

# Active Source-free Domain Adaptation in Open-set Medical Image Segmentation via Decomposed Uncertainty and Prototype Discrepancy

Jin Yang<sup>1</sup>, Yichi Zhang<sup>2</sup>, Peijie Qiu<sup>3</sup>, and Xiaobing Yu<sup>3</sup>

<sup>1</sup> Department of Radiation Oncology, Icahn School of Medicine at Mount Sinai, New York, NY, USA, 10029

<sup>2</sup> School of Data Science, Fudan University, Shanghai, China

<sup>3</sup> Mallinckrodt Institute of Radiology, Washington University School of Medicine in St. Louis, St. Louis, MO, USA, 63110  
yang.jin@wustl.edu

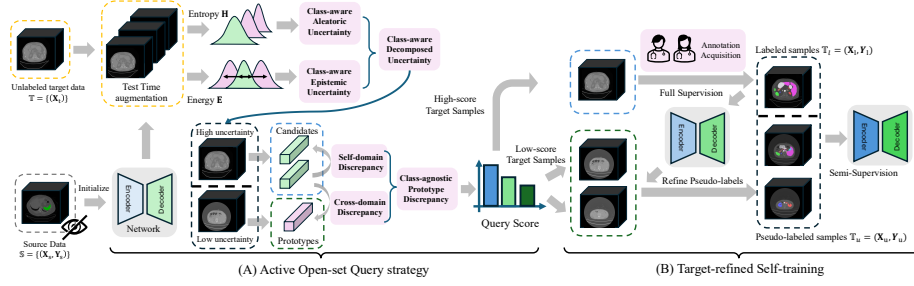
**Abstract.** Deep learning (DL) methods are challenged to demonstrate robust performance across different segmentation datasets due to domain shifts, but active domain adaptation techniques enhance their generalization performance by querying a few samples from target domains for adaptation training. However in clinical practice, target domains often include private classes of new anatomical structures or pathologies that are not presented in the source data, and existing methods implement closed-set segmentation where source and target domains have the same segmentation classes. Additionally, source data are often inaccessible during adaptation due to strict data privacy regulations. To address these limitations, we propose an Active Source-free Open-set Domain Adaptation (ASFOSDA) method which is the first work to implement active learning for adapting DL models in open-set medical image segmentation without the access to source data. This method employs an active open-set query strategy to select the most informative target samples for training models based on Class-aware Decomposed Uncertainty (CDU) and Class-agnostic Prototype Discrepancy (CPD). CDU measures sample aleatoric uncertainty and model epistemic uncertainty by employing test time augmentation in stochastic processes. CPD measures cross-domain and self-domain discrepancy for selecting diverse samples. Subsequently, to boost the adaptation performance by enhancing training samples, a Target-refined Self-training strategy is proposed to generate high-quality pseudo labels for unselected samples, thus combining them with labeled samples for a semi-supervised training. We evaluated our method on cross-domain open-set volumetric medical image segmentation tasks, and it outperformed state-of-the-art adaptation methods.

**Keywords:** Active Learning · Open-set Domain Adaptation · Medical Image Segmentation · Self-training.

## 1 Introduction

Deep learning (DL) methods achieve great success in automatic organ segmentation, but they are limited from generalizing across different data sources due to domain shifts [5,21]. Domain adaptation (DA) methods enhance DL methods' generalizability to address these shifts by transferring knowledge from labeled source domains to unlabeled target domains [26]. However, most DA methods implement "closed set" segmentation [18,22]. In clinical practice, target domains often include private or new classes of anatomical structures or pathologies, and these classes are not delineated in the source data. Thus, existing DA methods may lead to negative transfer when aligning features of unknown classes from source to target domains, degrading model adaptation performance [17]. Additionally, strict data privacy regulations may lead to the inaccessibility of source data. Source-free domain adaptation (SFDA) problems are those of adapting models to target domains without the access to the source data [2], and active learning (AL) is their feasible solution [24,4,28]. AL employs a query strategy to select informative samples for annotation from target domains, thus training models to maximize adaptation performance with minimal annotation overhead [16,10,25,23]. However, these active source-free domain adaptation (ASFDA) methods lack mechanisms to evaluate the information level of new classes, and no work is implemented to adapt models across domains in open-set medical image segmentation.

To address these limitations, we propose a novel **Active Source-free Open-set Domain Adaptation** (ASFOSDA) method for medical image segmentation. It employs a novel **active open-set query strategy** to select informative samples for model adaptation by evaluating the information levels of known and unknown classes based on two metrics: **Class-aware Decomposed Uncertainty** (CDU) and **Class-agnostic Prototype Discrepancy** (CPD). First, CDU measures class-aware aleatoric uncertainty (CAU) and class-aware epistemic uncertainty (CEU) from stochastic processes by employing test time augmentation within the target domain. CAU calculates predictive entropy from probability maps of target classes that are unknown to source data in the original and augmented images. Thus, querying samples by CAU to train models improves their reliability on modeling target patterns. CEU calculates variances of predictive energy for unknown classes between the original and augmented images, and querying samples by CEU for training models improves their capabilities of generalizing across new domains. Second, CPD selects representative samples from high-uncertainty ones to train models for further learning target data manifold. CPD measures source-target domain discrepancy by evaluating semantic distance between low-uncertainty and high-uncertainty samples; to avoid selecting samples with redundant knowledge, it also measures semantic dissimilarity among high-uncertainty target samples. Additionally, to improve adaptation performance by enhancing training samples, we implement a **Target-refined Self-training** to generate high-quality pseudo labels for unselected samples, thus adapting models with labeled samples and pseudo-labeled samples via a semi-supervised training. We evaluated the effectiveness of ASFOSDA



**Fig. 1. Active Source-free Open-set Domain Adaptation in medical image segmentation.** (A) **Active Open-set Query strategy:** After test time augmentation was employed, Class-aware Aleatoric Uncertainty and Class-aware Epistemic Uncertainty were measured, and combined to measure Class-aware Decomposed Uncertainty. Class-agnostic Prototype Discrepancy measured self-domain discrepancy and cross-domain discrepancy among low- and high-uncertainty samples. (B) **Target-refined Self-training:** Target samples with high scores were selected for annotation  $\mathbb{T}_l = \{(\mathbf{X}_l, \mathbf{Y}_l)\}$  and training the network, and then this network regenerated and refined pseudo-labels for unselected samples with low scores  $\mathbb{T}_u = \{(\mathbf{X}_u, \mathbf{Y}_u)\}$ . Finally, labeled and pseudo-labeled samples were combined for semi-supervised training.

on 3D volumetric multi-organ segmentation in CT and MR images by adapting models to segment organs with known and unknown classes. Our method achieved superior performance than state-of-the-art (SOTA) SFDA and ASFDA methods.

## 2 Methods

A source domain  $\mathbb{S}$  includes  $\mathcal{N}_s$  images  $\mathbf{X}_s = \{x_s | 1 \leq s \leq \mathcal{N}_s\}$  with annotation  $\mathbf{Y}_s = \{y_s | 1 \leq s \leq \mathcal{N}_s\}$  of  $\mathcal{C}_s$  organs, and a target domain  $\mathbb{T}$  includes  $\mathcal{N}_t$  images  $\mathbf{X}_t = \{x_t | 1 \leq t \leq \mathcal{N}_t\}$  without annotation. The goal of ASFOSDA is to derive a segmentation model  $\mathcal{F}(\boldsymbol{\theta})$  with source prior knowledge of segmenting  $\mathcal{C}_s$  organs to generate predicted segmentation masks  $\mathbf{Y}_t = \{y_t | 1 \leq t \leq \mathcal{N}_t\}$  of  $\mathcal{C}_t$  organs in target images ( $\mathcal{C}_t > \mathcal{C}_s$ ), thus minimizing the risk on the target domain without access to the source domain (Fig. 1). Before AL query, the model  $\mathcal{F}(\boldsymbol{\theta})$  is initialized by source data  $\mathbb{S} = \{(\mathbf{X}_s, \mathbf{Y}_s)\}$ . A labeled target set is empty as  $\mathbb{T}_l = \emptyset$ , and an unlabeled target set is initialized on the whole target set as  $\mathbb{T}_u = \mathbb{T} = \{\mathbf{X}_t\}$ . Subsequently, with a pre-defined query budget  $\mathcal{N}_l$  ( $\mathcal{N}_l \ll \mathcal{N}_t$ ), ASFOSDA employs an active open-set query strategy to select  $\mathcal{N}_l$  images  $\mathbf{X}_l = \{x_l | 1 \leq l \leq \mathcal{N}_l\}$  for annotation  $\mathbf{Y}_l = \{y_l | 1 \leq l \leq \mathcal{N}_l\}$  into the labeled target set  $\mathbb{T}_l$  from the unlabeled target set  $\mathbb{T}_u$ , where  $\mathbb{T}_l = \mathbb{T}_l \cup \{(\mathbf{X}_l, \mathbf{Y}_l)\}$  and  $\mathbb{T}_u = \mathbb{T}_u \setminus \{\mathbf{X}_l\} = \{\mathbf{X}_u\}$ . Subsequently, a Target-refined Self-training strategy adapts the model by selected samples with their annotation  $\{(\mathbf{X}_l, \mathbf{Y}_l)\}$  and unselected samples with their refined pseudo labels  $\{(\mathbf{X}_u, \mathbf{Y}_u)\}$  in a semi-supervised training.

## 2.1 Active open-set query strategy

**Class-aware Decomposed Uncertainty.** CDU evaluates predictive uncertainty by decomposing it into the measurement of class-aware aleatoric uncertainty (CAU) and class-aware epistemic uncertainty (CEU). CAU and CEU are measured from stochastic processes by employing test time augmentation (TTA). For an unlabeled target sample  $x_u$  from the set  $\mathbf{X}_u$ , TTA generates its augmented variations  $\mathbf{x}_u^M = \{x_u^1, \dots, x_u^m\}$  by applying image augmentation techniques, including flipping, rotations, brightness shifts, contrast shifts, and Gaussian noising.

The source-trained segmentation model  $\mathcal{F}(\boldsymbol{\Theta})$  is utilized to generate probability maps of the original sample as  $\mathbf{P}(x_u; \boldsymbol{\Theta}) = \{p_1, \dots, p_C\} \in \mathbb{R}^{C \times H \times W \times D}$  for all target segmentation classes  $\mathcal{C}_t$ , where  $H, W, D$  represent height, width, depth of the image, respectively. The number of channels  $C$  is equal to the number of segmented organs  $\mathcal{C}_t$  ( $C = |\mathcal{C}_t|$ ), and the target domain includes known source classes  $\mathcal{C}_s$  and unknown target classes  $\mathcal{C}_u = \mathcal{C}_t \setminus \mathcal{C}_s = \{c_1^u, \dots, c_{t-s}^u\}$ . Here, the known source classes refer to classes present in the source domain, whereas the unknown target classes correspond to classes that exist only in the target domain and are absent from the source domain. Class-aware predictive entropy  $H(x_u)$  of the original sample  $x_u$  is calculated for unknown classes  $\mathcal{C}_u$  from corresponding probability maps  $\mathbf{P}_u(x_u; \boldsymbol{\Theta}) = \{p_1^u, \dots, p_{t-s}^u\}$  as

$$H(x_u) = -\mathbf{P}_u(x_u; \boldsymbol{\Theta}) \log \mathbf{P}_u(x_u; \boldsymbol{\Theta}) = -\sum_{i=1}^{t-s} p_i^u \log p_i^u. \quad (1)$$

Subsequently, the model generates probability maps  $\mathbf{P}(x_u^m; \boldsymbol{\Theta}) \in \mathbb{R}^{C \times H \times W \times D}$  for augmented variations  $\mathbf{x}_u^M = \{x_u^1, \dots, x_u^m\}$ , and their class-aware predictive entropy  $H(\mathbf{x}_u^M)$  is calculated for unknown classes  $\mathcal{C}_u$  from their corresponding probability maps  $\{\mathbf{P}_1(x_u^1; \boldsymbol{\Theta}), \dots, \mathbf{P}_m(x_u^m; \boldsymbol{\Theta})\} = \{\{p_1^1, \dots, p_{t-s}^1\}, \dots, \{p_1^m, \dots, p_{t-s}^m\}\}$  as

$$H(\mathbf{x}_u^M) = \sum_{j=1}^m H(x_u^j) = -\sum_{j=1}^m \mathbf{P}_j(x_u^j; \boldsymbol{\Theta}) \log \mathbf{P}_j(x_u^j; \boldsymbol{\Theta}) = -\sum_{j=1}^m \sum_{i=1}^{t-s} p_i^j \log p_i^j. \quad (2)$$

Finally, CAU is calculated by averaging the predictive entropy of the original sample  $x_u$  and its augmented variations  $\mathbf{x}_u^M = \{x_u^1, \dots, x_u^m\}$  as

$$\text{CAU}(x_u) = \frac{1}{1+m} [H(x_u) + H(\mathbf{x}_u^M)]. \quad (3)$$

Energy measures how well models align with the domain data, thus evaluating epistemic uncertainty [19]. The model  $\mathcal{F}(\boldsymbol{\Theta})$  generates predicted logits of the original sample  $\mathbf{Z}(x_u; \boldsymbol{\Theta}) = \{z_1, \dots, z_C\} \in \mathbb{R}^{C \times H \times W \times D}$  for all target segmentation classes  $\mathcal{C}_t$ . Class-aware predictive energy  $E(x_u)$  of the original sample  $x_u$  is calculated for unknown classes  $\mathcal{C}_u$  from logits  $\mathbf{Z}_u(x_u; \boldsymbol{\Theta}) = \{z_1^u, \dots, z_{t-s}^u\}$  as

$$E(x_u) = -T \cdot \log e^{\mathbf{Z}_u(x_u; \boldsymbol{\Theta})/T} = -T \cdot \log \sum_{i=1}^{t-s} e^{z_i^u/T}. \quad (4)$$

Subsequently, the model generates predicted logits  $\mathbf{Z}(x_u^m; \boldsymbol{\Theta}) \in \mathbb{R}^{C \times H \times W \times D}$  for augmented variations  $\mathbf{x}_u^M = \{x_u^1, \dots, x_u^m\}$ , and their class-aware predictive energy  $E(\mathbf{x}_u^M)$  is calculated for unknown classes  $\mathcal{C}_t^u$  from their corresponding logit maps  $\{\mathbf{Z}_1(x_u^1; \boldsymbol{\Theta}), \dots, \mathbf{Z}_m(x_u^m; \boldsymbol{\Theta})\} = \{\{z_1^1, \dots, z_{t-s}^1\}, \dots, \{z_1^m, \dots, z_{t-s}^m\}\}$  as

$$E(\mathbf{x}_u^M) = \sum_{j=1}^m E(x_u^j) = - \sum_{j=1}^m T \cdot \log e^{\mathbf{Z}_j(x_u^j; \boldsymbol{\Theta})/T} = - \sum_{j=1}^m T \cdot \log \sum_{i=1}^{t-s} e^{z_i^j/T}. \quad (5)$$

$T$  is a temperature parameter, and set to 1 as commonly used. The average energy among the original sample and its augmented variations is calculated as  $\bar{E} = \frac{1}{1+m} (E(x_u) + E(\mathbf{x}_u^M))$ . Thus, CEU is calculated by calculating the variances of predictive energy between the original sample  $x_u$  and its augmented variations  $\mathbf{x}_u^M = \{x_u^1, \dots, x_u^m\}$  as

$$\text{CEU}(x_u) = \frac{1}{1+m} \left[ (E(x_u) - \bar{E})^2 + \sum_{j=1}^m (E(x_u^j) - \bar{E})^2 \right]. \quad (6)$$

Higher CAU and CEU scores demonstrate higher aleatoric and epistemic uncertainty, respectively. When combining them to calculate the CDU score, we apply min-max normalization followed by a uniform Quantile transformation (collectively denoted as “*Norm*”) to ensure that the two scores are mapped to the same range and follow uniform distributions as

$$\text{CDU}(x_u) = \text{Norm}(\text{CAU}(x_u)) + \text{Norm}(\text{CEU}(x_u)). \quad (7)$$

**Class-agnostic Prototype Discrepancy.** A source-trained model demonstrates low predictive uncertainty on target samples if it has been equipped with sufficient knowledge to extract accurate patterns from these samples. Thus, these samples are identified as source-like samples, and their feature embeddings generated by the source-trained model represent source domain knowledge. Thus, the encoder of the segmentation model  $\mathcal{F}(\boldsymbol{\Theta})$  is employed to generate feature embeddings of  $K$  target samples with low CDU score as prototypes  $\{\mathbf{L}_1, \dots, \mathbf{L}_K\}$  to represent source knowledge. In contrast, target samples with high uncertainty exhibit a significant distribution shift from the source domain, and these target samples are identified as target-specific samples to represent target knowledge. Thus, querying these samples for training enables models to learn target-specific knowledge for adaptation. The encoder generates feature embeddings of  $K$  target samples with high CDU score as target query candidates  $\{\mathbf{H}_1, \dots, \mathbf{H}_K\}$ . Subsequently, to select samples for describing knowledge gaps between source and target domains, the cross-domain discrepancy (CDD) is measured for high CDU sample  $x_u$  by calculating the semantic distance between its candidate  $\mathbf{H}_u$  and all prototypes  $\{\mathbf{L}_1, \dots, \mathbf{L}_K\}$  as

$$\text{CDD}(x_u) = \frac{1}{K} \sum_{j=1}^K \left( 1 - \frac{\mathbf{H}_u \cdot \mathbf{L}_j}{\|\mathbf{H}_u\| \cdot \|\mathbf{L}_j\|} \right). \quad (8)$$

To avoid selecting samples with redundant knowledge within target domain, the self-domain discrepancy (SDD) for the sample  $x_u$  is measured by calculating the semantic dissimilarity between its candidate  $\mathbf{H}_u$  and other candidates as

$$\text{SDD}(x_u) = \frac{1}{K-1} \sum_{i=1}^{K-1} \left(1 - \frac{\mathbf{H}_u \cdot \mathbf{H}_i}{\|\mathbf{H}_u\| \cdot \|\mathbf{H}_i\|}\right). \quad (9)$$

The CPD score of the sample  $x_u$  is calculated by combining CDD and SDD as

$$\text{CPD}(x_u) = \text{CDD}(x_u) + \text{SDD}(x_u). \quad (10)$$

## 2.2 Target-refined Self-training

To improve adaptation performance by enhancing training samples, the source-trained model is adapted by queried samples with their annotation and unselected samples with their pseudo labels via a semi-supervised training rather than solely using queried samples. However, pseudo labels generated by the source-trained model are in low-quality, and using these pseudo-labeled samples degrades adaptation performance. To address these limitations, the Target-refined self-training strategy implements semi-supervised adaptation in two stages. First, it trains the segmentation model  $\mathcal{F}(\Theta)$  with selected target samples with their annotation  $\{\mathbf{X}_l, \mathbf{Y}_l\}$  on the labeled target set  $\mathbb{T}_l$  via full supervision with a supervised loss function  $\mathcal{L}_{sup}(\mathbf{Y}_l, \mathbf{P}_l)$ . Subsequently, the target-trained model is employed to regenerate and refine pseudo-labels for unlabeled target samples  $\mathbf{X}_u$  on the unlabeled target set  $\mathbb{T}_u$ . These pseudo-labeled images  $\{\mathbf{X}_u, \mathbf{Y}_u\}$  and labeled images  $\{\mathbf{X}_l, \mathbf{Y}_l\}$  are combined to further adapt the model in a semi-supervised training via a joint loss function  $\mathcal{L} = \mathcal{L}_{sup}(\mathbf{Y}_l, \mathbf{P}_l) + \mathcal{L}_{unsup}(\mathbf{Y}_u, \mathbf{P}_u)$ .

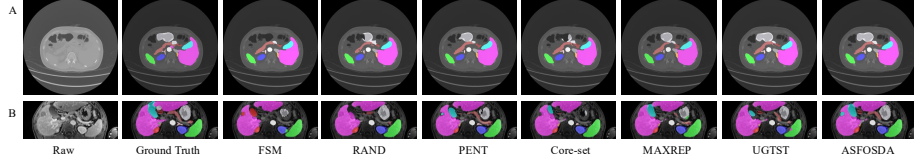
## 3 Experiments

**Datasets.** (i) **MSD Spleen:** 41 CT scans with spleen annotations [1]. (ii) **FLARE:** 361 CT scans with annotations of four abdominal organs (spleen, liver, kidneys, and pancreas) [11]. (iii) **AMOS CT:** 300 CT scans with annotations of 14 organs (spleen, liver, kidneys, pancreas, gall bladder, esophagus, stomach, aorta, postcava, right adrenal gland, left adrenal gland, duodenum, bladder, and prostate/uterus), where right and left kidneys were merged to maintain consistency with FLARE. (iv) **Abdominal MR:** 30 MR images with annotations of eight anatomical structures, including spleen, right kidney, left kidney, liver, pancreas, gallbladder, aorta, and inferior vena cava [27]. (v) **AMOS MRI:** 60 MR scans with 13 organ categories where bladder and prostate/uterus were excluded due to extremely imbalanced annotation [7].

**Implementation details.** The experiments were implemented using PyTorch on NVIDIA Tesla A100 PCI-E Passive Single GPU. The 3D U-Net and Swin UNETR were used as segmentation networks [12,6]. The combination of dice loss

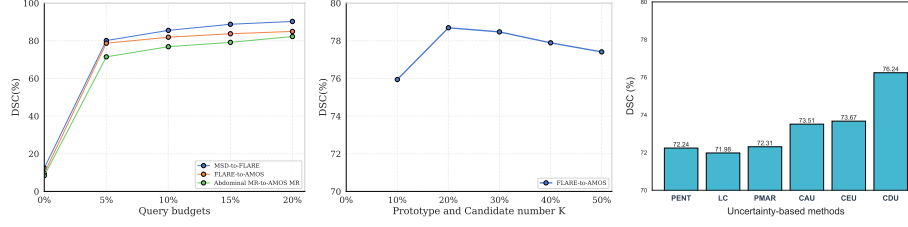
**Table 1.** Performance comparison in open-set medical image segmentation. The results were reported as Mean $\pm$ SD. **Bold** represents the best results. (\*:  $p < 0.01$  with the Mann-Whitney U test between ASFOSDA and other SFDA and ASFDA methods).

| Net        | Task  | Methods            | Open-set CT Segmentation         |                                 |                                  |                                 | Open-set MR Segmentation           |                                 |
|------------|-------|--------------------|----------------------------------|---------------------------------|----------------------------------|---------------------------------|------------------------------------|---------------------------------|
|            |       |                    | MSD Spleen $\rightarrow$ FLARE   |                                 | FLARE $\rightarrow$ AMOS CT      |                                 | Abdominal MR $\rightarrow$ AMOS MR |                                 |
|            |       |                    | DSC( $\%$ ) $\uparrow$           | 95HD(mm) $\downarrow$           | DSC( $\%$ ) $\uparrow$           | 95HD(mm) $\downarrow$           | DSC( $\%$ ) $\uparrow$             | 95HD(mm) $\downarrow$           |
| U-Net      |       | Source Only        | 12.36 $\pm$ 12.84                | 62.32 $\pm$ 20.72               | 9.65 $\pm$ 10.70                 | 68.59 $\pm$ 25.51               | 8.35 $\pm$ 10.23                   | 70.34 $\pm$ 30.34               |
|            |       | Target Supervision | 95.21 $\pm$ 1.40                 | 1.31 $\pm$ 1.12                 | 92.24 $\pm$ 1.64                 | 1.84 $\pm$ 1.13                 | 90.16 $\pm$ 1.21                   | 1.91 $\pm$ 1.17                 |
|            | SFDA  | DPL [3]            | 40.92 $\pm$ 15.45                | 44.45 $\pm$ 15.21               | 38.94 $\pm$ 17.98                | 48.36 $\pm$ 17.76               | 35.51 $\pm$ 18.53                  | 50.34 $\pm$ 19.41               |
|            |       | FSM [20]           | 43.36 $\pm$ 15.22                | 40.07 $\pm$ 14.23               | 40.58 $\pm$ 16.54                | 45.42 $\pm$ 16.91               | 36.63 $\pm$ 16.72                  | 49.27 $\pm$ 18.71               |
|            |       | TT-SFDA [14]       | 47.62 $\pm$ 14.39                | 38.31 $\pm$ 14.98               | 43.12 $\pm$ 15.19                | 40.18 $\pm$ 16.55               | 40.56 $\pm$ 16.75                  | 45.64 $\pm$ 18.38               |
|            |       | RAND               | 68.18 $\pm$ 10.90                | 10.94 $\pm$ 12.11               | 66.96 $\pm$ 12.46                | 11.97 $\pm$ 12.25               | 61.24 $\pm$ 13.43                  | 13.18 $\pm$ 15.63               |
|            | ASFDA | PENT [15]          | 72.24 $\pm$ 10.65                | 8.24 $\pm$ 9.26                 | 71.81 $\pm$ 10.31                | 8.67 $\pm$ 9.36                 | 65.83 $\pm$ 11.75                  | 11.92 $\pm$ 13.56               |
|            |       | LC [8]             | 71.98 $\pm$ 10.72                | 8.45 $\pm$ 9.38                 | 71.64 $\pm$ 11.10                | 8.69 $\pm$ 9.83                 | 65.73 $\pm$ 11.97                  | 12.17 $\pm$ 14.15               |
|            |       | PMAR [15]          | 72.31 $\pm$ 10.68                | 8.20 $\pm$ 9.81                 | 71.20 $\pm$ 11.78                | 8.87 $\pm$ 10.11                | 65.35 $\pm$ 12.27                  | 12.35 $\pm$ 14.17               |
|            |       | Core-set [13]      | 70.86 $\pm$ 10.48                | 9.01 $\pm$ 9.24                 | 70.28 $\pm$ 11.99                | 9.28 $\pm$ 10.37                | 65.25 $\pm$ 12.84                  | 12.46 $\pm$ 14.31               |
|            |       | MAXREP [24]        | 73.88 $\pm$ 9.32                 | 7.26 $\pm$ 8.56                 | 73.32 $\pm$ 10.46                | 7.56 $\pm$ 9.86                 | 67.78 $\pm$ 10.45                  | 10.86 $\pm$ 11.83               |
|            |       | CENT [9]           | 74.56 $\pm$ 9.11                 | 6.80 $\pm$ 7.91                 | 73.48 $\pm$ 9.65                 | 7.48 $\pm$ 8.82                 | 68.16 $\pm$ 10.32                  | 10.52 $\pm$ 10.34               |
|            |       | UGTST [10]         | 74.76 $\pm$ 8.66                 | 6.52 $\pm$ 7.15                 | 73.93 $\pm$ 9.72                 | 7.04 $\pm$ 7.51                 | 68.62 $\pm$ 10.01                  | 10.16 $\pm$ 9.99                |
|            |       | ASFOSDA (Ours)     | <b>80.11<math>\pm</math>8.01</b> | <b>5.20<math>\pm</math>6.22</b> | <b>78.69<math>\pm</math>8.87</b> | <b>5.75<math>\pm</math>6.38</b> | <b>71.46<math>\pm</math>8.42</b>   | <b>8.78<math>\pm</math>9.82</b> |
| Swin UNETR |       | Source Only        | 11.93 $\pm$ 13.04                | 64.84 $\pm$ 21.77               | 9.30 $\pm$ 10.91                 | 69.00 $\pm$ 26.24               | 8.16 $\pm$ 10.57                   | 72.31 $\pm$ 32.42               |
|            |       | Target Supervision | 94.88 $\pm$ 1.56                 | 1.51 $\pm$ 1.35                 | 91.88 $\pm$ 1.78                 | 1.89 $\pm$ 1.21                 | 90.05 $\pm$ 1.37                   | 1.93 $\pm$ 1.22                 |
|            | SFDA  | DPL [3]            | 39.93 $\pm$ 16.91                | 46.56 $\pm$ 16.76               | 37.69 $\pm$ 18.66                | 49.33 $\pm$ 17.89               | 33.83 $\pm$ 18.88                  | 55.34 $\pm$ 21.34               |
|            |       | FSM [20]           | 42.54 $\pm$ 16.44                | 41.53 $\pm$ 15.70               | 39.18 $\pm$ 17.42                | 47.62 $\pm$ 16.87               | 34.12 $\pm$ 18.35                  | 53.62 $\pm$ 20.66               |
|            |       | TT-SFDA [14]       | 47.37 $\pm$ 15.51                | 40.82 $\pm$ 14.32               | 42.81 $\pm$ 17.28                | 41.78 $\pm$ 16.61               | 38.35 $\pm$ 17.62                  | 48.55 $\pm$ 20.45               |
|            |       | RAND               | 67.91 $\pm$ 11.81                | 11.85 $\pm$ 12.45               | 65.87 $\pm$ 13.11                | 12.05 $\pm$ 12.36               | 60.84 $\pm$ 14.21                  | 15.34 $\pm$ 16.46               |
|            | ASFDA | PENT [15]          | 71.52 $\pm$ 11.33                | 8.77 $\pm$ 10.25                | 70.85 $\pm$ 11.22                | 9.03 $\pm$ 9.92                 | 65.29 $\pm$ 12.55                  | 12.38 $\pm$ 14.02               |
|            |       | LC [8]             | 71.44 $\pm$ 11.28                | 8.82 $\pm$ 10.90                | 70.78 $\pm$ 12.09                | 9.12 $\pm$ 9.98                 | 64.78 $\pm$ 12.78                  | 13.26 $\pm$ 14.15               |
|            |       | PMAR [15]          | 71.85 $\pm$ 11.12                | 8.71 $\pm$ 10.97                | 70.17 $\pm$ 12.91                | 9.48 $\pm$ 10.32                | 64.52 $\pm$ 13.32                  | 13.76 $\pm$ 15.81               |
|            |       | Core-set [13]      | 70.59 $\pm$ 10.68                | 9.25 $\pm$ 9.70                 | 69.99 $\pm$ 12.63                | 10.31 $\pm$ 10.88               | 64.24 $\pm$ 13.71                  | 13.82 $\pm$ 15.85               |
|            |       | MAXREP [24]        | 72.62 $\pm$ 10.51                | 7.82 $\pm$ 8.69                 | 72.34 $\pm$ 10.58                | 8.32 $\pm$ 9.72                 | 66.62 $\pm$ 12.23                  | 11.65 $\pm$ 12.89               |
|            |       | CENT [9]           | 73.85 $\pm$ 10.05                | 7.44 $\pm$ 8.72                 | 72.68 $\pm$ 10.16                | 8.23 $\pm$ 9.60                 | 66.94 $\pm$ 11.44                  | 11.54 $\pm$ 10.82               |
|            |       | UGTST [10]         | 73.99 $\pm$ 9.98                 | 6.88 $\pm$ 7.94                 | 72.92 $\pm$ 9.98                 | 7.88 $\pm$ 8.15                 | 67.73 $\pm$ 11.12                  | 10.93 $\pm$ 10.32               |
|            |       | ASFOSDA (Ours)     | <b>79.81<math>\pm</math>9.24</b> | <b>5.64<math>\pm</math>6.16</b> | <b>77.95<math>\pm</math>8.95</b> | <b>5.97<math>\pm</math>6.57</b> | <b>70.35<math>\pm</math>9.46</b>   | <b>9.32<math>\pm</math>9.73</b> |



**Fig. 2.** Qualitative comparison among results of SFDA, ASFDA, and our methods by adapting U-Net to (A) AMOS CT and (B) AMOS MR datasets.

and cross entropy loss was used in both supervised training during initialization and self-training during adaptation. Networks were initialized on source data for 1000 epochs and trained on target data for 500 epochs at each stage of self-training. The AdamW was utilized for optimization. The initial learning rate was set to 0.001 and decayed in a polynomial scheduler with a power of 0.9. The batch size was 2. Raw volumes were z-score normalized and scaled to the patches with the dimension of  $96 \times 96 \times 96$ . The performance was evaluated using Dice Similarity Coefficient (DSC; $\%$ ) and 95% Hausdorff distance (95HD;mm).



**Fig. 3.** Ablation studies of key hyper-parameters and components. (A) Impact of different target budgets on U-Net adaptation. (B) Impact of different number of prototypes and candidates  $K$  using 5% budget for U-Net FLARE-to-AMOS adaptation. (C) Performance comparison of uncertainty methods in U-Net MSD-to-FLARE adaptation.

**Table 2.** The results of ablation study on the active open-set query strategy for U-Net adaptation with 5% samples queried. **Bold** represents the best results. (\*:  $p < 0.01$  with the Mann-Whitney U test between different components).

| Components   |              |              |               | MSD Spleen $\rightarrow$ FLARE |                        | FLARE $\rightarrow$ AMOS CT |                        | Abdominal MR $\rightarrow$ AMOS MR |                        |
|--------------|--------------|--------------|---------------|--------------------------------|------------------------|-----------------------------|------------------------|------------------------------------|------------------------|
| CAU          | CEU          | CPD          | Self-training | DSC(%) $\uparrow$              | 95HD(mm) $\downarrow$  | DSC(%) $\uparrow$           | 95HD(mm) $\downarrow$  | DSC(%) $\uparrow$                  | 95HD(mm) $\downarrow$  |
| $\checkmark$ |              |              |               | 73.51 $\pm$ 9.73               | 7.42 $\pm$ 8.61        | 72.71 $\pm$ 10.86           | 8.17 $\pm$ 9.34        | 66.20 $\pm$ 11.27                  | 11.81 $\pm$ 12.44      |
|              | $\checkmark$ |              |               | 73.67 $\pm$ 9.58               | 7.34 $\pm$ 8.24        | 72.82 $\pm$ 10.72           | 8.08 $\pm$ 9.26        | 66.32 $\pm$ 11.41                  | 11.73 $\pm$ 12.51      |
|              | $\checkmark$ | $\checkmark$ |               | 76.24 $\pm$ 8.69               | 6.06 $\pm$ 7.12        | 74.30 $\pm$ 9.11            | 6.85 $\pm$ 7.40        | 69.28 $\pm$ 9.89                   | 9.66 $\pm$ 10.02       |
|              | $\checkmark$ | $\checkmark$ | $\checkmark$  | 78.46 $\pm$ 8.34               | 5.83 $\pm$ 6.68        | 76.48 $\pm$ 8.82            | 6.45 $\pm$ 7.12        | 70.55 $\pm$ 9.44                   | 9.02 $\pm$ 9.91        |
| $\checkmark$ | $\checkmark$ | $\checkmark$ | $\checkmark$  | <b>80.11</b> $\pm$ 8.01        | <b>5.20</b> $\pm$ 6.22 | <b>78.69</b> $\pm$ 8.87     | <b>5.75</b> $\pm$ 6.38 | <b>71.46</b> $\pm$ 8.42            | <b>8.78</b> $\pm$ 9.82 |

**Experimental results.** We compared the performance of ASFOSDA with (i) Source Only: lower bound performance where source-trained models were employed to target domains without adaptation; (ii) Target Supervision: upper bound performance where source-trained models were trained by all target samples with annotation; (iii) SFDA methods: DPL [3], FSM [20], and TT-SFDA [14]; (iv) ASFDA methods with 5% target samples queried: Random Selection (RAND), Predictive Entropy (PENT) [15], Least Confidence (LC) [8], Probability Margin (PMAR) [15], Core-set [13], Maximum Representation (MAXREP) [24], Cluster Entropy (CENT) [9], and UGTST [10]. For better adaptation, one target sample with annotation was selected randomly for source initialization in all methods. To ensure a fair comparison, all methods were initialized using the same settings, and were trained using the same adaptation protocols.

With a 5% target sample annotation budget, our method consistently outperformed existing SFDA and ASFDA methods across both 3D U-Net and Swin UNETR (Table 1 and Fig. 2). In the MSD Spleen-to-FLARE (single to four organs) transition, our approach achieved over 84% of the supervised upper bound for both networks. Similarly, for the FLARE-to-AMOS CT (four to 14 organ) task, ASFOSDA enabled the models to reach 85.31% and 84.84% of their respective upper bounds. Finally, in MR image adaptation (eight to 13 organs), ASFOSDA enabled U-Net and Swin UNETR to maintain robust performance, achieving 82.62% and 78.12% of the upper bound.

**Ablation study.** Comparing performance of different components from Active Open-set Query strategy in three adaptation tasks demonstrated its effectiveness (Table 2). The performance of U-Net demonstrated consistent improvement as the query budget increased from 0% to 20% (Fig. 3(A)). Employing 20% prototypes and candidates  $K$  achieved the best trade-off between segmentation accuracy and computational overhead (Fig. 3(B)). Employing CAU, CEU, and CDU to evaluate uncertainty for querying target samples achieved superior performance than other uncertainty-based methods (Fig. 3(C)).

## 4 Conclusion

We proposed an Active Source-Free Open-Set Domain Adaptation method for medical image segmentation. Our method achieved superior performance on volumetric multi-organ segmentation than SOTA domain adaptation methods.

**Acknowledgments.** The authors have no acknowledgments to declare.

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

## References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. *Nature communications* **13**(1), 4128 (2022)
2. Bateson, M., Kervadec, H., Dolz, J., Lombaert, H., Ayed, I.B.: Source-free domain adaptation for image segmentation. *Medical Image Analysis* **82**, 102617 (2022)
3. Chen, C., Liu, Q., Jin, Y., Dou, Q., Heng, P.A.: Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 225–235. Springer (2021)
4. Gaillochet, M., Desrosiers, C., Lombaert, H.: Active learning for medical image segmentation with stochastic batches. *Medical Image Analysis* **90**, 102958 (2023)
5. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering* **69**(3), 1173–1185 (2021)
6. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: *International MICCAI brainlesion workshop*. pp. 272–284. Springer (2021)
7. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems* **35**, 36722–36732 (2022)
8. Li, M., Sethi, I.K.: Confidence-based active learning. *IEEE transactions on pattern analysis and machine intelligence* **28**(8), 1251–1261 (2006)

9. Liu, X., Araki, K., Harada, S., Yoshizawa, A., Terada, K., Kurata, M., Nakajima, N., Abe, H., Ushiku, T., Bise, R.: Cluster entropy: Active domain adaptation in pathological image segmentation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2023)
10. Luo, Z., Luo, X., Gao, Z., Wang, G.: An uncertainty-guided tiered self-training framework for active source-free domain adaptation in prostate segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 107–117. Springer (2024)
11. Ma, J., Zhang, Y., Gu, S., An, X., Wang, Z., Ge, C., Wang, C., Zhang, F., Wang, Y., Xu, Y., et al.: Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. *Medical Image Analysis* **82**, 102616 (2022)
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
13. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. In: International Conference on Learning Representations (2018), <https://openreview.net/forum?id=H1aIuk-RW>
14. VS, V., Valanarasu, J.M.J., Patel, V.M.: Target and task specific source-free domain adaptive image segmentation. In: Medical Imaging with Deep Learning (2024), <https://openreview.net/forum?id=Ym30gCtKqN>
15. Wang, D., Shang, Y.: A new active labeling method for deep learning. In: 2014 International joint conference on neural networks (IJCNN). pp. 112–119. IEEE (2014)
16. Wang, H., Chen, J., Zhang, S., He, Y., Xu, J., Wu, M., He, J., Liao, W., Luo, X.: Dual-reference source-free active domain adaptation for nasopharyngeal carcinoma tumor segmentation across multiple hospitals. *IEEE Transactions on Medical Imaging* (2024)
17. Wang, W., Wang, H., Ran, Z.Y., He, R.: Learning robust feature transformation for domain adaptation. *Pattern Recognition* **114**, 107870 (2021)
18. Wang, Y., Zhou, Y., Shen, W., Park, S., Fishman, E.K., Yuille, A.L.: Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. *Medical image analysis* **55**, 88–102 (2019)
19. Xie, B., Yuan, L., Li, S., Liu, C.H., Cheng, X., Wang, G.: Active learning for domain adaptation: An energy-based approach. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 8708–8716 (2022)
20. Yang, C., Guo, X., Chen, Z., Yuan, Y.: Source free domain adaptation for medical image segmentation with fourier style mining. *Medical Image Analysis* **79**, 102457 (2022)
21. Yang, J., Marcus, D.S., Sotiras, A.: Adapting medical vision foundation models for volumetric medical image segmentation via active learning and selective semi-supervised fine-tuning. arXiv preprint arXiv:2509.10784 (2025)
22. Yang, J., Qiu, P., Zhang, Y., Marcus, D.S., Sotiras, A.: D-net: Dynamic large kernel with dynamic feature fusion for volumetric medical image segmentation. *Biomedical Signal Processing and Control* **113**, 108837 (2026)
23. Yang, J., Yu, X., Qiu, P., Marcus, D., Sotiras, A.: Active source-free cross-domain and cross-modality adaptation for volumetric medical image segmentation by image sensitivity and organ heterogeneity sampling. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 3–12. Springer (2025)

24. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive annotation: A deep active learning framework for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 399–407. Springer (2017)
25. Yang, Z., Li, Y., Liu, Y.: Continual source-free active domain adaptation for nasopharyngeal carcinoma tumor segmentation across multiple hospitals. *Neural Networks* p. 107869 (2025)
26. Zhang, T., Li, K., Gu, S., Heng, P.A.: Enhancing source-free domain adaptation in medical image segmentation via regulated model self-training. *Medical Image Analysis* **102**, 103543 (2025)
27. Zhou, A., Liu, Z., Tieu, A., Patel, N., Sun, S., Yang, A., Choi, P., Lee, H.C., Tordjman, M., Deyer, L., et al.: Mrannotator: multi-anatomy and many-sequence mri segmentation of 44 structures. *Radiology Advances* **2**(1), uma035 (2025)
28. Zhou, T., Yang, J., Cui, L., Zhang, N., Chai, S.: Sbc-al: Structure and boundary consistency-based active learning for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 283–293. Springer (2024)