect of reliable ratio C. In the training phase, we split the candidate pseudo labels into reliable pseudo labels and unreliable pseudo labels by C according to the confidence scores. As shown in Fig. 3(a), we analyze the influence of different ratios. If we directly set the fixed threshold, the performance is worse than dynamic, which is mainly caused by the lack of box regression optimization for the unlabeled data. However, assigning too many pseudo labels as reliable can be detrimental due to the presence of noisy boxes. In all experiments, we set $C = 0.2$.

Effect of queue K. In ASM, we aim to calculate the similarity of the detection results by maintaining a queue that stores the permutation matrices. Fig. 4(b) demonstrates that as the queue length increases, the detection effectiveness also increases, but at a decreasing rate. To balance computational efficiency and detection effectiveness, we set $K = 20$.

Visualization Analysis. The detection visualization results for the 4C-A and EPV datasets are given in Fig. 3 and more detection visualization results can be illustrated in Fig.1 in Appendix‡. From Fig. 3, we can observe that our method is always consistent with the ground truth, and some misses and misdetections occur in the advanced comparison methods. For example, on the 4C-A dataset, Unbiased Teacher (Chen et al. 2022a) misdetects and misses RIB, Label Matching (Liu, Ma, and Kira 2022) misdetects IVS and CR, Consistent Teacher (Wang et al. 2023b) misdetects the position of RIB, Semi-DETR (Zhang et al. 2023) misdetects the position of RIB DAO, and our method detects all correctly. In addition, as illustrated in Fig. 2 in Appendix‡, we give an instance-level feature visualization by t-SNE (Van der Maaten and Hinton 2008) on each test dataset, and we observe that our method can effectively distinguish the multiple medical structures on each test set compared with the supervised method.

Conclusion

In this paper, we propose a novel semi-supervised learning method for multi-structure detection in ultrasound images, consisting of Anatomical Structure Matching (ASM) and Adaptive Prior Knowledge Transfer (APKT). ASM constructs the graph for the topological representation of the ultrasound image, which builds the connectivity among different detected structures, bridges their relationships, and aligns the teacher and student. In APKT, we explored the prior distribution knowledge of different anatomical structures and assigned more weight to those that are small, irregular, and difficult to detect, which helps to enhance the detection accuracy in complex medical scenarios. We achieved satisfactory results on all five publicly available datasets, demonstrating that our method can serve as a potentially popular tool for medical semi-supervised detection tasks. In the future, we will extend the proposed method to more medical anatomical structure detection scenarios.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China (Grant No. 2022YFF0606302), the National Natural Science Foundation of China (Grant No. 62227808 and 62306003), and the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), under Grant No.GML-KF-24-29.

References

Bai, Y.; Chen, D.; Li, Q.; Shen, W.; and Wang, Y. 2023. Bidirectional copy-paste for semi-supervised medical image segmentation. In *CVPR*, 11514–11524.

Basak, H.; and Yin, Z. 2023. Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In *CVPR*, 19786–19797.

Chaitanya, K.; Erdil, E.; Karani, N.; and Konukoglu, E. 2023. Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. *Medical Image Analysis*, 87: 102792.

Chen, B.; Chen, W.; Yang, S.; Xuan, Y.; Song, J.; Xie, D.; Pu, S.; Song, M.; and Zhuang, Y. 2022a. Label matching semi-supervised object detection. In *CVPR*, 14381–14390.

Chen, B.; Li, P.; Chen, X.; Wang, B.; Zhang, L.; and Hua, X.-S. 2022b. Dense learning based semi-supervised object detection. In *CVPR*, 4815–4824.

Chen, Z.; Zhuo, W.; Wang, T.; Cheng, J.; Xue, W.; and Ni, D. 2023. Semi-Supervised Representation Learning for Segmentation on Medical Volumes and Sequences. *IEEE Transactions on Medical Imaging*.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *NeurIPS*, 26.

Elakkiya, R.; Subramaniyaswamy, V.; Vijayakumar, V.; and Mahanti, A. 2021. Cervical cancer diagnostics healthcare system using hybrid object detection adversarial networks. *IEEE Journal of Biomedical and Health Informatics*, 26(4): 1464–1471.

Gao, Z.; Hong, B.; Li, Y.; Zhang, X.; Wu, J.; Wang, C.; Zhang, X.; Gong, T.; Zheng, Y.; Meng, D.; et al. 2023. A semi-supervised multi-task learning framework for cancer classification with weak annotation in whole-slide images. *Medical Image Analysis*, 83: 102652.

Ge, Y.; Zhou, Q.; Wang, X.; Shen, C.; Wang, Z.; and Li, H. 2023. Point-teaching: Weakly semi-supervised object detection with point annotations. In *AAAI*, volume 37, 667–675.

Girshick, R. 2015. Fast r-cnn. In *ICCV*, 1440–1448.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hua, W.; Liang, D.; Li, J.; Liu, X.; Zou, Z.; Ye, X.; and Bai, X. 2023. SOOD: Towards semi-supervised oriented object detection. In *CVPR*, 15558–15567.

Jeong, J.; Lee, S.; Kim, J.; and Kwak, N. 2019. Consistency-based semi-supervised learning for object detection. *NeurIPS*, 32.

Kar, P.; Chudasama, V.; Onoe, N.; and Wasnik, P. 2023. Revisiting Class Imbalance for End-to-end Semi-Supervised Object Detection. In *CVPR*, 4569–4578.

Lecouat, B.; Chang, K.; Foo, C.-S.; Unnikrishnan, B.; Brown, J. M.; Zenati, H.; Beers, A.; Chandrasekhar, V.; Kalpathy-Cramer, J.; and Krishnaswamy, P. 2018. Semi-supervised deep learning for abnormality classification in retinal images. *arXiv preprint arXiv:1812.07832*.

Li, H.; Wu, Z.; Shrivastava, A.; and Davis, L. S. 2022. Rethinking pseudo labels for semi-supervised object detection. In *AAAI*, volume 36, 1314–1322.

Li, J.; Lin, X.; Zhang, W.; Tan, X.; Li, Y.; Han, J.; Ding, E.; Wang, J.; and Li, G. 2023. Gradient-based sampling for class imbalanced semi-supervised object detection. In *CVPR*, 16390–16400.

Li, W.; Liu, X.; and Yuan, Y. 2022. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *CVPR*, 5291–5300.

Lin, Q.; Zhou, Y.; Shi, S.; Zhang, Y.; Yin, S.; Liu, X.; Peng, Q.; Huang, S.; Jiang, Y.; Cui, C.; et al. 2022. How much can AI see in early pregnancy: A multi-center study of fetus head characterization in week 10–14 in ultrasound using deep learning. *Computer Methods and Programs in Biomedicine*, 226: 107170.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *CVPR*, 2117–2125.

Liu, C.; Zhang, W.; Lin, X.; Zhang, W.; Tan, X.; Han, J.; Li, X.; Ding, E.; and Wang, J. 2023a. Ambiguity-resistant semi-supervised learning for dense object detection. In *CVPR*, 15579–15588.

Liu, L.; Zhang, B.; Zhang, J.; Zhang, W.; Gan, Z.; Tian, G.; Zhu, W.; Wang, Y.; and Wang, C. 2023b. Mixteacher: Mining promising labels with mixed scale teacher for semi-supervised object detection. In *CVPR*, 7370–7379.

Liu, Y.-C.; Ma, C.-Y.; He, Z.; Kuo, C.-W.; Chen, K.; Zhang, P.; Wu, B.; Kira, Z.; and Vajda, P. 2020. Unbiased Teacher for Semi-Supervised Object Detection. In *ICLR*.

Liu, Y.-C.; Ma, C.-Y.; and Kira, Z. 2022. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *CVPR*, 9819–9828.

Loh, D. R.; Yong, W. X.; Yapeter, J.; Subburaj, K.; and Chandramohanadas, R. 2021. A deep learning approach to the screening of malaria infection: Automated and rapid cell counting, object detection and instance segmentation using Mask R-CNN. *Computerized Medical Imaging and Graphics*, 88: 101845.

Luo, X.; Chen, J.; Song, T.; and Wang, G. 2021. Semi-supervised medical image segmentation through dual-task consistency. In *AAAI*, volume 35, 8801–8809.

Luo, X.; Wang, G.; Liao, W.; Chen, J.; Song, T.; Chen, Y.; Zhang, S.; Metaxas, D. N.; and Zhang, S. 2022. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Medical Image Analysis*, 80: 102517.

Nie, Y.; Fang, C.; Cheng, L.; Lin, L.; and Li, G. 2023. Adapting object size variance and class imbalance for semi-supervised object detection. In *AAAI*, volume 37, 1966–1974.

Pu, B.; Li, K.; Li, S.; and Zhu, N. 2021. Automatic fetal ultrasound standard plane recognition based on deep learning and IIoT. *IEEE Transactions on Industrial Informatics*, 17(11): 7771–7780.

Pu, B.; Wang, L.; Yang, J.; He, G.; Dong, X.; Li, S.; Tan, Y.; Chen, M.; Jin, Z.; Li, K.; et al. 2024. M3-UDA: A New Benchmark for Unsupervised Domain Adaptive Fetal Cardiac Structure Detection. In *CVPR*, 11621–11630.

Shehzadi, T.; Hashmi, K. A.; Stricker, D.; and Afzal, M. Z. 2024. Sparse semi-detr: Sparse learnable queries for semi-supervised object detection. In *CVPR*, 5840–5850.

Shi, Y.; Zhang, J.; Ling, T.; Lu, J.; Zheng, Y.; Yu, Q.; Qi, L.; and Gao, Y. 2021. Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(3): 608–620.

Sohn, K.; Zhang, Z.; Li, C.-L.; Zhang, H.; Lee, C.-Y.; and Pfister, T. 2020. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*.

Sohn, K.; Zhang, Z.; Li, C.-L.; Zhang, H.; Lee, C.-Y.; and Pfister, T. 2021. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*.

Tang, Y.; Chen, W.; Luo, Y.; and Zhang, Y. 2021. Humble teachers teach better students for semi-supervised object detection. In *CVPR*, 3132–3141.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 30.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, K.; Zhuang, J.; Li, G.; Fang, C.; Cheng, L.; Lin, L.; and Zhou, F. 2023a. De-biased teacher: Rethinking iou matching for semi-supervised object detection. In *AAAI*, volume 37, 2573–2580.

Wang, X.; Yang, X.; Zhang, S.; Li, Y.; Feng, L.; Fang, S.; Lyu, C.; Chen, K.; and Zhang, W. 2023b. Consistent-teacher: Towards reducing inconsistent pseudo-targets in semi-supervised object detection. In *CVPR*, 3240–3249.

Wu, Y.; Ge, Z.; Zhang, D.; Xu, M.; Zhang, L.; Xia, Y.; and Cai, J. 2022. Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 81: 102530.

Xu, M.; Zhang, Z.; Hu, H.; Wang, J.; Wang, L.; Wei, F.; Bai, X.; and Liu, Z. 2021. End-to-end semi-supervised object detection with soft teacher. In *CVPR*, 3060–3069.

Yang, X.; Wei, Q.; Zhang, C.; Zhou, K.; Kong, L.; and Jiang, W. 2020. Colon polyp detection and segmentation based on improved MRCNN. *IEEE Transactions on Instrumentation and Measurement*, 70: 1–10.

You, C.; Dai, W.; Min, Y.; Liu, F.; Clifton, D.; Zhou, S. K.; Staib, L.; and Duncan, J. 2024. Rethinking semi-supervised medical image segmentation: A variance-reduction perspective. *NeurIPS*, 36.

Yu, L.; Wang, S.; Li, X.; Fu, C.-W.; and Heng, P.-A. 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *MICCAI*, 605–613. Springer.

Zeng, Q.; Xie, Y.; Lu, Z.; and Xia, Y. 2023. Pefat: Boosting semi-supervised medical image classification via pseudo-loss estimation and feature adversarial training. In *CVPR*, 15671–15680.

Zhang, J.; Lin, X.; Zhang, W.; Wang, K.; Tan, X.; Han, J.; Ding, E.; Wang, J.; and Li, G. 2023. Semi-detr: Semi-supervised object detection with detection transformers. In *CVPR*, 23809–23818.

Zhang, L.; Sun, Y.; and Wei, W. 2023. Mind the gap: Polishing pseudo labels for accurate semi-supervised object detection. In *AAAI*, volume 37, 3463–3471.

Zhang, W.; Zhu, L.; Hallinan, J.; Zhang, S.; Makmur, A.; Cai, Q.; and Ooi, B. C. 2022. Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In *CVPR*, 20666–20676.

Zhang, Y.; Yang, L.; Chen, J.; Fredericksen, M.; Hughes, D. P.; and Chen, D. Z. 2017. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *MICCAI*, 408–416. Springer.

Zhou, H.; Ge, Z.; Liu, S.; Mao, W.; Li, Z.; Yu, H.; and Sun, J. 2022. Dense teacher: Dense pseudo-labels for semi-supervised object detection. In *ECCV*, 35–50. Springer.

Zhou, H.-Y.; Wang, C.; Li, H.; Wang, G.; Zhang, S.; Li, W.; and Yu, Y. 2021a. SSMD: semi-supervised medical image detection with adaptive consistency and heterogeneous perturbation. *Medical Image Analysis*, 72: 102117.

Zhou, Q.; Yu, C.; Wang, Z.; Qian, Q.; and Li, H. 2021b. Instant-teaching: An end-to-end semi-supervised object detection framework. In *CVPR*, 4081–4090.

Zhou, Z.-H.; and Zhou, Z.-H. 2021. Semi-supervised learning. *Machine Learning*, 315–341.