Active Source-Free Cross-Domain and Cross-Modality Adaptation for Volumetric Medical Image Segmentation by Image Sensitivity and Organ Heterogeneity Sampling

Jin Yang[1], Xiaobing Yu[1], Peijie Qiu[1], Daniel Marcus[1], and Aristeidis Sotiras[1,2]

[1] Mallinckrodt Institute of Radiology, Washington University School of Medicine in St. Louis, St. Louis, MO, USA, 63110
[2] Institute for Informatics, Data Science and Biostatistics, Washington University School of Medicine in St. Louis, St. Louis, MO, USA, 63110
yang.jin@wustl.edu

Abstract. Deep learning (DL) methods have achieved great success in medical image segmentation, but they are challenged to demonstrate robust performance across different datasets due to domain and modality gaps. The Source-Free Domain Adaptation techniques adapt DL models to generalize across domains without access to source data, and active learning is implemented to actively query informative target samples to fine-tune models, thus improving their generalization. However, only a few Active Source-Free Domain Adaptation methods have been proposed. Additionally, existing methods focus on same-modality adaptation and lack mechanisms to address modality gaps, thus limiting their applicability. To address these limitations, we propose a novel Active Source-Free Cross-Domain and Cross-Modality Adaptation method for medical image segmentation. This method adapts models across different domains and modalities by employing a novel Active Test Time Sample Query strategy to jointly implement Image Sensitivity Query (ISQ) and Organ Heterogeneity Query (OHQ). ISQ is designed to evaluate samples' image-level modality agnostic informativeness, thus querying informative samples from different domains and modalities. OHQ is proposed to query samples with large foreground diversity by measuring the uncertainty-weighted organ boundary discontinuity and uncertainty-weighted organ interior abnormality, thus avoiding the influence of modality-specific background noise. A Dynamic Image-to-Organ Scaling mechanism is proposed to dynamically fuse the results of ISQ and OHQ for sample querying. We evaluated our method on cross-domain and cross-modality volumetric pancreas segmentation tasks. Our method outperformed other state-of-the-art methods on adaptation from a CT domain to another larger CT domain, T1-weighted MR and T2-weighted MR domains.

Keywords: Active Source-Free Domain Adaptation · Active Learning · Medical Image Segmentation · Data Efficient Learning

1 Introduction

Various deep learning (DL) based methods have achieved great success in automatic organ segmentation [10]. However, they are limited from generalizing across different data sources due to domain gaps [18]. To address this issue, Domain Adaptation methods have been proposed to enhance DL methods' generalizability to the unlabeled target domain by leveraging knowledge from labeled source domain data [3]. However, strict data privacy regulations in clinical practice may lead to the source data annotations' inaccessibility and source knowledge's unavailability [17]. Thus, Source-free Domain Adaptation (SFDA) techniques were proposed to adapt models to target domains without access to the source data [8,21,20]. Moreover, utilizing a small set of data with annotations to train models in the target domain improves their adaptation performance [6]. Active learning (AL) is implemented to determine the optimal way to select samples for annotations, and these AL-based methods employ query strategies to choose the most informative samples for training, thus enabling models to achieve optimal performance with minimal annotation cost [4,5,11,24]. Therefore, Active Source-Free Domain Adaptation (ASFDA) methods are designed to adapt models to the target domain by actively querying informative samples without knowledge of the source domain. However, only a few ASFDA methods have been proposed for medical image segmentation [18,9,19]. Additionally, these methods investigated the adaptation of segmentation networks across domains with the same modality by exploring predictive uncertainty and feature similarity for querying. Thus, they lack mechanisms to utilize modality-agnostic information or organ-specific characteristics to query samples from different domains and modalities.

To tackle these limitations, we propose a novel **Active Source-Free Cross-Domain and Cross-Modality Adaptation** method for medical image segmentation. This method adapts segmentation models to generalize across different domains and modalities by employing a novel Active Test Time Sample Query strategy. This strategy queries target samples to fine-tune models by jointly implementing **Image Sensitivity Query** (ISQ) and **Organ Heterogeneity Query** (OHQ). ISQ is implemented to evaluate image-level modality-agnostic informativeness of samples. The information level of each sample is measured by calculating the KL divergence between its probabilistic predictions and that of its modality-agnostic perturbations. OHQ is implemented to query samples with large foreground diversity by measuring **uncertainty-weighted organ boundary discontinuity** (OBD) and **uncertainty-weighted organ interior abnormality** (OIA). OBD is proposed to measure the complexity of boundary regions of the target organ, and OIA can measure the variability of the interior regions of the target organ. Lastly, we propose a **Dynamic Image-to-Organ Scaling** mechanism to dynamically fuse results of ISQ and OHQ to score samples for querying. This mechanism is designed to prioritize image-level informativeness for querying at the start of adaptation and gradually shift focus on organ-level informativeness by reducing relative weighting between ISQ and OHQ. We evaluated the effectiveness of our method on cross-domain and

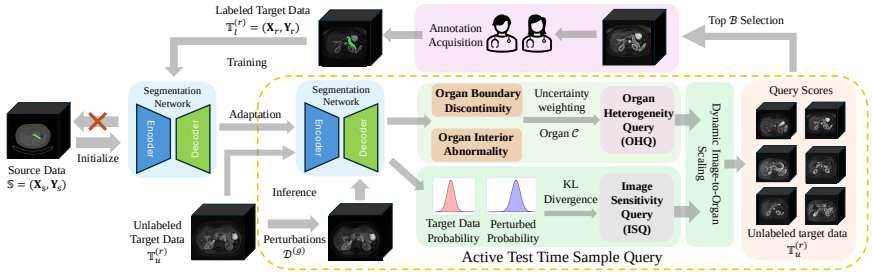

Fig. 1. Active Source-Free Cross-Domain and Cross-Modality Adaptation for medical image segmentation. The segmentation network was trained on the source domain $\mathbb{S}$ and adapted to the target domain $\mathbb{T}$. Image Sensitivity Query (ISQ) was evaluated by calculating the KL divergence between probabilistic predictions of unlabeled target data $\mathbb{T}_u^{(r)}$ and their perturbations $\mathcal{D}^{(g)}$. The Organ Heterogeneity Query (OHQ) was evaluated by measuring uncertainty-weighted Organ Boundary Discontinuity and uncertainty-weighted Organ Interior Abnormality. Subsequently, the results of ISQ and OHQ were fused via Dynamic Image-to-Organ Scaling to generate query scores for unlabeled data $\mathbb{T}_u^{(r)}$. Finally, top $\mathcal{B}$ samples were queried for annotations, and these labeled target data $\mathbb{T}_l^{(r)}$ were utilized to re-train the network.

cross-modality 3D volumetric pancreas segmentation by adapting networks from a CT source domain to another larger CT target domain, an MR T1w target domain, and an MR T2w target domain. Our method achieved superior adaptation performance compared to other commonly used ASFDA methods.

2 Methods

2.1 Overall Method

Given a source domain $\mathbb{S}$ with $\mathcal{N}_s$ labeled data of a specific modality $\mathcal{M}_s$ and a target domain $\mathbb{T}$ with $\mathcal{N}_t$ unlabeled data of another modality $\mathcal{M}_t$, the goal of the Active Source-Free Cross-Domain and Cross-Modality Adaptation is to derive a segmentation model with source prior knowledge to minimize the risk on the target domain without access to the source domain (Fig. 1). During adaptation, an Active Test Time Sample Query strategy is implemented to query a small number of samples from the target domain for annotations to boost the network's performance on this domain and modality. Specifically, images $\boldsymbol{X}_s = \{x_s | 1 \leq s \leq \mathcal{N}_s; \mathcal{M}_s\}$ and their labels $\boldsymbol{Y}_s = \{y_s | 1 \leq s \leq \mathcal{N}_s\}$ from the source domain $\mathbb{S} = \{(\boldsymbol{X}_s, \boldsymbol{Y}_s)\}$ are utilized to train the segmentation network $\mathscr{F}(\boldsymbol{\Theta}; (\boldsymbol{X}, \boldsymbol{Y}))$. Subsequently, we adapt this source-trained network with prior knowledge $\mathscr{F}(\boldsymbol{\Theta}; (\boldsymbol{X}_s, \boldsymbol{Y}_s))$ to the target domain without access to source data. To maximize performance and minimize risks during this adaptation, an AL procedure is implemented to query a small number of $\mathcal{N}_{AL}$ informative samples for annotations during test time ($\mathcal{N}_{AL} << \mathcal{N}_t$). The budget of annotated target

samples at each round is pre-defined as $\mathcal{B}$. At the beginning of the $\mathcal{R}$-round query-and-adaptation alternation, the network is source-trained as $\mathscr{F}^{(0)}(\boldsymbol{\Theta};(\boldsymbol{X}_0,\boldsymbol{Y}_0))$ from $\mathscr{F}(\boldsymbol{\Theta};(\boldsymbol{X}_s,\boldsymbol{Y}_s))$. At the first round $r=1$, we actively query $\mathcal{B}$ instances for annotations from the unlabeled target domain $\mathbb{T}_u^{(0)}$ based on the Active Test Time Sample Query strategy by implementing Image Sensitivity Query and Organ Heterogeneity Query. These samples $\boldsymbol{X}_1=\{x_1,...,x_{\mathcal{B}}\}$ with their annotations $\boldsymbol{Y}_1=\{y_1,...,y_{\mathcal{B}}\}$ are collected to generate a labeled target domain $\mathbb{T}_l^{(1)}$ and utilized to optimize the model $\mathscr{F}^{(0)}(\boldsymbol{\Theta};(\boldsymbol{X}_0,\boldsymbol{Y}_0))$ to $\mathscr{F}^{(1)}(\boldsymbol{\Theta};(\boldsymbol{X}_1,\boldsymbol{Y}_1))$. The target domain is split by $\mathbb{T}_u^{(1)}=\mathbb{T}_u^{(0)}\backslash\mathbb{T}_l^{(1)}$. In subsequent rounds ($r>1$), $\mathcal{B}$ samples $\boldsymbol{X}_r=\{x_1,...,x_{\mathcal{B}}\}$ and their annotations $\boldsymbol{Y}_r=\{y_1,...,y_{\mathcal{B}}\}$ are queried to generate $\mathbb{T}_l^{(r)}=\mathbb{T}_l^{(r-1)}\cup\{(\boldsymbol{X}_r,\boldsymbol{Y}_r);\mathcal{M}_t\}$ to continuously optimize the network as $\mathscr{F}^{(r)}(\boldsymbol{\Theta};(\boldsymbol{X}_r,\boldsymbol{Y}_r))$, and the unlabeled set is updated by $\mathbb{T}_u^{(r)}=\mathbb{T}_u^{(r-1)}\backslash\mathbb{T}_l^{(r)}$. The iterations will be terminated until the labeled target samples reach the pre-defined annotation budgets $\mathcal{N}_{AL}$.

2.2 Active Test Time Sample Query Strategy

Image Sensitivity Query. We implement an ISQ to query the most informative samples by evaluating their image-level modality-agnostic informativeness (Fig. 1). Their informativeness is measured based on the model's epistemic uncertainty. Epistemic uncertainty arises from a model's lack of knowledge and its limitations in learning from data and generalizing to new situations. When a model is less confident in its predictions for a sample, that sample is likely to lie away from the distribution of learned knowledge, indicating it contains more unlearned information. Such samples are thus more informative and valuable for model training. ISQ is designed to quantify this uncertainty by measuring the difference between the model's original and perturbed probabilistic predictions. Thus, querying high ISQ samples targets those that contribute most to the model's epistemic uncertainty and limit its generalization to the target domain.

For an image $\mathcal{I}\in\mathbb{R}^{D\times H\times W}$, its modality-agnostic perturbations can be constructed by applying image texture transformation techniques $\mathcal{D}^{(g)}(\mathcal{I})$ to re-sample it ($g\in\{1,2,...,\mathcal{G}\}$). If a sample is more informative to the network, its perturbations will be more likely to be mispredicted. Thus, the informativeness of a sample can be described by the differences between its prediction results and the predictions of its perturbations. In our method, we applied four perturbation techniques ($\mathcal{G}=4$), including adding zero-centered additive Gaussian noise with $U(0,0.1)$ variance, Gaussian blurring by a $U(0.5,1.5)$ kernel, brightness enhancement, and contrast enhancement. The probabilistic predictions of this image and its modality-agnostic perturbed predictions are generated from the segmentation network as $\mathscr{F}(\boldsymbol{\Theta};\mathcal{I})$ and $\mathscr{F}(\boldsymbol{\Theta};\mathcal{D}^{(g)}(\mathcal{I}))$. The score of ISQ($\mathcal{I}$) is evaluated by calculating Kullback-Leibler (KL) divergence to measure the distance between the probabilistic output of the original sample and that of

modality-agnostic perturbations

$$\begin{aligned}\mathrm{ISQ}(\mathcal{I}) &= \sum_{g}^{\mathcal{G}} \mathrm{KL}\big[\mathscr{F}(\boldsymbol{\Theta};\mathcal{I})||\mathscr{F}(\boldsymbol{\Theta};\mathcal{D}^{(g)}(\mathcal{I}))\big] \\ &= \sum_{g=0}^{4} \mathscr{F}(\boldsymbol{\Theta};\mathcal{I}) \log \frac{\mathscr{F}(\boldsymbol{\Theta};\mathcal{I})}{\mathscr{F}(\boldsymbol{\Theta};\mathcal{D}^{(g)}(\mathcal{I}))}. \end{aligned} \tag{1}$$

Organ Heterogeneity Query. Calculating image-level similarities among samples may misestimate their diversity since these methods may query samples with a large diversity in the background while ignoring the heterogeneity of organs, limiting networks from capturing organ-related information. Thus, we implement an OHQ to query samples based on the diversity of foreground using two metrics: **uncertainty-weighted organ boundary discontinuity** and **uncertainty-weighted organ interior abnormality** (Fig. 1).

Since most mis-segmented or over-segmented regions are within the boundaries of organs, OBD is most likely to indicate the complexity of boundary regions and the difficulty of segmentation. Thus, querying samples with the largest organ boundary discontinuity will select samples with the highest complexity of boundary regions, thus significantly benefiting the network from learning diverse foreground representations. Specifically, this OBD is evaluated by averaging the L1-norm difference between the intensity values of all boundary voxels belonging to this organ and its neighbors. A specific organ $\mathcal{C}$ has $\mathcal{N}$ boundary voxel points where i-th voxel has an intensity value $\mathcal{U}_i$, and each boundary voxel has $\mathcal{K}$ neighbor voxels (from foreground and background) with the intensity value $\mathcal{U}_j$. The number of voxels for a specific organ varies from sample to sample, and to avoid its influence, the $\mathrm{OBD}^{(\mathcal{C})}$ is normalized by the number of boundary voxels and their neighbor voxels. Subsequently, boundary regions in predictions present a large uncertainty. Thus, after predictive entropy is normalized by the number of boundary voxels $\mathcal{N}$ and the total number of neighbor voxels from the background $\mathcal{H}$, weighting $\mathrm{OBD}^{(\mathcal{C})}$ with this voxel-wise normalized entropy facilitates the assessment of potential false predictions

$$\mathrm{OBD}^{(\mathcal{C})} = \bigg(-\frac{1}{\mathcal{N}+\mathcal{H}} \sum_{i=1}^{\mathcal{N}+\mathcal{H}} p_i \log p_i \bigg) \bigg(\frac{1}{\mathcal{N}\times\mathcal{K}} \sum_{i=1}^{\mathcal{N}} \sum_{j=1}^{\mathcal{K}} ||\mathcal{U}_i^{(\mathcal{C})} - \mathcal{U}_j^{(\mathcal{C})}||_1 \bigg). \tag{2}$$

The organ from a specific sample may present different image intensity values from others if it demonstrates large textural heterogeneity in the interior regions. Additionally, the organ with large heterogeneity in the interior region may show variations in image intensity values due to anatomical abnormality, such as masses (cysts or tumors). Thus, evaluating OIA enables querying samples whose organs show large heterogeneity in interior regions. Specifically, after boundary voxels are excluded, the number of interior voxels of a specific organ $\mathcal{C}$ are counted as $\mathcal{L}$ with i-th voxel of the intensity value $\mathcal{V}_i$. OIA is evaluated by

calculating the variations in intensity values of interior voxels. OIA is normalized by the number of interior voxels and weighted by the voxel-wise normalized predictive entropy to assess potential false predictions in large structures

$$\mathrm{OIA}^{(\mathcal{C})} = \Big(-\frac{1}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} p_i \log p_i \Big) \Big(\frac{1}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} ||\mathcal{V}_i - \frac{\sum_{j=1}^{\mathcal{L}} \mathcal{V}_j}{\mathcal{L}}||_1 \Big). \tag{3}$$

The OHQ is calculated by adding the scores of OBD and OIA as

$$\mathrm{OHQ}(\mathcal{I}) = \mathrm{OBD}^{(\mathcal{C})} + \mathrm{OIA}^{(\mathcal{C})}. \tag{4}$$

Dynamic Image-to-Organ Scaling. We propose a Dynamic Image-to-Organ Scaling mechanism to fuse $\mathrm{ISQ}(\mathcal{I})$ and $\mathrm{OHQ}(\mathcal{I})$ dynamically by changing their relative weights during AL iterations r $(1 \leq r \leq \mathcal{R})$. Firstly, we apply the Max-Min normalization (*Norm*) to ensure the scores of $\mathrm{ISQ}(\mathcal{I})$ and $\mathrm{OHQ}(\mathcal{I})$ are in the same range. Subsequently, we assign a larger weight to $\mathrm{ISQ}(\mathcal{I})$ at the first several rounds to query samples with more image-level modality agnostic information, thus helping the network quickly capture domain-specific and modality-specific knowledge. Subsequently, we gradually increase the weight for $\mathrm{OHQ}(\mathcal{I})$ while reducing that of $\mathrm{ISQ}(\mathcal{I})$. This is designed to allow the network to focus on the diversity of the target organ within the specific domain

$$\mathrm{Q}(\mathcal{I}) = \Big(1 - \frac{r}{\mathcal{R}+1}\Big) \Big[\mathit{Norm}\big(\mathrm{ISQ}(\mathcal{I})\big) \Big] + \Big(\frac{r}{\mathcal{R}+1}\Big) \Big[\mathit{Norm}\big(\mathrm{OHQ}(\mathcal{I})\big) \Big]. \tag{5}$$

3 Experiments

Datasets. We utilized four public pancreas segmentation datasets. The National Institutes of Health (NIH) dataset comprises 80 contrast-enhanced 3D abdominal CT scans with the annotations of normal pancreas [13]. The Medical Segmentation Decathlon (MSD) dataset consists of 281 abdominal CT scans with annotations of pancreatic parenchyma and pancreatic mass [1]. The third and fourth datasets consist of 162 T1-weighted (T1w) and T2-weighted (T2w) abdominal MRI series with annotations of the pancreas from PanSegData [23].

Implementation details. The experiments were implemented using PyTorch on NVIDIA Tesla A100 PCI-E Passive Single GPU with 40GB of GDDR5 memory. The 3D U-Net and Attention U-Net were used as segmentation networks [12,14]. The combination of Dice loss and Cross-entropy loss was used as the loss function. We trained the network on source data for 1000 epochs and fine-tuned it on target data for 500 epochs in each iteration. The batch size was 2. The SGD was utilized for optimization. The initial learning rate was set to 0.01 and decayed in a polynomial scheduler with a power of 0.9. Raw volumes were z-score normalized and scaled to the patches with the dimension of $64 \times 224 \times 224$. The

Table 1. Performance comparison between ours and other ASFDA methods when adapting U-Net and Attention U-Net trained on NIH source data to MSD target data (NIH → MSD). The results were reported as Mean±SD. **Bold** and underline represent the best and the second best result. (Lower bound 0%: source-training without ADA; Upper bound 100%: fully supervised training)

Nets	Mtc	Methods	Low Bound	Query Budgets (Number of Iterations)					Upper Bound
			0%	4% (r=1)	8% (r=2)	12% (r=3)	16% (r=4)	20% (r=5)	100%
U-Net	DSC (↑)	RAND	42.30±16.48	72.62±18.21	79.53±11.29	81.58±10.24	83.10±9.79	84.00±8.57	92.48±1.64
		SENT	42.30±16.48	73.59±18.00	80.98±12.48	81.91±9.53	83.14±9.05	84.03±8.84	92.48±1.64
		LC	42.30±16.48	67.36±24.60	78.23±15.42	80.29±13.53	82.84±12.59	83.71±11.97	92.48±1.64
		SMAR	42.30±16.48	67.53±24.10	78.68±14.14	80.69±13.38	82.86±12.92	83.73±10.64	92.48±1.64
		Core-set	42.30±16.48	71.62±17.35	79.48±11.22	81.36±11.79	83.08±9.97	83.90±9.87	92.48±1.64
		MREP	42.30±16.48	69.89±25.60	79.10±17.43	80.73±12.30	82.54±11.29	83.39±10.12	92.48±1.64
		BADGE	42.30±16.48	74.60±18.86	81.49±12.93	82.70±8.12	84.56±8.79	84.84±8.87	92.48±1.64
		Ours	42.30±16.48	**76.33**±15.28	**82.01**±10.48	**83.28**±7.54	**85.02**±9.11	**86.12**±7.90	92.48±1.64
	95HD (↓)	RAND	25.06±12.85	15.91±22.64	9.00±12.57	7.62±11.24	6.41±9.35	5.74±7.18	1.69±1.44
		SENT	25.06±12.85	15.54±21.74	8.94±12.72	7.65±10.14	6.28±9.27	5.72±7.99	1.69±1.44
		LC	25.06±12.85	19.28±29.77	10.85±15.20	8.98±14.69	7.09±12.85	6.26±10.68	1.69±1.44
		SMAR	25.06±12.85	19.45±29.27	9.68±15.07	8.85±13.98	6.93±12.53	6.23±10.20	1.69±1.44
		Core-set	25.06±12.85	17.30±21.36	9.37±10.21	7.25±11.10	6.69±9.84	6.20±9.45	1.69±1.44
		MREP	25.06±12.85	18.36±26.35	9.44±15.54	7.61±11.44	7.30±10.16	6.73±9.88	1.69±1.44
		BADGE	25.06±12.85	14.13±20.57	8.13±12.80	6.98±8.37	5.61±7.65	5.43±8.29	1.69±1.44
		Ours	25.06±12.85	**11.40**±13.15	**7.36**±10.08	**6.10**±8.02	**5.17**±7.23	**4.68**±6.47	1.69±1.44
Attention U-Net	DSC (↑)	RAND	42.10±16.40	66.10±21.24	76.87±17.21	79.45±12.44	82.47±11.66	82.59±10.46	92.00±1.91
		SENT	42.10±16.40	68.39±18.99	78.78±12.42	81.15±10.69	83.18±10.30	83.92±9.05	92.00±1.91
		LC	42.10±16.40	65.01±24.99	77.14±15.54	80.31±12.49	81.57±11.62	82.69±10.72	92.00±1.91
		SMAR	42.10±16.40	65.73±24.26	77.57±15.09	80.71±12.45	81.79±11.07	83.28±10.33	92.00±1.91
		Core-set	42.10±16.40	71.72±20.09	79.29±12.76	81.45±10.99	82.91±9.32	83.88±9.38	92.00±1.91
		MREP	42.10±16.40	62.31±26.05	74.12±17.12	78.16±13.23	80.52±12.18	81.83±12.00	92.00±1.91
		BADGE	42.10±16.40	74.29±16.77	80.81±11.15	82.33±9.72	83.66±9.73	84.54±8.89	92.00±1.91
		Ours	42.10±16.40	**75.99**±16.49	**81.32**±10.09	**83.35**±8.39	**84.93**±8.36	**85.79**±8.10	92.00±1.91
	95HD (↓)	RAND	25.35±16.29	20.50±21.02	12.64±17.94	9.54±10.09	7.36±10.24	7.48±9.87	1.76±1.50
		SENT	25.35±16.29	18.10±20.47	10.66±12.48	8.48±9.85	6.86±9.76	6.15±8.66	1.76±1.50
		LC	25.35±16.29	21.55±21.97	12.05±15.64	9.98±10.31	8.18±10.22	7.51±9.18	1.76±1.50
		SMAR	25.35±16.29	21.27±21.17	11.29±15.77	9.60±10.08	8.13±10.02	7.21±9.92	1.76±1.50
		Core-set	25.35±16.29	16.10±19.86	9.62±12.56	7.65±9.89	7.05±9.47	6.36±9.05	1.76±1.50
		MREP	25.35±16.29	22.51±28.35	13.75±18.54	10.05±11.26	9.13±11.22	7.86±10.09	1.76±1.50
		BADGE	25.35±16.29	14.81±21.10	8.92±10.86	7.38±9.27	6.48±9.87	5.95±8.17	1.76±1.50
		Ours	25.35±16.29	**12.89**±18.72	**7.55**±9.78	**6.42**±8.49	**6.15**±11.85	**5.42**±6.69	1.76±1.50

segmentation performance was evaluated using two metrics (Mtc): Dice Similarity Coefficient (DSC;%) and 95th percentile Hausdorff Distance (95HD;mm). We set $\mathcal{B} = 4\%$ and $\mathcal{R} = 5$ in the NIH-to-MSD adaptation, and $\mathcal{B} = 5\%$ and $\mathcal{R} = 5$ in the CT-to-T1w and CT-to-T2w adaptation. To ensure the reliability of the experimental results, we performed the experiments three times and recorded the average value for each setting.

Experimental results. We compared the performance of our method with seven well-known active query methods, including Random Selection (**RAND**), Softmax Entropy (**SENT**) [16], Least Confidence (**LC**) [7], Softmax Margin (**SMAR**) [16], core-set approach (**Core-set**) [15], Maximum Representation (**MREP**) [22], and Batch Active Diverse Gradient Embeddings (**BADGE**) [2]. Our method outperformed other methods on NIH-to-MSD adaptation (Table 1), CT-to-T1w, and CT-to-T2w adaptations (Table 2 and Fig. 2). Specifically, when U-Net and Attention U-Net were adapted from the NIH CT domain to the MSD

Table 2. Performance comparison between ours and other ASFDA methods when adapting U-Net trained on NIH CT source data to PanSegData T1w (CT → T1w) and T2w (CT → T2w) MR target data. The results were reported as Mean±SD. **Bold** and underline represent the best and the second best result. (Lower bound 0%: source-training without ADA; Upper bound 100%: fully supervised training)

Tasks	Mtc	Methods	Low Bound	Query Budgets (Number of Iterations)					Upper Bound
			0%	5% (r=1)	10% (r=2)	15% (r=3)	20% (r=4)	25% (r=5)	100%
CT→ MR T1w	DSC (↑)	RAND	3.83±10.21	60.45±19.64	69.58±19.84	72.96±19.10	76.11±19.32	78.78±18.49	95.17±1.17
		SENT	3.83±10.21	61.49±19.96	70.23±19.53	74.38±18.80	76.32±18.80	79.02±18.73	95.17±1.17
		LC	3.83±10.21	57.54±22.12	67.26±21.14	72.14±22.20	74.38±21.50	77.93±19.67	95.17±1.17
		SMAR	3.83±10.21	58.97±21.78	67.71±20.50	72.70±20.68	75.42±19.60	78.41±18.29	95.17±1.17
		Core-set	3.83±10.21	63.53±18.55	73.09±17.82	77.18±17.35	79.32±16.23	81.15±15.32	95.17±1.17
		MREP	3.83±10.21	57.03±27.91	68.92±20.94	73.70±19.65	76.28±19.94	78.52±18.89	95.17±1.17
		BADGE	3.83±10.21	62.46±19.31	70.25±19.52	74.79±19.51	76.89±18.02	79.36±17.72	95.17±1.17
		Ours	3.83±10.21	**65.07**±17.76	**74.95**±16.29	**78.38**±15.66	**81.01**±15.59	**83.09**±15.07	95.17±1.17
	95HD (↓)	RAND	77.82±28.97	15.89±15.45	9.25±13.66	7.54±9.19	6.80±9.71	5.32±8.62	1.00±0.00
		SENT	77.82±28.97	14.64±16.12	9.03±10.79	6.96±10.33	5.94±10.55	5.28±10.08	1.00±0.00
		LC	77.82±28.97	16.06±17.12	11.05±15.57	8.64±16.33	6.96±12.12	6.30±11.90	1.00±0.00
		SMAR	77.82±28.97	16.30±18.67	10.33±14.49	8.11±16.07	6.86±11.29	6.26±10.73	1.00±0.00
		Core-set	77.82±28.97	12.73±13.95	8.26±10.13	5.64±6.80	5.52±6.96	4.73±7.63	1.00±0.00
		MREP	77.82±28.97	16.92±21.79	9.33±13.75	7.01±9.23	6.36±9.33	6.21±8.82	1.00±0.00
		BADGE	77.82±28.97	13.18±14.59	8.69±10.09	6.66±8.55	5.77±8.80	5.11±8.32	1.00±0.00
		Ours	77.82±28.97	**11.62**±12.48	**6.76**±9.81	**5.61**±5.62	**4.80**±7.29	**4.01**±7.03	1.00±0.00
CT→ MR T2w	DSC (↑)	RAND	2.30±7.19	43.59±22.82	60.17±27.45	68.48±19.02	73.16±17.45	76.28±15.86	95.15±1.37
		SENT	2.30±7.19	46.32±22.01	63.49±18.30	69.47±20.48	73.83±16.91	77.45±15.32	95.15±1.37
		LC	2.30±7.19	41.14±27.95	59.11±20.92	67.47±18.53	72.58±16.76	75.88±16.70	95.15±1.37
		SMAR	2.30±7.19	41.82±23.48	59.61±19.57	67.68±17.32	73.00±16.13	75.95±15.16	95.15±1.37
		Core-set	2.30±7.19	53.53±19.95	66.69±17.39	71.33±16.55	75.23±15.75	78.43±14.14	95.15±1.37
		MREP	2.30±7.19	44.39±23.58	61.14±19.47	68.68±18.07	73.64±19.38	76.45±15.04	95.15±1.37
		BADGE	2.30±7.19	50.23±20.68	64.65±19.35	71.24±17.48	74.95±15.62	77.69±14.65	95.15±1.37
		Ours	2.30±7.19	**58.41**±19.75	**69.62**±16.54	**76.57**±14.62	**78.62**±14.89	**80.13**±14.18	95.15±1.37
	95HD (↓)	RAND	65.80±49.65	24.27±21.58	14.57±20.57	9.97±11.58	8.42±11.25	6.70±8.77	1.00±0.08
		SENT	65.80±49.65	20.07±20.22	12.65±14.01	9.38±14.51	7.44±10.93	5.83±8.62	1.00±0.08
		LC	65.80±49.65	25.62±27.06	15.24±15.74	11.27±12.59	9.87±10.90	7.21±9.49	1.00±0.08
		SMAR	65.80±49.65	25.04±21.28	14.78±15.89	10.69±11.74	9.86±10.59	6.76±9.07	1.00±0.08
		Core-set	65.80±49.65	16.79±15.38	10.45±13.33	7.67±10.46	5.92±9.97	5.10±8.20	1.00±0.08
		MREP	65.80±49.65	23.19±24.26	13.33±15.38	9.67±13.60	8.10±10.13	6.56±8.69	1.00±0.08
		BADGE	65.80±49.65	17.11±17.73	12.64±16.49	8.60±11.14	6.80±9.82	5.33±8.66	1.00±0.08
		Ours	65.80±49.65	**15.55**±14.13	**8.73**±12.16	**6.74**±11.30	**4.82**±9.78	**4.70**±8.22	1.00±0.08

CT domain, our method achieved better performance than other ASFDA methods at each round. Additionally, our method adapted the U-Net and Attention U-Net to achieve over 93.12% and 93.25% upper bound performance with only 20% samples queried. Moreover, when U-Net was adapted from the NIH CT domain to the T1w MR and the T2w MR domains, our method demonstrated superior performance than other methods at each round. Additionally, U-Net achieved 87.31% and 84.21% upper bound performance with only 25% samples queried for cross-modality adaptation by our method.

Ablation study. To evaluate the effectiveness of the Active Test Time Sample Query strategy, we compared the performance of different strategies on the CT-to-T1w when querying 25% samples (Table 3). Specifically, our method achieved the best performance, demonstrating the effectiveness of our query strategy.

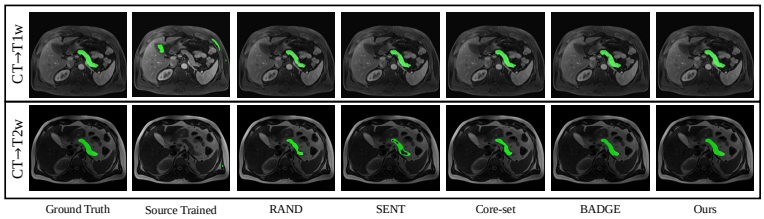

Fig. 2. Qualitative comparison among results of source-trained network before adaptation and networks fine-tuned by 25% samples queried by other and our methods.

Table 3. The results of ablation study on the Active Test Time Query strategy.

Strategies	ISQ	OHQ	ISQ+OHQ	ISQ+OHQ+Scaling
DSC (↑)	80.36±17.02	78.85±18.32	82.58±15.83	**83.09**±15.07
95HD (↓)	4.95±8.21	5.21±8.86	4.24±7.68	**4.01**±7.53

4 Conclusion

We proposed an Active Source-Free Cross-domain and Cross-modality Adaptation method for medical image segmentation by employing an Active Test Time Query strategy. Our method achieved superior performance on cross-domain and cross-modality volumetric pancreas segmentation compared to other methods.

Acknowledgment. Computations were performed on the Washington University RCIF funded by NIH S10 1S10OD025200-01A1 and 1S10OD030477-01.

Disclosure of Interests. The authors have no competing interests.

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. Ash, J.T., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A.: Deep batch active learning by diverse, uncertain gradient lower bounds. In: International Conference on Learning Representations (2020)
3. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering **69**(3), 1173–1185 (2021)
4. Hiasa, Y., Otake, Y., Takao, M., Ogawa, T., Sugano, N., Sato, Y.: Automated muscle segmentation from clinical ct using bayesian u-net for personalized musculoskeletal modeling. IEEE transactions on medical imaging **39**(4), 1030–1040 (2019)
5. Kadir, M.A., Alam, H.M.T., Sonntag, D.: Edgeal: an edge estimation based active learning approach for oct segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 79–89. Springer (2023)

6. Li, J., Yu, Z., Du, Z., Zhu, L., Shen, H.T.: A comprehensive survey on source-free domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
7. Li, M., Sethi, I.K.: Confidence-based active learning. IEEE transactions on pattern analysis and machine intelligence **28**(8), 1251–1261 (2006)
8. Liang, J., Hu, D., Wang, Y., He, R., Feng, J.: Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 8602–8617 (2021)
9. 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)
10. Mishra, S., Zhang, Y., Chen, D.Z., Hu, X.S.: Data-driven deep supervision for medical image segmentation. IEEE Transactions on Medical Imaging **41**(6), 1560–1574 (2022)
11. Qu, C., Zhang, T., Qiao, H., Tang, Y., Yuille, A.L., Zhou, Z., et al.: Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks. Advances in Neural Information Processing Systems **36** (2024)
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. Roth, H.R., Farag, A., Turkbey, E., Lu, L., Liu, J., Summers, R.M.: Data from pancreas-ct. the cancer imaging archive. IEEE Transactions on Image Processing **5** (2016)
14. Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D.: Attention gated networks: Learning to leverage salient regions in medical images. Medical image analysis **53**, 197–207 (2019)
15. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 (2017)
16. 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)
17. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)
18. 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)
19. Wang, H., Luo, X., Chen, W., Tang, Q., Xin, M., Wang, Q., Zhu, L.: Advancing uwf-slo vessel segmentation with source-free active domain adaptation and a novel multi-center dataset. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 75–85. Springer (2024)
20. Wang, Y., Liang, J., Zhang, Z.: A curriculum-style self-training approach for source-free semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
21. 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)

22. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive annotation: A deep active learning framework for biomedical image segmentation. In: Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20. pp. 399–407. Springer (2017)
23. Zhang, Z., Keles, E., Durak, G., Taktak, Y., Susladkar, O., Gorade, V., Jha, D., Ormeci, A.C., Medetalibeyoglu, A., Yao, L., et al.: Large-scale multi-center ct and mri segmentation of pancreas with deep learning. Medical Image Analysis **99**, 103382 (2025)
24. 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)