Dual-Reference Source-Free Active Domain Adaptation for Nasopharyngeal Carcinoma Tumor Segmentation across Multiple Hospitals

Hongqiu Wang, Jian Chen, Shichen Zhang, Yuan He, Jinfeng Xu, Mengwan Wu, Jinlan He, Wenjun Liao, Xiangde Luo

***Abstract*— Nasopharyngeal carcinoma (NPC) is a prevalent and clinically significant malignancy that predominantly impacts the head and neck area. Precise delineation of the Gross Tumor Volume (GTV) plays a pivotal role in ensuring effective radiotherapy for NPC. Despite recent methods that have achieved promising results on GTV segmentation, they are still limited by lacking carefully-annotated data and hard-to-access data from multiple hospitals in clinical practice. Although some unsupervised domain adaptation (UDA) has been proposed to alleviate this problem, unconditionally mapping the distribution distorts the underlying structural information, leading to inferior performance. To address this challenge, we devise a novel Sourece-Free Active Domain Adaptation framework to facilitate domain adaptation for the GTV segmentation task. Specifically, we design a dual reference strategy to select domain-invariant and domain-specific representative samples from a specific target domain for annotation and model fine-tuning without relying on source-domain data. Our approach not only ensures data privacy but also reduces the workload for oncologists as it just requires annotating a few representative samples from the target domain and does not need to access the source data. We collect a large-scale clinical dataset comprising 1057 NPC patients from five hospitals to validate our approach. Experimental results show that our method outperforms the previous active learning (e.g., AADA and MHPL) and UDA (e.g., Tent and CPR) methods, and achieves comparable results to the fully supervised upper bound, even with few annotations, highlighting the significant medical utility of our approach. In addition, there is no public dataset about multi-center NPC segmentation, we will release code and dataset for future research (*Git*).**

This work was supported by the National Natural Science Foundation of China under Grant 82203197. Corresponding author: Xiangde Luo (xiangde.luo@std.uestc.edu.cn).

H. Wang and S. Zhang are with the Department of Systems Hub, Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511400, China.

J. Chen is with the Department of Radiology, University of Cambridge, Cambridge CB2 1TN, UK.

Y. He is with the Department of Radiation Oncology, Anhui Provincial Hospital, University of Science and Technology of China, Hefei 23000, China.

J. Xu is with the Department of Radiation Oncology, Nanfang Hospital, Southern Medical University, Guangzhou 510515, China.

M. Wu is with the Cancer Center, Sichuan Provincial People's Hospital, Chengdu 610041, China.

J. He is with the Department of Radiation Oncology, West China Hospital, Sichuan University, Chengdu 610041, China.

W. Liao and X. Luo are with the Department of Radiation Oncology, Sichuan Cancer Hospital and Institute, University of Electronic Science and Technology of China, Chengdu 610072, China. X. Luo is also with Shanghai Artificial Intelligence Laboratory, Shanghai 200030, China.

***Index Terms*— GTV Segmentation, active learning, domain adaptation, nasopharyngeal carcinoma.**

I. Introduction

NASOPHARYNGEAL carcinoma (NPC) is a prevalent malignancy affecting the head and neck region, and Intensity-Modulated Radiation Therapy (IMRT) has emerged as a preferred radiation technique for its treatment [1]–[3]. IMRT has demonstrated noticeable advancements in enhancing the 5-year locoregional control rate and reducing radiation-associated toxicities among NPC patients [4]. Accurate delineation of the GTV based on Magnetic Resonance Imaging (MRI) is critical in radiation therapy, particularly in IMRT for NPC [5]–[7]. Our focus is on MRI-based automatic GTV segmentation for its detailed soft tissue characterization, and we acknowledge the indispensable role of CT in radiation planning [8], [9], which complements the MRI. However, GTV contouring for NPC poses significant challenges due to its complex anatomical structures and diverse tumor invasion pathways. Manual GTV delineation is time-consuming, labor-intensive, and prone to errors and inter-observer variability [10]. These factors can adversely impact the accuracy of GTV contouring and potentially result in treatment failures [11].

Recently, the application of deep learning techniques has substantially advanced the field of medical auto-delineation, yielding successful models for various cancer types [12]–[14], including the NPC [15]. These models have demonstrated highly promising segmentation outcomes, showcasing their potential for precise and efficient delineation in clinical practice.

Despite the remarkable progress in GTV segmentation methods and the availability of data for NPC, there are still several challenges that hinder the widespread clinical adoption of deep learning techniques in this domain. Firstly, GTV segmentation tasks typically require a large amount of well-annotated MRI training samples, which necessitates extensive manual effort and time-consuming annotation processes. Secondly, deep models often lack generalizability across different data sources, as variations in equipment vendors, imaging resolution, slice thickness, and delineation styles among medical centers can impact model performance. Thirdly, strict data privacy regulations and diverse information systems in clinical settings limit data sharing, posing a challenge for transfer learning and domain adaptation approaches that rely on access to both source and target domain annotated data [16], [17].

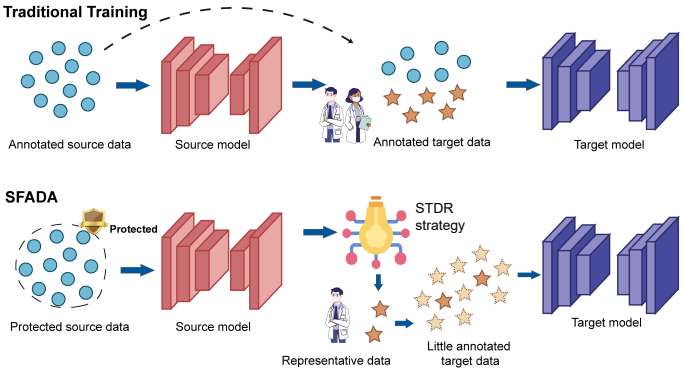

Fig. 1. Comparison of traditional training and our Source-Free Active Domain Adaptation (SFADA) training. Our approach safeguards the source data while demanding only a minimal annotation effort.

Addressing these challenges is paramount in facilitating the practical application of deep learning-based GTV segmentation methods for NPC radiotherapy.

We acknowledge the existence of UDA methods that partially address the aforementioned challenges. However, based on our primary experiments and research, we have found that while applying widely investigated UDA methods does yield some improvement compared to directly using a source domain model without fine-tuning, this improvement is quite limited. Notably, there still exists a substantial performance gap between UDA methods and fully supervised trained models. For instance, in some scenarios (medical image segmentation), this gap could potentially exceed 10% [18]–[20]. This considerable gap greatly reduces the practical value of computer-aided segmentation. Hence, there is a pressing need to enhance the effectiveness of domain adaptation techniques in medical image segmentation by exploring more practical and feasible approaches that can effectively bridge this performance gap.

In this paper, we pioneer the introduction of a novel approach called Source-Free Active Domain Adaptation for cross-domain GTV segmentation across different medical centers. This method is specifically designed to address the unique requirements and challenges in the field of medicine as shown in Fig. 1. To begin with, our approach eliminates the need for accessing the source data, ensuring strict data privacy compliance and adherence to hospital regulations. Meanwhile, it reduces the complexity of data transmission, resulting in significant time and resource savings. Additionally, our method facilitates knowledge transfer of model segmentation across diverse healthcare domains, promoting seamless collaboration and improving overall efficiency. Another pivotal aspect of our approach is the utilization of the proposed Source-domain and Target-domain Dual-Reference (STDR) strategy, enabling the active selection of representative domain-invariant and domain-specific samples for recommendation to radiation oncologists for annotation. This enables the model to acquire representations specific to the target domain while efficiently preserving the foundational segmentation knowledge that is common across both the source and target domains. It is important to highlight that this strategy is highly practical, as these selected samples constitute only a small fraction of the overall dataset. This approach effectively reduces the workload and saves valuable time for radiation oncologists, while enabling radiation oncologists to concentrate their efforts on the most representative and information-rich key samples for segmentation. Furthermore, subsequent experimental results demonstrate that our method attains comparable outcomes to those achieved through fully supervised training in the target domain. This underscores the remarkable practical value of our approach and its potential for real-world applications.

To substantiate the efficacy of our approach, we assemble a comprehensive dataset comprising MRI scans from 1057 NPC patients, sourced from five diverse medical centers for experiments on cross-domain GTV segmentation. Additionally, we conducted a thorough comparison with state-of-the-art (SOTA) domain adaptation methods (both with and without access to source data), as well as active learning methods to demonstrate the superiority of our method. The main contributions could be summarized as follows:

- To the best of our knowledge, this work is the first attempt to develop a Source-Free Active Domain Adaptation method for cross-domain GTV segmentation, which explores a new feasible solution for the practical implementation of computer-aided GTV segmentation, and remarkably improves the accuracy of segmentation.
- We propose an STDR to actively select representative samples for annotation and further reduce label costs without relying on source-domain data.
- We employ a semi-supervised learning strategy, leveraging both unlabeled and labeled samples chosen through our STDR strategy, to jointly train the model. This allows the model to learn from more diverse examples, therefore, enhancing the segmentation accuracy.
- Our approach outperforms other SOTA domain adaptation methods on five hospitals' clinical NPC datasets. Notably, the proposed method with a few labeled representative samples can match the performance of fully supervised training on the target domain. Moreover, we build the first multi-center NPC GTV dataset for public research.

II. Related Work

A. GTV Segmentation in NPC

In recent years, there have been numerous studies focusing on the segmentation of NPC [15]. These studies can be categorized into two main approaches: conventional machine learning-based methods and deep learning-based methods. Zhou *et al.* [21] introduce a segmentation technique for NPC based on a two-class support vector machine (SVM). This method aims to find an optimal hyperplane to effectively separate the different classes in the MR images. Huang *et al.* [22] present two region-based methods, the metric-based similarity learning and the discriminative classification-based with kernel learning for NPC segmentation. Men *et al.* [23] propose an end-to-end deep deconvolutional neural network to achieve pixel segmentation in CT images for NPC patients. Chen *et al.* [24] propose a multi-encoder single-decoder network for accurate segmentation of NPC using multi-modal MRI data. Liao *et al.* [25] evaluate the clinical value of a semi-supervised

learning framework for GTV delineation in NPC, demonstrating its accuracy and potential to assist oncologists in improving contouring accuracy and reducing contouring time. Li *et al.* [15] propose a comprehensive framework for primary NPC tumor segmentation, incorporating position enhancement, scale enhancement, and boundary enhancement modules. The results demonstrate the effectiveness of the approach in achieving accurate segmentation outcomes. While great progress has been made in NPC segmentation, effective methods for cross-center GTV segmentation remain lacking. Hence, our study focuses on leveraging segmentation knowledge from the source domain to the target domain, enabling comparable performance with minimal annotated target domain data.

B. Domain adaptation

Unsupervised Domain Adaptation: UDA is commonly trained using labeled data from the source domain and unlabeled data from the target domain. UDA encompasses various strategies to address domain shifts between source and target domains [26], [27], including Feature-level alignment (e.g., Statistic Divergence Alignment [28] and Normalization Statistics [29]), Input-level alignment (e.g., Generative Domain Mapping [30]), Self-training [31] and Self-supervision [32]. Tsai *et al.* [33] propose an adversarial learning domain adaptation algorithm, utilizing structured pixel-level predictions for spatial information encoding. Vu *et al.* [34] introduce a novel approach that maximizes prediction certainty in the target domain, improving UDA performance. Tang *et al.* [35] tackle instability by introducing a weak-strong augmented mean teacher learning approach. Huai *et al.* [36] propose a context-aware pseudo-label refinement method for UDA.

Semi-supervised Domain Adaptation: Semi-supervised approaches employ a completely annotated source dataset alongside a target dataset with partial labeling to facilitate domain adaptation [37]–[39]. Kim *et al.* [40] propose to align features by reducing the intra-domain differences by three strategies: attraction, perturbation, and exploration. Saito *et al.* [41] introduce a Minimax Entropy method, which employs adversarial optimization for refining an adaptable few-shot model. He *et al.* [42] develop a semi-supervised adversarial network that allows the knowledge transfer from the labeled videos to the heterogeneous audio domain.

Source-Free Domain Adaptation: Source-free domain adaptation (SFDA) is a technique that enables model adaptation to target domains without relying on source data. Following the comprehensive survey by Yu *et al.* [43], SFDA transfers knowledge and adapts models to target domains, addressing data availability and privacy concerns. This approach is particularly valuable in scenarios where source data is scarce or inaccessible, such as in medical settings [44]. Wang *et al.* [17] introduce a confidence optimization technique that leverages entropy-based model predictions for SFDA. Fleuret *et al.* [45] propose to mitigate prediction uncertainty in target data by minimizing the entropy and maximizing the robustness. Chen *et al.* [46] propose a denoised pseudo-labeling approach for SFDA by incorporating pixel-level and class-level denoising schemes. Yang *et al.* [47] develop an SFDA framework using Fourier Style Mining and Contrastive Domain Distillation. Bateson*et al.* [18] propose a new loss for unlabeled target-domain data, combining Shannon entropy with Kullback–Leibler divergence to align segmentation class-ratios with an anatomical prior. Hu *et al.* [48] propose a prompt learning, which prompts the source model to generate reliable predictions for target data. Besides the above learning-based methods, the classic intensity standardization technique, histogram matching (HistMatch), can also alleviate the domain gap by matching the intensity distribution between source and target domains [49], [50].

C. Active learning

Active learning is a strategy that aims to achieve optimal performance while minimizing annotation costs. It makes efficient utilization of the annotation budget by focusing on the most valuable samples for improving performance [51]. Various strategies have been developed for sample selection. These strategies include uncertainty-based methods that focus on selecting samples with high uncertainty [52], [53], diversity-based methods that aim to maximize the diversity of selected samples [54], [55], representativeness-based methods that prioritize samples representing different regions of the data distribution [56], [57], and strategies that aim to maximize the expected label changes [58], [59]. These diverse approaches provide different perspectives on selecting informative samples in active learning and these methods have been widely utilized in computer vision tasks, such as image classification and image segmentation. In this study, we propose the integration of active learning into domain adaptation to reduce the domain gap. Active learning offers the advantage of minimal annotation cost, making it feasible in many medical scenarios considering the potential performance improvements.

Currently, the application of active learning to the domain adaptation problem remains relatively limited. Su *et al.* [60] introduce an Active Adversarial Domain Adaptation (AADA) strategy, which combines domain adversarial learning and active learning. AADA takes into account both the uncertainty and diversity of the samples during the selection process, reducing labeling costs on the target domain on object classification and detection tasks. However, this adversarial-based learning approach relies on access to source data. Wang *et al.* [61] present a minimal happy point learning strategy for sample selection, which actively explores and utilizes minimal happy points as most informative samples, and validates its effectiveness on multiple categorical datasets. In this paper, we introduce a novel Source-Free Active Domain Adaptation method specifically designed for medical scenarios, addressing the challenge of cross-center medical image segmentation.

III. Methodology

In this chapter, we begin by presenting the construction of our cross-center GTV segmentation dataset. Subsequently, we provide a detailed description of our Source-domain and Target-domain Dual-Reference (STDR) strategy (as shown in Fig. 2), which assists in selecting the most representative

TABLE I
THE DATASET COMPRISES DATA ACQUIRED FROM FIVE DIVERSE MEDICAL CENTERS, INCLUDING INFORMATION ON THE VOLUME, VENDORS, AND MAGNETIC FIELD OF THE DATA.

Data source	Patients	Vendors (Patients)	Magnetic field (Patients)
Sichuan Cancer Hospital (SCH)	52	Siemens (52)	1.5T (23), 3T (29)
Anhui Provincial Hospital (APH)	146	GE (113), Siemens (27), Philips (6)	1.5T (119), 3T (27)
Sichuan Provincial People's Hospital (SPH)	208	Siemens (208)	1.5T (10), 3T (198)
West China Hospital (WCH)	284	GE (284)	3T (284)
Sothern Medical University (SMU)	367	GE (334), Siemens (24), Philips (9)	1.5T (336), 3T (31)

samples and helps us to learn about domain-invariant and domain-specific knowledge. In addition, Fig. 5 describes the semi-supervised workflow. This allows the network to effectively utilize the abundant unlabeled data and improve the generalization capability of the model. Finally, we provide details on the training and implementation of our model.

A. Dataset Construction

We acquired the NPC MRI datasets with annotated labels from five renowned medical institutions, focusing specifically on patients diagnosed with NPC. The MR images are obtained by utilizing different scanners with 1.5-T or 3.0-T, with a large range of inter-slice thickness (range, 1.0-8.0 mm) and a middle range in-plane spacing (range, 0.47-1.67 mm). Strict selection criteria are applied, including histological confirmation of NPC and prior MRI evaluations of the nasopharynx and neck before initiating anticancer treatments. The MRI sequences of this dataset consisted of contrast-enhanced T1-weighted sequences as well as unenhanced T1- and T2-weighted sequences. Since we are focused on single-modality SFADA segmentation, to exclude interference, all our experiments use only contrast-enhanced T1-weighted MR images. The basic information regarding the data distribution and vendors of the multi-center NPC MRI datasets is provided in Table I. Notably, the t-SNE visualization [62] depicted in Fig. 3 demonstrates significant variations in the distributions of the different medical center domains. Our primary objective is to overcome potential domain shift challenges and achieve accurate GTV delineations for patients within the designated medical center.

B. Source-domain and Target-domain Dual-Reference

Fig. 2 presents a schematic illustration of our STDR framework. Our Source-Free Active Domain Adaptation task differs significantly from past active learning for domain adaptation as we lack access to the source data during model fine-tuning. This distinguishes our approach from methods that can be jointly trained using both source and target domain data. In cross-center GTV segmentation, a significant challenge arises from the domain gap in data distribution between different centers, while still sharing some common structural segmentation knowledge. However, learning on a small number of similar active samples selected from the target data easily leads to overfitting [63], [64]. Therefore, our proposed STDR strategy serves as a reference by jointly considering the data from the source and target domains to select the source domain representative samples P^s and the target domain representative samples P^t. This approach offers two key advantages. During fine-tuning in the target domain, samples P^s reinforce the sharing segmentation knowledge without accessing the source data and aid in learning domain-invariant representations. Simultaneously, samples P^t facilitate effective migration of the model to the target domain data, thereby overcoming potential domain shift challenges. Fig. 4 illustrates some examples of MRI images selected by STDR.

The input for the entire framework consists of the following components: the MR images of NPC patients from the source center, denoted as I^s: $\Omega_s \subset \mathcal{R}^2 \rightarrow \mathcal{R}$, their respective manual labels Y^s for each pixel $i \in \Omega_s$, where $y_s(i) \in \{0, 1\}$, and the MR images I^t from the target center. In the SFDA scenario, we begin by training the segmentation network on source images I^s and their corresponding ground truth labels Y^s in a fully supervised manner. This entails minimizing the loss function with respect to the network parameters θ_s, which is formulated as below:

$$L_s(\theta_s, \Omega_s) = \frac{1}{|\Omega_s|} \sum_{i=1}^{S} \Psi(y_s(i), p_s(i, \theta_s)). \tag{1}$$

The softmax output of the segmentation network at pixel i in the source image I^s is denoted as $p_s(i, \theta) \in [0, 1]$. We utilize a composite loss function Ψ defined as follows:

$$\begin{aligned} \Psi(y_s(i), p_s(i, \theta)) = &- \sum_{i=1}^{N} y_s(i) \log p_s(i, \theta) \\ &+ 1 - \frac{2 \sum_{i=1}^{N} y_s(i) \hat{y}_s(i)}{\sum_{i=1}^{N} y_s(i)^2 + \sum_{i=1}^{N} \hat{y}_s(i)^2}, \end{aligned} \tag{2}$$

Where $\hat{y}_s(i)$ denotes the i_{th} element of the predicted masks, and N represents the overall count of elements.

Dual-Reference Selection. Considering the presence of multiple radiation oncologists performing GTV outlining and the probable utilization of diverse imaging devices from various vendors in a medical center, it is essential to select multiple reference points R^s to effectively capture this data domain's distributional characteristics. These multiple reference points R^s ensure a comprehensive representation of the data and accounts for potential variations introduced by different radiation oncologists and imaging equipment. We initially freeze the above pre-trained segmentation network. Then, we merge the feature embeddings from the penultimate layer of the network with the corresponding predicted masks to help the model focus on the features predicted to be GTV regions. The projection is as follows:

$$F^s(I^s(i)) = \frac{1}{\|N^s\|} MaxP_f(AvgP_c(\hat{Y}^s(i) \otimes f_E(I^s(i)))), \tag{3}$$

where $\|N^s\|$ represents the total number of pixels belonging to the GTV, $\hat{Y}_s(i)$ denotes the i_{th} predicted mask, and $f_E(I^s(i)) \in R^{C \times H \times W}$ represents the output feature embedding from the penultimate layer of i_{th} MRI image. $MaxP_f$ is the max-pooling operation on image features with spatial dimensions $R^{H \times W}$, while $AvgP_c$ is the average pooling operation along the channel dimension R^C, reducing the feature tensor from three dimensions to two dimensions. Notably, in our dense prediction task, we select a small kernel size in Max pooling to extract features from multiple image patches. This approach provides a more comprehensive representation

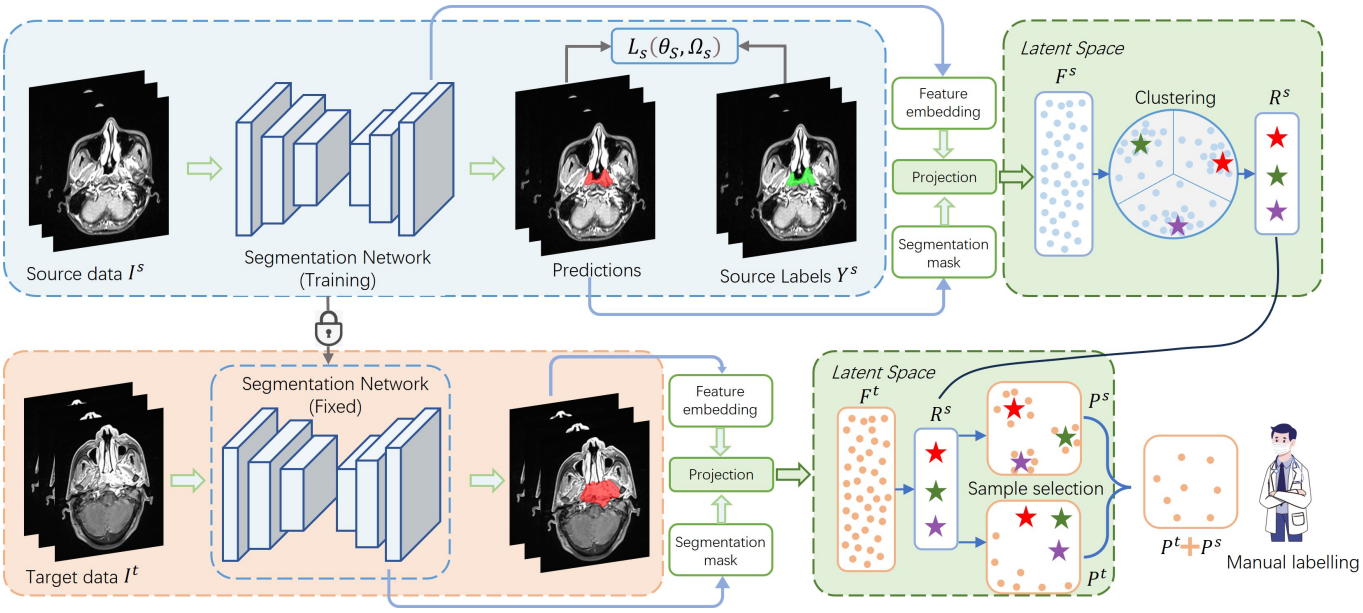

Fig. 2. Overview of the STDR strategy. Given the source MR images $\boldsymbol{I^s}$ accompanied by its ground truth masks $\boldsymbol{Y^s}$, our STDR trains the segmentation network in the source domain first. The segmented features are then mapped to the latent space as $\boldsymbol{F^s}$, and clustering algorithms derive the reference point $\boldsymbol{R^s}$. The data in the target domain undergoes similar manipulation to obtain $\boldsymbol{F^t}$. We then select two types of representative samples for manual labeling by radiation oncologists.

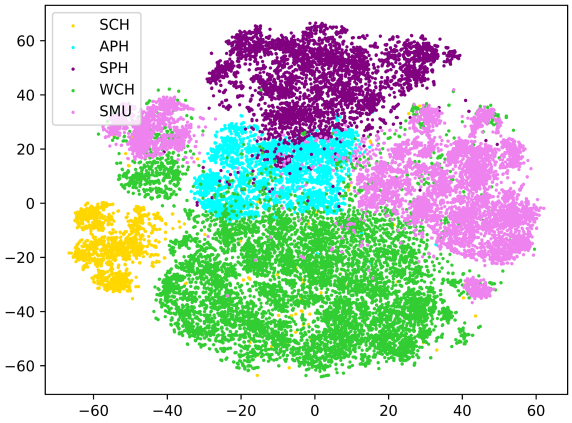

Fig. 3. Visualization (t-SNE [62]) of the multiple medical centers' distribution. The dots of different colors indicate the latent space representations of the samples from different medical centers, respectively.

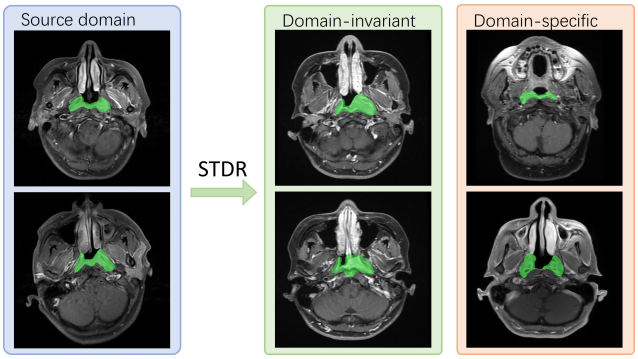

Fig. 4. Examples of source domain (SPH), domain-invariant and domain-specific samples (APH). The source domain images present higher similarity with the domain-invariant images, while the discrepancies with the domain-specific images become more obvious.

compared to pooling the entire image into a single element. The final F^s is a flattening of this two-dimensional vector (a 1×256 vector here).

Subsequently, we employ the clustering method (the widely-used K-means [65]) to group feature vectors from all source images into K distinct clusters. This process is crucial for our methodology as it helps identify representative characteristics within the source data, which are then used for further analysis and comparison. Clustering is done aiming to minimize the following error:

$$\sum_{k=1}^{K} \sum_{i \in R_k} \|F^s(I^s(i)) - R_k^s\|_2^2, \tag{4}$$

where $\|\cdot\|_2^2$ refers to the Euclidean distance, and R_k^s denotes the representative point of the cluster R_k, as:

$$R_k^s = \frac{1}{\|R_k\|} \sum_{i \in R_k} F^s(I^s(i)). \tag{5}$$

The $\|R_k\|$ represents the number of images belonging to cluster R_k. The centroids R^s serve as essential reference points for comparing and selecting target images during the active sample selection process.

Active Sample Selection. As depicted in Fig. 2, the model's parameters trained in the source domain are kept fixed and applied to process MR images in the target domain. The resulting features are then mapped to the latent space to obtain F^t, following the procedure described in Equation 3. Next, we compute the Euclidean distances between $F^t(I^t(i))$ and all source-domain references R^s, and designate the smallest distance as the measure of the target-domain sample's similarity to the source domain:

$$Similarity(I^t(i)) = \min_k \left\|F^t(I^t(i)) - R_k^s\right\|_2^2, \tag{6}$$

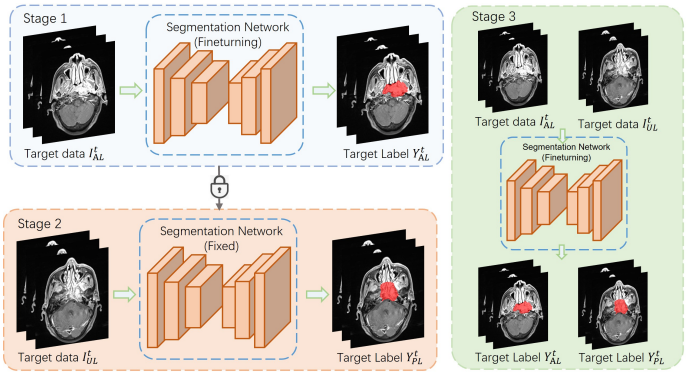

Fig. 5. Detailed pipeline of the proposed semi-supervised method. The whole process consists of three stages.

where $F^t(\cdot)$ is similar to Equation 3 and represents the projection function for data in the target domain. In our source-free setting, the active learning sample selection encompasses both target domain representative samples P^t and source domain representative samples P^s. This approach ensures the preservation of the sharing domain-invariant segmentation knowledge from the source domain, even during the migration to the target domain. the $Selected(P^s, P^t)$ is specified in the following way:

$$Selected = Min(Similarity)[\frac{p}{2}\%]+Max(Similarity)[\frac{p}{2}\%], \tag{7}$$

where the $Similarity$ represents a list of the similarities of all the target domain data, and $p\%$ represents the proportion of samples to be selected for active learning. Our subsequent ablation experiments further demonstrate that this approach effectively preserves shared domain-invariant segmentation knowledge, leading to improved segmentation accuracy.

C. Semi-supervised Learning for Source-free Active Domain Adaptation

In our designed semi-supervised procedure, we employ a three-stage approach. The stage 1 entails fully supervised fine-tuning of the model parameters using actively labeled samples I^t_{AL}. The detailed loss function is as follows:

$$L_s(\theta_t, \Omega_t) = \frac{1}{|\Omega_t|}\sum_{i=1}^{T}\Psi(y_t(i), p_t(i, \theta_t)), \tag{8}$$

where the θ_t represent the fine-turning network parameters, $y_t(i) \in \{0, 1\}$ denotes the pixel labels from Y^t_{AL} and $p_t(i, \theta_t)$ is the predicted value of the network. The Ψ is the same as Equation 2. Subsequently, in stage 2, we keep these fine-tuned model parameters fixed and utilize them to infer pseudo-labels Y^t_{PL} for the unlabeled data I^t_{UL}. Finally, in stage 3, we jointly fine-tune the model parameters using the inferred pseudo-labels Y^t_{PL} and the actively labeled samples Y^t_{AL}. It should be noted that here we utilize the identical loss function as in stage 1, with the exception that it incorporates more comprehensive data.

This multi-stage process allows us to effectively leverage both labeled and unlabeled data. The initial fine-tuning of the actively labeled samples enables the model to learn the feature distribution of the target domain while preserving the shared domain-invariant segmentation knowledge. This establishes a strong foundation for the model, enabling it to generate higher-quality pseudo-labels. By using the fixed model to generate pseudo-labels Y^t_{AL} for the unlabeled data I^t_{UL}, we expand the training dataset and gain additional valuable information. Finally, incorporating the inferred pseudo-labels alongside the actively selected samples in the last fine-tuning stage facilitates further refinement of the model, leading to improved performance and generalization. Our subsequent ablation experiments will further validate this aspect.

D. Model Training and Implementation Details

To maintain objectivity during evaluation, we partition each dataset randomly into three distinct subsets (train: valid: test) using a 7: 1: 2 ratio. In our setting, 70% of the data from each medical center is allocated for model training and parameter tuning. Subsequently, we select the model with the best performance on the validation set and report its results on the test set. Considering the potential benefits of larger datasets in terms of richer segmentation samples and knowledge, we opt to designate the SMU center with the largest data volume, as our source domain. Subsequently, we utilized the datasets from the remaining four centers as target domains for Source-Free Active Domain Adaptation. Furthermore, to assess the generalization and robustness of our proposed method, we conduct additional experiments using data from the SPH center as the source domain for training. Subsequently, we migrate the trained model to other datasets using our approach to evaluate its performance and effectiveness.

Considering the huge different thicknesses of MR images from different hospitals, we segment GTV slice by slice in the axis slices and then stack them as volumetric predictions. We use 2D U-Net [66] as the baseline for all methods, the implementations are based on the PyMIC [67]. All experiments are executed on an NVIDIA RTX 3090 GPU with 24 GB of memory. To standardize the training process, all model inputs are adjusted to a uniform resolution of 256×256 and normalized the intensity to zero mean and unit variance. Furthermore, to boost the model's robustness, random rotation, flip, and Gaussian noise are applied for data augmentation. All models are trained using the SGD optimizer and a batch size of 32. The cluster count (denoted as K) is set to 10 according to the ablation study, and the $p\%$ is set to 20% in our experiments. When training with all data in the source or target domain, all models are trained with $30k$ iterations. When training with actively selected samples, the data size is small, and therefore $20k$ iterations are trained. The initial learning rate is set to 0.03 and decays exponentially at a factor of 0.9 after each iteration. To be fair comparisons, we re-implement all comparison methods with the same backbone (U-Net) and run them in the same settings. For quantitative evaluation, we measure the volumetric-level Dice Similarity Coefficient, DSC (%), 95% Hausdorff distance, HD95 (mm), and Average Surface Distance, ASD (mm) between the ground truth and predictions. Generally, a superior method should yield a higher DSC score and lower values for HD95 and ASD.

TABLE II

QUANTITATIVE RESULTS OF THE MODEL TRAINED AND TESTED ON DIFFERENT DATASETS.

Model	Dataset	DSC (%)	HD95 (mm)	ASD (mm)
U-Net	SCH	72.19 ± 5.97	9.25 ± 3.52	2.22 ± 0.86
U-Net	APH	86.79 ± 5.14	4.49 ± 3.35	1.13 ± 1.03
U-Net	SPH	80.16 ± 8.09	4.42 ± 1.63	1.42 ± 0.71
U-Net	WCH	85.87 ± 4.85	4.15 ± 2.47	1.17 ± 0.59
U-Net	SMU	85.31 ± 7.45	5.88 ± 14.83	0.76 ± 0.69

TABLE III

QUANTITATIVE RESULTS WITH MODELS TRAINED ON DIFFERENT DATASETS AND TESTED ON OTHER DATASETS.

Model	Adaptation setting	DSC (%)	HD95 (mm)	ASD (mm)
U-Net	SPH $\rightarrow$ SCH	59.11 ± 8.98	31.35 ± 11.99	3.13 ± 1.73
U-Net	SPH $\rightarrow$ APH	78.25 ± 9.42	9.14 ± 8.49	2.68 ± 2.95
U-Net	SPH $\rightarrow$ WCH	74.76 ± 11.19	13.09 ± 12.78	4.02 ± 4.44
U-Net	SPH $\rightarrow$ SMU	73.71 ± 13.55	12.89 ± 20.91	3.41 ± 5.92
U-Net	SMU $\rightarrow$ SCH	63.35 ± 9.98	42.88 ± 17.29	2.49 ± 2.69
U-Net	SMU $\rightarrow$ APH	81.45 ± 8.12	8.31 ± 8.97	2.60 ± 3.37
U-Net	SMU $\rightarrow$ SPH	77.51 ± 8.49	9.42 ± 17.59	2.96 ± 4.95
U-Net	SMU $\rightarrow$ WCH	77.40 ± 8.47	7.94 ± 7.07	2.37 ± 1.70

IV. EXPERIMENTS AND RESULTS

In this section, we first introduce the performance of the models trained in the respective data centers, and the performance of the trained models directly applied to other data centers. Subsequently, we compare our approach with state-of-the-art methods comprehensively, and finally, we report ablation experimental results to demonstrate the effectiveness of our design.

A. Model training on source domain data

Initially, we train the models on the training sets of the five medical centers separately. The selected models from the validation sets are subsequently evaluated on the corresponding test sets of each medical center. The results of these evaluations are presented in Table II. The results indicate that even the same model and training strategy could produce varying outcomes across different data centers. Specifically, the segmentation result at SCH achieved a DSC of only 72.19%, which may be attributed to more intricate patient conditions and smaller dataset sizes. In contrast, the other four centers yielded a DSC of over 80%. In addition, our findings suggest that a larger dataset does not necessarily yield superior results. For instance, the performance at APH, which only had 146 patients, outperforms that at SPH, which had 208 patients.

Table III presents the performances of deploying the segmentation model to other medical centers directly, after training it on SPH and SMU datasets. Our results demonstrate that the existence of domain gaps leads to a decrease in DSC accuracy of 5-10% in the majority of cases, such as SPH $\rightarrow$ APH and SMU $\rightarrow$ SCH. In some cases, however, we observed a DSC accuracy degradation of more than 10%, such as SPH $\rightarrow$ SCH and SPH $\rightarrow$ SMU. This further confirms that deploying computer-aided GTV segmentation algorithms in different medical centers requires active source-free domain adaptation of the model. In addition, we can also find that the model transferred from SMU performed better than the one from SPH. This observation inspires us that in practical applications, we can adapt the model trained on a dataset with a larger volume of data to smaller data centers using source-free active domain adaptation to yield superior results.

TABLE IV

QUANTITATIVE RESULTS WITH MODELS TRAINED ON SPH DATASET AND ADAPTED ON THE OTHERS. THE $*$ DENOTES p-VALUE < 0.05 IN ALL PAIRED t-TEST, INDICATING OUR METHOD SIGNIFICANTLY OUTPERFORMED THE OTHERS.

Model	Target	Source data	DSC (%)	HD95 (mm)	ASD (mm)
HistMatch [50]	SCH	No	57.84 ± 13.54	27.55 ± 12.15	4.33 ± 5.84
HistMatch [50]	APH	No	77.82 ± 13.57	8.03 ± 8.76	2.54 ± 2.13
HistMatch [50]	WCH	No	72.69 ± 10.84	10.50 ± 10.52	3.29 ± 2.67
HistMatch [50]	SMU	No	72.44 ± 17.93	9.70 ± 16.98	2.10 ± 4.59
AdaptSeg [33]	SCH	Yes	59.63 ± 11.03	27.84 ± 17.52	7.31 ± 5.12
AdaptSeg [33]	APH	Yes	77.87 ± 15.26	7.62 ± 7.19	1.97 ± 1.43
AdaptSeg [33]	WCH	Yes	74.95 ± 8.64	10.74 ± 7.71	3.64 ± 2.45
AdaptSeg [33]	SMU	Yes	74.38 ± 12.58	12.05 ± 21.64	3.11 ± 4.97
AdvEnt [34]	SCH	Yes	60.14 ± 10.16	20.49 ± 16.21	4.65 ± 2.72
AdvEnt [34]	APH	Yes	79.71 ± 11.37	7.13 ± 5.67	2.18 ± 1.63
AdvEnt [34]	WCH	Yes	75.27 ± 8.22	9.94 ± 6.69	3.19 ± 2.08
AdvEnt [34]	SMU	Yes	75.56 ± 11.47	10.41 ± 19.32	2.79 ± 4.68
UncertainDA [45]	SCH	No	61.74 ± 11.24	21.29 ± 16.91	4.84 ± 2.72
UncertainDA [45]	APH	No	80.03 ± 11.56	6.81 ± 5.22	2.01 ± 1.76
UncertainDA [45]	WCH	No	75.88 ± 8.81	12.14 ± 12.50	3.65 ± 4.02
UncertainDA [45]	SMU	No	76.02 ± 12.67	9.45 ± 18.96	2.06 ± 4.57
Tent [17]	SCH	No	62.37 ± 9.74	26.02 ± 17.27	5.45 ± 5.37
Tent [17]	APH	No	80.73 ± 10.46	6.85 ± 6.15	1.99 ± 1.81
Tent [17]	WCH	No	75.81 ± 10.31	11.24 ± 10.51	3.51 ± 3.57
Tent [17]	SMU	No	75.88 ± 12.15	8.17 ± 17.52	1.94 ± 3.14
DPL [46]	SCH	No	62.88 ± 10.36	21.29 ± 14.56	4.27 ± 3.16
DPL [46]	APH	No	80.96 ± 12.14	6.23 ± 5.15	1.78 ± 1.49
DPL [46]	WCH	No	77.26 ± 9.83	8.96 ± 5.73	2.81 ± 1.55
DPL [46]	SMU	No	77.24 ± 10.53	9.22 ± 18.77	1.85 ± 3.23
FSM [47]	SCH	No	62.85 ± 10.80	21.71 ± 18.17	5.42 ± 4.64
FSM [47]	APH	No	80.07 ± 8.62	6.25 ± 3.93	1.83 ± 1.28
FSM [47]	WCH	No	75.56 ± 8.58	10.30 ± 6.66	3.01 ± 1.68
FSM [47]	SMU	No	76.23 ± 14.25	9.93 ± 17.44	2.14 ± 3.63
AdaMI [18]	SCH	No	62.24 ± 11.43	21.85 ± 19.63	5.32 ± 4.21
AdaMI [18]	APH	No	79.88 ± 11.81	6.09 ± 4.83	1.74 ± 1.39
AdaMI [18]	WCH	No	77.07 ± 9.78	8.35 ± 5.98	2.74 ± 1.42
AdaMI [18]	SMU	No	77.13 ± 14.88	8.79 ± 15.38	2.28 ± 2.75
ProSFDA: [48]	SCH	No	63.11 ± 9.75	21.68 ± 20.06	6.13 ± 5.99
ProSFDA: [48]	APH	No	80.97 ± 11.67	6.37 ± 3.95	1.72 ± 1.34
ProSFDA: [48]	WCH	No	77.81 ± 8.58	8.24 ± 5.45	2.93 ± 1.81
ProSFDA: [48]	SMU	No	77.65 ± 9.72	8.42 ± 15.13	2.09 ± 2.71
CBMT [35]	SCH	No	62.74 ± 10.33	21.57 ± 18.06	6.09 ± 6.59
CBMT [35]	APH	No	80.17 ± 12.31	5.90 ± 3.98	1.77 ± 1.51
CBMT [35]	WCH	No	77.28 ± 9.54	8.47 ± 5.60	2.85 ± 1.38
CBMT [35]	SMU	No	77.36 ± 9.84	9.29 ± 20.16	2.05 ± 3.09
CPR [36]	SCH	No	63.42 ± 9.02	21.72 ± 19.69	5.81 ± 5.54
CPR [36]	APH	No	81.33 ± 10.60	6.05 ± 4.12	1.73 ± 1.14
CPR [36]	WCH	No	78.02 ± 8.85	8.14 ± 5.32	2.73 ± 1.31
CPR [36]	SMU	No	77.99 ± 9.77	8.77 ± 15.16	1.82 ± 2.65
Ours	SCH	No	$\mathbf{70.99} \pm 6.59*$	$\mathbf{8.57} \pm 2.39*$	$\mathbf{2.03} \pm 0.84*$
Ours	APH	No	$\mathbf{85.69} \pm 4.54*$	$\mathbf{4.58} \pm 2.24*$	$\mathbf{1.18} \pm 0.78*$
Ours	WCH	No	$\mathbf{83.59} \pm 6.76*$	$\mathbf{4.85} \pm 2.79*$	$\mathbf{1.37} \pm 0.92*$
Ours	SMU	No	$\mathbf{84.13} \pm 7.65*$	$\mathbf{6.17} \pm 15.33*$	$\mathbf{0.86} \pm 0.95*$

B. Comparison with State-of-the-art Methods

To ensure a comprehensive evaluation of our method's performance, we extended our analysis beyond merely comparing it with other SOTA domain adaptation methods (Tables IV and V). We also conducted a performance comparison with SOTA active learning methods (Tables VI and VII). Under both comparisons, we conducted extensive experiments containing two kinds of data settings (SPH $\rightarrow$ other datasets, SMU $\rightarrow$ other datasets). Our methods consistently demonstrated superior performance across all experiments, affirming the effectiveness of our approach.

1) Comparison with State-of-the-art Domain adaptation methods: We conduct a comparative analysis between our approach and SOTA domain adaptation methods to assess their performance. This evaluation encompasses methods with (AdaptSeg [33], AdvEnt [34]) and without (HistMatch [50], UncertainDA [45], Tent [17], DPL [46], FSM [47], AdaMI [18], ProSFDA [48], CBMT [35] and CPR [36]) access to source domain data. As anticipated, our proposed method exhibits significant enhancements compared to the other domain adaptation methods in both Table IV and Table V. This outcome underscores the impactful role of strategically selected active samples, in tandem with the semi-supervised

TABLE V

QUANTITATIVE RESULTS WITH MODELS TRAINED ON SMU DATASET AND ADAPTED ON THE OTHERS. THE $*$ DENOTES p-VALUE < 0.05 IN ALL PAIRED t-TEST, INDICATING OUR METHOD SIGNIFICANTLY OUTPERFORMED THE OTHERS.

Model	Target	Source data	DSC (%)	HD95 (mm)	ASD (mm)
HistMatch [50]	SCH	No	62.66 ± 10.49	41.66 ± 16.74	3.14 ± 4.94
HistMatch [50]	APH	No	80.25 ± 9.35	7.88 ± 8.59	2.36 ± 2.96
HistMatch [50]	SPH	No	75.72 ± 12.88	8.06 ± 12.01	1.96 ± 1.81
HistMatch [50]	WCH	No	74.01 ± 15.16	7.56 ± 5.29	1.78 ± 1.29
AdaptSeg [33]	SCH	Yes	63.56 ± 10.57	41.38 ± 15.13	2.65 ± 2.73
AdaptSeg [33]	APH	Yes	82.11 ± 7.15	8.26 ± 8.67	2.80 ± 2.21
AdaptSeg [33]	SPH	Yes	76.87 ± 9.62	8.85 ± 11.99	2.57 ± 2.07
AdaptSeg [33]	WCH	Yes	77.95 ± 8.57	7.39 ± 4.86	2.00 ± 1.19
AdvEnt [34]	SCH	Yes	64.06 ± 7.57	36.18 ± 14.12	2.90 ± 1.76
AdvEnt [34]	APH	Yes	80.91 ± 6.31	6.91 ± 5.93	2.33 ± 1.97
AdvEnt [34]	SPH	Yes	76.71 ± 8.99	6.97 ± 11.29	1.90 ± 1.75
AdvEnt [34]	WCH	Yes	78.37 ± 9.03	7.15 ± 4.84	1.95 ± 1.48
UncertainDA [45]	SCH	No	63.59 ± 8.32	28.73 ± 12.28	4.38 ± 2.85
UncertainDA [45]	APH	No	81.17 ± 7.11	7.96 ± 4.05	2.24 ± 1.28
UncertainDA [45]	SPH	No	77.88 ± 9.58	6.44 ± 11.56	1.89 ± 1.79
UncertainDA [45]	WCH	No	78.16 ± 9.10	6.67 ± 4.44	1.87 ± 1.07
Tent [17]	SCH	No	64.43 ± 7.55	29.25 ± 16.34	3.39 ± 2.23
Tent [17]	APH	No	80.49 ± 7.58	7.94 ± 7.14	2.48 ± 2.56
Tent [17]	SPH	No	78.46 ± 9.15	7.37 ± 12.02	2.18 ± 2.44
Tent [17]	WCH	No	78.13 ± 8.79	6.46 ± 4.09	1.65 ± 0.99
DPL [46]	SCH	No	64.51 ± 7.70	24.74 ± 7.27	3.83 ± 1.85
DPL [46]	APH	No	81.43 ± 6.74	6.38 ± 5.68	1.97 ± 2.02
DPL [46]	SPH	No	78.83 ± 8.38	6.89 ± 12.67	1.94 ± 2.95
DPL [46]	WCH	No	78.92 ± 8.42	6.26 ± 3.79	1.72 ± 1.04
FSM [47]	SCH	No	63.82 ± 9.15	22.65 ± 13.07	5.56 ± 3.44
FSM [47]	APH	No	82.08 ± 12.40	6.94 ± 5.85	2.61 ± 1.77
FSM [47]	SPH	No	78.16 ± 7.91	7.12 ± 12.79	2.56 ± 2.76
FSM [47]	WCH	No	78.45 ± 7.45	6.61 ± 3.95	1.94 ± 1.18
AdaMI [18]	SCH	No	63.76 ± 7.32	25.86 ± 17.61	5.31 ± 4.68
AdaMI [18]	APH	No	81.67 ± 8.52	6.04 ± 4.24	2.25 ± 1.84
AdaMI [18]	SPH	No	78.04 ± 8.57	7.32 ± 13.45	2.30 ± 2.26
AdaMI [18]	WCH	No	78.43 ± 9.16	6.37 ± 3.86	1.93 ± 0.96
ProSFDA: [48]	SCH	No	64.05 ± 9.05	21.58 ± 14.41	4.35 ± 3.39
ProSFDA: [48]	APH	No	82.29 ± 8.05	6.36 ± 4.93	2.34 ± 1.94
ProSFDA: [48]	SPH	No	78.61 ± 9.68	6.36 ± 6.78	1.97 ± 0.93
ProSFDA: [48]	WCH	No	79.04 ± 7.05	6.46 ± 3.67	1.86 ± 1.07
CBMT [35]	SCH	No	64.65 ± 9.17	26.71 ± 19.82	5.01 ± 3.96
CBMT [35]	APH	No	81.93 ± 8.35	6.52 ± 4.99	2.41 ± 1.96
CBMT [35]	SPH	No	78.63 ± 7.84	7.04 ± 12.33	2.01 ± 2.66
CBMT [35]	WCH	No	79.05 ± 7.14	6.72 ± 4.02	2.02 ± 1.18
CPR [36]	SCH	No	64.36 ± 10.79	20.24 ± 13.84	3.49 ± 2.39
CPR [36]	APH	No	82.45 ± 8.37	6.05 ± 4.03	2.38 ± 1.95
CPR [36]	SPH	No	78.91 ± 8.51	7.29 ± 4.57	2.07 ± 1.99
CPR [36]	WCH	No	79.25 ± 7.07	6.59 ± 3.85	1.89 ± 1.28
Ours	SCH	No	**72.58** $\pm 6.72*$	**8.06** $\pm 2.06*$	**2.00** $\pm 0.89*$
Ours	APH	No	**86.01** $\pm 5.10*$	**4.38** $\pm 2.51*$	**1.21** $\pm 1.04*$
Ours	SPH	No	**80.97** $\pm 7.24*$	**4.02** $\pm 1.62*$	**1.03** $\pm 0.58*$
Ours	WCH	No	**84.71** $\pm 6.04*$	**4.48** $\pm 2.23*$	**1.24** $\pm 0.77*$

pipeline, in achieving substantial performance advancements. As an illustration, our method demonstrates a superiority over the best DSC achieved by other approaches, surpassing them by 8.11% and 8.07% points on SPH $\rightarrow$ SCH and SMU $\rightarrow$ SCH, respectively. Fig. 6 additionally provides an intuitive visual representation of segmentation instances using various methods.

2) Comparison with State-of-the-art Active learning methods: Meanwhile, considering that our active domain adaptation framework inherently incorporates the concept of active learning, it is reasonable to conduct a comparison with other SOTA active learning methods. Therefore, we opt to conduct a comparison between our approach and random selection, alongside four SOTA active learning methods. The details are as follows: (i) Random Selection: Samples are randomly chosen from the target domain with uniform probability; (ii) Adversarial [33]: Leveraging the trained discriminator following [33], target samples are selected with the lowest predicted probabilities, indicating those that exhibit the most pronounced divergence from the source domain; (iii) Entropy [34]: The AdvEnt method [34] is employed to calculate the prediction map entropy for each sample within the target domain, and those samples with the highest entropy are selected for manual annotation; (iv) AADA [60]: AADA considers both sample uncertainty and diversity when making selections; (v) MHPL [61]: MHPL finds neighbor-chaotic, individual-different, and target-like points as representative samples.

TABLE VI

EXPERIMENTAL RESULTS ON VARIOUS ACTIVE SAMPLE SELECTION METHODS FROM SPH DATASET TO OTHER DATASETS. THE $*$ DENOTES p-VALUE < 0.05 IN ALL PAIRED t-TEST, INDICATING OUR METHOD SIGNIFICANTLY OUTPERFORMED THE OTHERS.

Model	Target	DSC (%)	HD95 (mm)	ASD (mm)	NSD (%)
Random	SCH	62.28 ± 9.72	30.48 ± 10.59	**2.64** ± 1.61	49.05 ± 6.37
Random	APH	81.19 ± 8.69	6.55 ± 4.82	2.01 ± 1.5	68.97 ± 12.44
Random	WCH	78.05 ± 7.40	6.53 ± 6.20	1.92 ± 1.94	60.90 ± 10.95
Random	SMU	78.75 ± 10.26	7.36 ± 14.77	1.37 ± 2.45	68.79 ± 13.01
Adversarial [33]	SCH	63.21 ± 7.23	24.75 ± 8.15	3.42 ± 1.61	49.47 ± 6.26
Adversarial [33]	APH	82.21 ± 6.32	8.27 ± 7.43	2.35 ± 1.81	69.40 ± 11.19
Adversarial [33]	WCH	79.00 ± 7.87	7.07 ± 5.20	2.21 ± 1.42	62.68 ± 10.80
Adversarial [33]	SMU	79.44 ± 9.39	7.17 ± 12.57	1.58 ± 1.51	69.36 ± 12.63
Entropy [34]	SCH	65.02 ± 8.67	18.03 ± 7.34	3.34 ± 1.87	54.45 ± 11.13
Entropy [34]	APH	83.45 ± 5.59	7.59 ± 9.45	1.82 ± 2.12	73.51 ± 12.03
Entropy [34]	WCH	80.97 ± 7.98	7.50 ± 5.98	2.22 ± 1.37	66.79 ± 10.12
Entropy [34]	SMU	80.27 ± 10.34	7.05 ± 14.74	1.43 ± 2.02	70.83 ± 12.54
AADA [60]	SCH	65.08 ± 9.82	14.45 ± 5.90	2.88 ± 1.67	54.25 ± 8.87
AADA [60]	APH	84.14 ± 5.05	7.87 ± 7.66	1.91 ± 1.95	73.34 ± 11.94
AADA [60]	WCH	81.21 ± 7.23	6.52 ± 3.81	2.01 ± 1.05	67.43 ± 11.45
AADA [60]	SMU	80.78 ± 8.60	6.33 ± 12.53	**1.12** ± 1.62	71.47 ± 12.80
MHPL [61]	SCH	67.35 ± 8.21	14.99 ± 8.61	3.41 ± 2.02	59.54 ± 9.15
MHPL [61]	APH	84.04 ± 5.17	5.77 ± 5.78	1.75 ± 1.42	74.02 ± 11.78
MHPL [61]	WCH	80.40 ± 8.12	6.35 ± 4.74	1.81 ± 1.11	67.75 ± 13.38
MHPL [61]	SMU	81.61 ± 8.32	7.76 ± 16.41	1.33 ± 1.09	73.36 ± 12.29
STDR	SCH	**69.55** $\pm 6.82*$	**10.64** $\pm 3.36*$	2.83 ± 1.83	**61.78** $\pm 10.05*$
STDR	APH	**85.05** $\pm 5.25*$	**4.98** $\pm 3.07*$	**1.36** $\pm 1.01*$	**75.21** $\pm 11.86*$
STDR	WCH	**82.81** $\pm 6.51*$	**5.39** $\pm 2.77*$	**1.58** $\pm 0.88*$	**71.17** $\pm 10.47*$
STDR	SMU	**83.51** $\pm 8.05*$	**5.81** ± 13.10	1.26 ± 1.56	**76.35** $\pm 12.22*$

TABLE VII

EXPERIMENTAL RESULTS ON VARIOUS ACTIVE SAMPLE SELECTION METHODS FROM SMU DATASET TO OTHER DATASETS. THE $*$ DENOTES p-VALUE < 0.05 IN ALL PAIRED t-TEST, INDICATING OUR METHOD SIGNIFICANTLY OUTPERFORMED THE OTHERS.

Model	Target	DSC (%)	HD95 (mm)	ASD (mm)	NSD (%)
Random	SCH	66.67 ± 7.51	15.06 ± 7.28	3.35 ± 2.23	59.31 ± 11.61
Random	APH	82.66 ± 6.37	7.28 ± 9.62	2.26 ± 1.79	71.74 ± 12.13
Random	SPH	78.89 ± 8.79	6.66 ± 11.57	1.81 ± 1.88	64.74 ± 13.49
Random	WCH	79.77 ± 8.01	7.58 ± 3.95	2.38 ± 1.15	67.59 ± 11.83
Adversarial [33]	SCH	67.67 ± 7.20	19.85 ± 9.63	3.12 ± 2.31	59.45 ± 10.47
Adversarial [33]	APH	82.47 ± 7.23	7.16 ± 6.96	2.03 ± 1.94	72.03 ± 11.98
Adversarial [33]	SPH	79.07 ± 7.79	9.08 ± 12.71	2.42 ± 2.51	66.16 ± 11.53
Adversarial [33]	WCH	81.18 ± 8.29	8.83 ± 10.41	2.83 ± 3.56	68.35 ± 11.47
Entropy [34]	SCH	67.78 ± 6.88	17.68 ± 8.55	4.77 ± 2.03	57.72 ± 10.07
Entropy [34]	APH	83.24 ± 7.01	7.41 ± 8.46	2.15 ± 2.80	72.81 ± 12.61
Entropy [34]	SPH	79.11 ± 8.82	6.99 ± 11.75	1.85 ± 1.91	65.60 ± 13.67
Entropy [34]	WCH	81.88 ± 7.38	6.03 ± 6.68	1.99 ± 2.74	69.74 ± 11.37
AADA [60]	SCH	67.96 ± 5.15	13.86 ± 9.73	2.53 ± 1.67	59.20 ± 8.29
AADA [60]	APH	83.21 ± 5.83	6.31 ± 7.03	1.83 ± 2.18	72.16 ± 11.71
AADA [60]	SPH	79.77 ± 7.78	6.50 ± 5.79	1.77 ± 1.36	65.95 ± 12.99
AADA [60]	WCH	81.73 ± 8.35	5.70 ± 3.63	1.74 ± 1.44	68.89 ± 12.06
MHPL [61]	SCH	68.71 ± 7.59	12.15 ± 5.66	2.61 ± 1.85	60.26 ± 9.94
MHPL [61]	APH	83.49 ± 8.94	5.65 ± 5.09	1.70 ± 1.79	73.79 ± 15.66
MHPL [61]	SPH	78.55 ± 7.43	5.86 ± 4.55	1.73 ± 1.12	65.59 ± 11.97
MHPL [61]	WCH	82.92 ± 6.57	5.69 ± 3.54	1.61 ± 0.93	71.26 ± 12.83
STDR	SCH	**70.99** $\pm 6.46*$	**10.72** $\pm 3.37*$	**2.47** ± 1.07	**63.10** $\pm 8.84*$
STDR	APH	**85.49** $\pm 6.09*$	**4.63** $\pm 3.06*$	**1.24** $\pm 1.02*$	**76.01** $\pm 12.98*$
STDR	SPH	**80.00** ± 8.10	**4.49** $\pm 2.64*$	**1.64** ± 1.46	**66.36** ± 13.15
STDR	WCH	**84.11** $\pm 6.64*$	**4.95** $\pm 2.75*$	**1.35** $\pm 0.88*$	**73.63** $\pm 11.46*$

For equitable comparisons, the identical experimental framework is utilized across all comparative experiments. Each experiment involves selecting only 20% active samples for labeling and training. Meanwhile, we further used a new Normalized Surface Dice (NSD) metric with a margin of 1 mm to evaluate the segmentation results. These experimental results in Table VI and Table VII demonstrate that the devised STDR strategy provides the best segmentation performance. This proves the superiority of the proposed strategy and the effectiveness of considering domain-invariant and domain-specific representations in sample selection.

C. Ablation Studies

To ascertain the effectiveness of the key components within our proposed methodology, we perform ablation experiments,

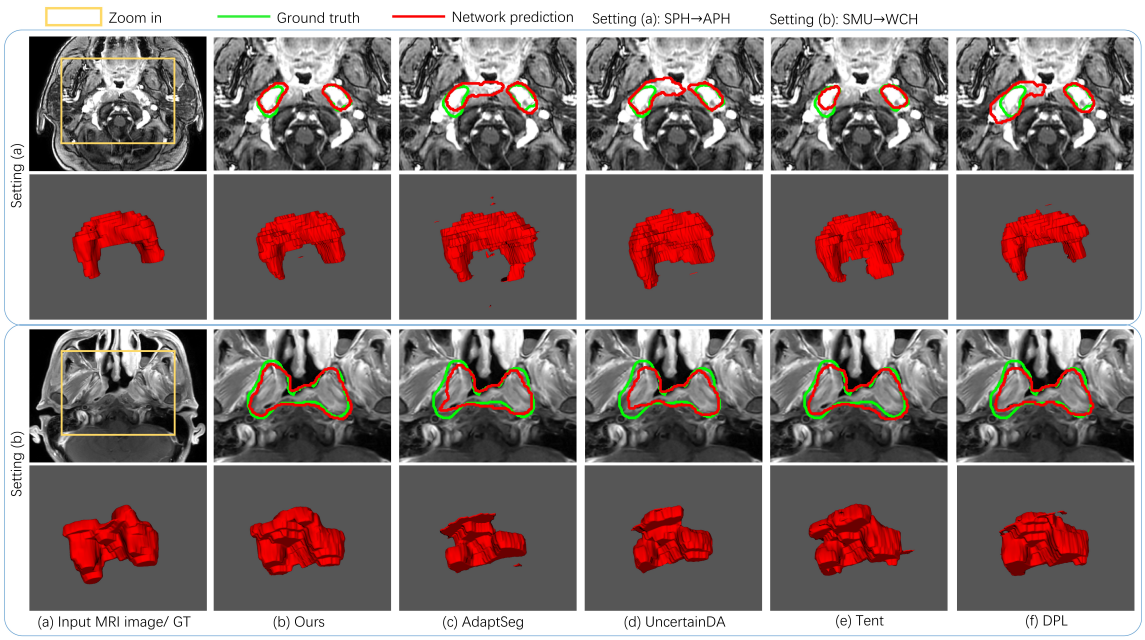

Fig. 6. Visual comparisons of (b) our method with other SOTA methods. The first and third rows showcase 2D slice visualizations, while the second and fourth rows offer 3D comparisons. Contours exhibit markedly improved alignment with ground truth both at the 2D and 3D levels following the application of our method, as compared to other methods.

TABLE VIII
RESULTS OF ABLATION EXPERIMENTS FROM SPH DATASET TO OTHER DATASETS.

Model	Target	DSC (%)	HD95 (mm)	ASD (mm)
STDR-α	SCH	62.66 ± 11.33	16.48 ± 16.03	4.54 ± 4.02
STDR-α	APH	81.46 ± 8.97	$5.64 \pm .344$	1.57 ± 1.19
STDR-α	WCH	79.51 ± 9.11	5.63 ± 3.23	1.58 ± 0.93
STDR-α	SMU	77.03 ± 13.11	8.45 ± 16.58	1.23 ± 1.34
STDR-β	SCH	69.28 ± 7.78	14.93 ± 9.12	3.43 ± 2.71
STDR-β	APH	84.38 ± 5.41	6.52 ± 7.48	1.89 ± 1.88
STDR-β	WCH	81.79 ± 7.31	6.06 ± 3.89	1.77 ± 0.97
STDR-β	SMU	82.66 ± 9.14	6.69 ± 15.83	1.17 ± 1.41
STDR	SCH	69.55 ± 6.82	10.64 ± 3.36	2.83 ± 1.83
STDR	APH	85.05 ± 5.25	4.98 ± 3.07	1.36 ± 1.01
STDR	WCH	82.81 ± 6.51	5.39 ± 2.77	1.58 ± 0.88
STDR	SMU	83.51 ± 8.05	**5.81** ± 13.10	1.26 ± 1.56
Ours (STDR+Semi)	SCH	**70.99** ± 6.59	**8.57** ± 2.39	**2.03** ± 0.84
Ours (STDR+Semi)	APH	**85.69** ± 4.54	**4.58** ± 2.24	**1.18** ± 0.78
Ours (STDR+Semi)	WCH	**83.59** ± 6.76	**4.85** ± 2.79	**1.37** ± 0.92
Ours (STDR+Semi)	SMU	**84.13** ± 7.65	6.17 ± 15.33	**0.86** ± 0.95

TABLE IX
RESULTS OF ABLATION EXPERIMENTS FROM SMU DATASET TO OTHER DATASETS.

Model	Target	DSC (%)	HD95 (mm)	ASD (mm)
STDR-α	SCH	65.64 ± 7.84	16.02 ± 9.75	2.98 ± 1.55
STDR-α	APH	81.57 ± 7.89	6.78 ± 4.65	1.94 ± 1.16
STDR-α	SPH	76.80 ± 8.51	5.29 ± 2.56	1.67 ± 0.99
STDR-α	WCH	79.44 ± 8.27	6.48 ± 4.97	1.68 ± 1.12
STDR-β	SCH	68.21 ± 7.16	10.58 ± 3.35	2.99 ± 1.67
STDR-β	APH	83.57 ± 6.57	6.17 ± 8.25	1.93 ± 2.63
STDR-β	SPH	79.86 ± 7.97	5.42 ± 4.92	1.48 ± 1.21
STDR-β	WCH	83.55 ± 6.69	6.53 ± 5.33	1.70 ± 1.14
STDR	SCH	70.99 ± 6.46	10.72 ± 3.37	2.47 ± 1.07
STDR	APH	85.49 ± 6.09	4.63 ± 3.06	1.24 ± 1.02
STDR	SPH	80.00 ± 8.10	4.49 ± 2.64	1.64 ± 1.46
STDR	WCH	84.11 ± 6.64	4.95 ± 2.75	1.35 ± 0.88
Ours (STDR+Semi)	SCH	**72.58** ± 6.72	**8.06** ± 2.06	**2.00** ± 0.89
Ours (STDR+Semi)	APH	**86.01** ± 5.10	**4.38** ± 2.51	**1.21** ± 1.04
Ours (STDR+Semi)	SPH	**80.97** ± 7.24	**4.02** ± 1.62	**1.03** ± 0.58
Ours (STDR+Semi)	WCH	**84.71** ± 6.04	**4.48** ± 2.23	**1.24** ± 0.77

leading to the identification of four distinct configurations: (i) STDR-α: we select the 20% samples with the largest $Similarity$ in Eq. 6 and focus on domain-invariant representations; (ii) STDR-β: we choose the 20% samples with the smallest $Similarity$ in Eq. 6 and focus on domain-specific representations; (iii) STDR: we follow Eq. 7 to take both into consideration; (iv) Ours (STDR+Semi): we combine the STDR strategy with our semi-supervised pipeline.

As shown in Tables VIII and IX, both STDR-α and STDR-β achieve good performance in the two adaptation settings, and in comparison, STDR-β outperform, which illustrates that representative samples of the target domain are necessary to significantly improve the performance. Nevertheless, when comparing STDR-β with STDR, the experimental outcomes consistently indicate that STDR outperforms overall (p-value < 0.05). This suggests that, for further enhancing model performance, relying solely on domain-specific representative samples is insufficient. It's essential to incorporate domain-invariant samples, enabling more effective fine-tuning of model parameters, preservation of shared segmentation knowledge, and improvement in model generalization, leading to superior results. This integrated approach is crucial for achieving better outcomes. Finally, it's evident that the performance of Ours (STDR+Semi) consistently surpasses that of STDR across nearly all metrics (p-value < 0.05), providing compelling evidence for the efficacy of integrating the semi-supervised pipeline into the active domain adaptation.

To assess the robustness of the STDR strategy, we randomly selected two adaptation settings and carried out comparative experiments using various proportions of active samples, as shown in Table X. To eliminate any potential biases arising from unlabeled samples and the usage of semi-supervised learning methods, we exclusively employed actively labeled target samples during the experiment. As depicted in Table X, the model's performance exhibits a consistent improvement with the increase in the proportion of samples. We observe a

TABLE X
EXPERIMENTAL RESULTS WITH DIFFERENT NUMBERS OF ACTIVE SAMPLES.

Adaptation setting	Percent	DSC (%)	HD95 (mm)	ASD (mm)
SPH → SCH	10%	64.61 ± 9.59	11.53 ± 4.21	2.96 ± 1.67
SPH → SCH	20%	69.55 ± 6.82	10.64 ± 3.36	2.83 ± 1.83
SPH → SCH	40%	70.88 ± 6.75	9.96 ± 4.49	2.47 ± 1.21
SPH → SCH	60%	71.73 ± 6.79	9.22 ± 4.28	2.66 ± 1.62
SPH → SCH	80%	73.57 ± 5.59	8.82 ± 4.29	2.27 ± 1.19
SPH → SCH	100%	74.31 ± 5.57	8.57 ± 4.68	2.09 ± 1.01
SPH → APH	10%	84.05 ± 5.60	5.37 ± 4.24	1.48 ± 1.35
SPH → APH	20%	85.05 ± 5.25	4.98 ± 3.07	1.36 ± 1.01
SPH → APH	40%	85.62 ± 4.76	4.98 ± 3.13	1.35 ± 1.07
SPH → APH	60%	86.02 ± 4.39	4.60 ± 2.47	1.33 ± 0.87
SPH → APH	80%	86.81 ± 4.19	4.05 ± 1.99	1.18 ± 0.93
SPH → APH	100%	87.39 ± 4.74	3.97 ± 2.19	0.99 ± 0.75

TABLE XI
ABLATION EXPERIMENTS OF DIFFERENT CLUSTER NUMBERS IN THE STDR METHOD FROM SPH TO APH

Cluster count	DSC (%)	HD95 (mm)	ASD (mm)	NSD (%)
1	83.81 ± 5.93	7.31 ± 8.46	2.06 ± 2.51	72.91 ± 12.31
5	84.41 ± 5.37	5.80 ± 6.67	1.56 ± 1.84	73.84 ± 12.32
10	**85.05** ± 5.25	**4.98** ± 3.07	**1.36** ± 1.01	**74.73** ± 11.77
20	84.71 ± 5.04	5.97 ± 6.57	1.56 ± 1.71	74.42 ± 11.52
30	84.69 ± 5.01	5.43 ± 4.36	1.52 ± 1.26	74.62 ± 11.36

significant performance boost when introducing new actively labeled (AL) samples, especially with a small sample size. However, this performance improvement gradually levels off as the number of samples increases. Thus, our choice of 20% of the samples is a trade-off between labeling cost and segmentation performance. Compared to fully labeling all data, our performance is only 4.76% and 2.34% lower in DSC for SPH → SCH and SPH → APH, respectively.

We have also investigated the impact of the number of cluster centers on the segmentation performance under the setting of the SPH → APH. We experimented with varying numbers of single and multiple clusters, with the specific outcomes illustrated in Table XI. We observed that a single cluster center did not yield optimal results, possibly due to the complex characteristics of the data distribution, which could be attributed to differences in vendors, magnetic fields, and radiation oncologists. We settled on using 10 clusters, as it demonstrated superior performance in our tests.

V. DISCCUSION

Accurate GTV segmentation is pivotal for effective radiotherapy in patients with NPC [68], [69]. Despite the proliferation of GTV segmentation algorithms [15], [21]–[24], [70], their practical implementation faces a significant challenge: models trained in the source domain experience substantial performance degradation when applied in a new medical center. Combining Tables II and III, it's clear that the algorithms consistently exhibit a reduction of 5% to 10%, or even more than 10%, in terms of DSC. There is also a noticeable degradation in performance on HD95 and ASD metrics. For example, in adaptations like SMU → SCH or SPH → SCH, HD95 shows an error variation exceeding $20mm$, indicating significant performance discrepancies. Furthermore, the privacy and security of medical data, coupled with hospital regulations, create significant constraints. Even in cases where certain medical centers possess ample well-labeled data, others can't leverage this resource, primarily due to the prominent domain gap evident between different data domains. This disparity is visually represented in Fig. 3.

TABLE XII
QUANTITATIVE RESULTS OF THE UNIVERSEG MODEL (SUPPORT SIZE = 64) TRAINED AND TESTED ON DIFFERENT DATASETS.

Model	Dataset	DSC (%)	HD95 (mm)	ASD (mm)
UniverSeg [71]	SCH	22.31 ± 12.68	30.58 ± 12.24	12.63 ± 5.64
UniverSeg [71]	APH	43.22 ± 15.08	31.50 ± 16.30	11.01 ± 5.40
UniverSeg [71]	SPH	34.73 ± 15.36	30.79 ± 17.90	11.71 ± 4.32
UniverSeg [71]	WCH	53.55 ± 14.98	29.23 ± 21.05	9.03 ± 6.05
UniverSeg [71]	SMU	19.00 ± 17.43	40.21 ± 17.79	18.81 ± 12.26

These challenges highlight the need for innovative domain adaptation techniques to bridge this gap, and our proposed solution is a Source-Free Active Domain Adaptation. Our approach holds multiple advantages over existing domain adaptation methods. Firstly, it demands only a limited number of reference vectors, bypassing the need for source data access. This not only ensures data privacy and security but also facilitates easy transmission and utilization. Secondly, our approach offers greater clinical utility. As evident in Tables IV and V, previous methods indeed enhance segmentation model performance on the target domain, but they exclusively rely on unlabeled target data. The absence of ground truth for supervised training can introduce the target-domain distribution distortions, resulting in relatively modest improvements, typically in the range of 1-3% for the DSC. Indeed, within a real clinical context, the active labeling of small data sets is both practically feasible and reasonably efficient in terms of time. The major significance of our approach becomes apparent as its ability to harness this limited data, enabling us to achieve a performance level comparable to that of a fully supervised-trained model specific to the target domain. This underscores the tremendous value our method contributes to the clinical deployment of computer-aid-segmentation networks. For instance, combining Tables II, IV, and V, we can find that in the SPH → SCH, our method's performance (70.99%, $8.57mm$, $2.03mm$ for DSC, HD95, ASD) is comparable with the results achieved by directly training on SCH (72.19%, $9.25mm$, $2.22mm$ for DSC, HD95, ASD), albeit with a slightly lower DSC and superior HD95 and ASD values. Meanwhile, in the SMU → SCH setting, our method (72.58%, $8.06mm$, $2.00mm$ for DSC, HD95, ASD) clearly outperforms.

Except for learning-based SFDA strategies, we also conducted experiments to explore classic intensity matching techniques (HistMatch) for source-free domain adaptation. Specifically, we followed previous works [49], [50] that employed HistMatch to adjust the intensity distribution of the target dataset to match the source dataset. The results of this exploration are detailed in Tables IV and V, showing that most metrics (such as HD95, ASD) experienced an improvement. However, DSC showed a decrease, possibly due to the loss or distortion of information during HistMatch, even as intensities align more closely. Different metrics measure different aspects of performance, which explains the variation in outcomes. In addition, given the increasing research enthusiasm for the foundation models recently [72], we also investigated the performance of representative work, UniverSeg [71] (with the maximum support size = 64) on the real multi-center clinical

TABLE XIII
EXPERIMENTAL RESULTS ON MR IMAGES OF OUR METHOD ON DIFFERENT MAGNET STRENGTHS.

Adaptation setting	Magnetic field	DSC (%)	HD95 (mm)	ASD (mm)
SPH → SCH	1.5T (5)	69.57 ± 7.75	8.03 ± 1.18	2.10 ± 0.89
SPH → SCH	3T (6)	72.19 ± 5.15	9.03 ± 2.97	1.97 ± 0.80
SPH → SCH	all (11)	70.99 ± 6.59	8.57 ± 2.39	2.03 ± 0.84
SPH → APH	1.5T(24)	85.49 ± 4.50	4.66 ± 2.36	1.23 ± 0.79
SPH → APH	3T (6)	86.47 ± 4.61	4.27 ± 1.63	0.99 ± 0.68
SPH → APH	all (30)	85.69 ± 4.54	4.58 ± 2.24	1.18 ± 0.78

dataset. The quantitative results are presented in Table XII, which can be found that the current performance of the model in the field of NPC tumor segmentation still needs to be further improved, proving the practical value of source-free active domain adaptation.

We performed a preliminary analysis of our method's performance across different magnetic field strengths, as presented in Table XIII. Our findings suggest that segmentation results on 3T strength tend to be superior to those on 1.5T strength. This improvement is likely attributable to the higher signal-to-noise ratio (SNR) and overall better image quality provided by 3T MR images [73], [74]. These attributes of 3T MR images facilitate the deep learning model to segment the targets more accurately [75].

This work also has some limitations in the aspects of applications and technical. Both the development and validation of our deep learning model are conducted exclusively using MRI data due to its superior soft-tissue contrast. While CT is more commonly employed in treatment planning, it's common practice to generate GTVs on MRI and subsequently map them to planning CT through image fusion. However, it would be more clinically pertinent if validation could encompass both MRI and CT datasets. Additionally, using Euclidean distance as a quantization metric performs well with low-dimensional features in the K-Means clustering and high-dimensional representations, it may lead to biased results due to one or a few dominant features [76]. Although our feature representation is 1×256, which is not high compared to our majority of data centers (except for SCH) with thousands of MRI image samples [77], there might still be room for optimization in this aspect by employing other metrics. Besides, our semi-supervised learning relies on the accuracy of pseudo-labels generated after the first training phase. While there is potential for errors in these labels, our method still demonstrated performance improvements, as seen in Tables VIII and IX. However, we currently do not explore measures to handle incorrect labels, which could limit performance. In addition, this work was only evaluated on the NPC GTV segmentation task and lacked the robustness and generalization evaluation to other tasks, it may potentially be adapted for various segmentation tasks. Another point is that although our method eliminates the need to access the source data directly, thereby effectively protecting the privacy and security of the source domain data (so-called source-free), it does require the transmission of a limited number of reference vectors. Addressing the above limitations will be a focus in our future work to enhance the robustness and effectiveness.

Another substantial hurdle in this research realm is the scarcity of suitably annotated multi-center GTV datasets. To the best of our knowledge, a major portion of MRI-based GTV segmentation relies on proprietary datasets, and the availability of multi-center datasets remains limited. In response to this challenge, we've taken proactive measures to curate meticulously annotated multi-center GTV datasets, collaborating closely with radiation oncologists. The newly proposed datasets can align with domain adaptation research analytical and practical prerequisites. Importantly, we will release an open-source NPC GTV dataset with more than 150 patients from multiple hospitals after the peer review, thereby fostering further progress in the field and encouraging active participation from fellow researchers.

VI. CONCLUSION

In this paper, we introduce a novel Source-Free Active Domain Adaptation for GTV segmentation in cross-center NPC, rigorously validated with data from 1,057 patients at five medical centers. We devise an STDR strategy to discern representative samples of domain-invariant and domain-specific in the target medical center. This selection process hinges on the embedding distance within the latent space, and training with these actively labeled samples significantly improves the model's performance on the target domain. Furthermore, we develop a semi-supervised learning process that combines it with active domain adaptation to fully utilize the remaining unlabeled samples to enhance segmentation. Numerous experimental results show that our network outperforms SOTA methods. Simultaneously, our approach achieves performance near fully supervised training with minimal labeling, showcasing significant clinical medical utility value. In the future, we plan to consider validating our method on other segmentation tasks (such as different imaging modalities and other tumors). We will also focus on strategies to detect and correct pseudo-label inaccuracies, enhancing our semi-supervised learning.

REFERENCES

[1] M. L. Chua, J. T. Wee *et al.*, "Nasopharyngeal carcinoma," *The Lancet*, vol. 387, no. 10022, pp. 1012–1024, 2016.
[2] N. Lee, P. Xia, J. M. Quivey, K. Sultanem, I. Poon, C. Akazawa, P. Akazawa, V. Weinberg, and K. K. Fu, "Intensity-modulated radiotherapy in the treatment of nasopharyngeal carcinoma: an update of the ucsf experience," *IJROBP*, vol. 53, no. 1, pp. 12–22, 2002.
[3] X. Luo, W. Liao, Y. He *et al.*, "Deep learning-based accurate delineation of primary gross tumor volume of nasopharyngeal carcinoma on heterogeneous magnetic resonance imaging: A large-scale and multi-center study," *Radiotherapy and Oncology*, vol. 180, p. 109480, 2023.
[4] Y.-P. Chen, A. T. Chan, Q.-T. Le *et al.*, "Nasopharyngeal carcinoma," *The Lancet*, vol. 394, no. 10192, pp. 64–80, 2019.
[5] M. K. Kam, P. M. Teo, R. M. Chau *et al.*, "Treatment of nasopharyngeal carcinoma with intensity-modulated radiotherapy: the hong kong experience," *IJROBP*, vol. 60, no. 5, pp. 1440–1450, 2004.
[6] A. A. K. A. Razek and A. King, "Mri and ct of nasopharyngeal carcinoma," *AJOR*, vol. 198, no. 1, pp. 11–18, 2012.
[7] X. Luo, J. Fu, Y. Zhong, S. Liu, B. Han, M. Astaraki, S. Bendazzoli, I. Toma-Dasu, Y. Ye, Z. Chen *et al.*, "Segrap2023: A benchmark of organs-at-risk and gross tumor volume segmentation for radiotherapy planning of nasopharyngeal carcinoma," *arXiv preprint arXiv:2312.09576*, 2023.
[8] X. Bai, Y. Hu, G. Gong, Y. Yin, and Y. Xia, "A deep learning approach to segmentation of nasopharyngeal carcinoma using computed tomography," *Biomedical Signal Processing and Control*, vol. 64, p. 102246, 2021.

[9] Y. Zhang, X. Ye, J. Ge, D. Guo, D. Zheng, H. Yu, Y. Chen, G. Yao, Z. Lu, A. Yuille *et al.*, "Deep learning-based multi-modality segmentation of primary gross tumor volume in ct and mri for nasopharyngeal carcinoma," *International Journal of Radiation Oncology, Biology, Physics*, vol. 117, no. 2, p. e498, 2023.

[10] L. Lin, Q. Dou, Y.-M. Jin *et al.*, "Deep learning for automated contouring of primary tumor volumes by mri for nasopharyngeal carcinoma," *Radiology*, vol. 291, no. 3, pp. 677–686, 2019.

[11] S. Chen, D. Yang, X. Liao *et al.*, "Failure patterns of recurrence and metastasis after intensity-modulated radiotherapy in patients with nasopharyngeal carcinoma: results of a multicentric clinical study," *Frontiers in Oncology*, vol. 11, p. 5730, 2022.

[12] S. Wang, M. Liu, J. Lian, and D. Shen, "Boundary coding representation for organ segmentation in prostate cancer radiotherapy," *TMI*, vol. 40, no. 1, pp. 310–320, 2020.

[13] D. Jin, D. Guo *et al.*, "Deeptarget: Gross tumor and clinical target volume segmentation in esophageal cancer radiotherapy," *MedIA*, vol. 68, p. 101909, 2021.

[14] M. Tian, H. Wang, X. Liu *et al.*, "Delineation of clinical target volume and organs at risk in cervical cancer radiotherapy by deep learning networks," *Medical Physics*, 2023.

[15] Y. Li, T. Dan *et al.*, "Npcnet: jointly segment primary nasopharyngeal carcinoma tumors and metastatic lymph nodes in mr images," *TMI*, vol. 41, no. 7, pp. 1639–1650, 2022.

[16] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[17] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," *ICLR*, 2021.

[18] M. Bateson, H. Kervadec, J. Dolz, H. Lombaert, and I. B. Ayed, "Source-free domain adaptation for image segmentation," *Medical Image Analysis*, vol. 82, p. 102617, 2022.

[19] H. Wang, S. Zhang, X. Luo, W. Liao, and L. Zhu, "Advancing delineation of gross tumor volume based on magnetic resonance imaging by performing source-free domain adaptation in nasopharyngeal carcinoma," in *International Workshop on Computational Mathematics Modeling in Cancer Analysis*. Springer, 2023, pp. 71–80.

[20] Z. Li, C. Li, X. Luo, Y. Zhou, J. Zhu, C. Xu, M. Yang, Y. Wu, and Y. Chen, "Towards source-free cross tissues histopathological cell segmentation via target-specific finetuning," *IEEE Transactions on Medical Imaging*, 2023.

[21] J. Zhou, K. L. Chan, P. Xu, and V. F. Chong, "Nasopharyngeal carcinoma lesion segmentation from mr images by support vector machine," in *ISBI*, 2006, pp. 1364–1367.

[22] W. Huang, K. L. Chan, and J. Zhou, "Region-based nasopharyngeal carcinoma lesion segmentation from mri using clustering-and classification-based methods with learning," *JDI*, vol. 26, pp. 472–482, 2013.

[23] K. Men, X. Chen *et al.*, "Deep deconvolutional neural network for target segmentation of nasopharyngeal cancer in planning computed tomography images," *Frontiers in oncology*, vol. 7, p. 315, 2017.

[24] H. Chen, Y. Qi, Y. Yin *et al.*, "Mmfnet: A multi-modality mri fusion network for segmentation of nasopharyngeal carcinoma," *Neurocomputing*, vol. 394, pp. 27–40, 2020.

[25] W. Liao, J. He *et al.*, "Automatic delineation of gross tumor volume based on magnetic resonance imaging by performing a novel semisupervised learning framework in nasopharyngeal carcinoma," *IJROBP*, vol. 113, no. 4, pp. 893–902, 2022.

[26] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *TIST*, vol. 11, no. 5, pp. 1–46, 2020.

[27] X. Liu, C. Yoo, F. Xing, H. Oh, G. El Fakhri, J.-W. Kang, J. Woo *et al.*, "Deep unsupervised domain adaptation: A review of recent advances and perspectives," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[28] X. Liu, Y. Han, S. Bai, Y. Ge, T. Wang, X. Han, S. Li, J. You, and J. Lu, "Importance-aware semantic segmentation in self-driving with discrete wasserstein training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 629–11 636.

[29] X. Liu, F. Xing, G. El Fakhri, and J. Woo, "Memory consistent unsupervised off-the-shelf model adaptation for source-relaxed medical image segmentation," *Medical image analysis*, vol. 83, p. 102641, 2023.

[30] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*. Pmlr, 2018, pp. 1989–1998.

[31] K. Mei, C. Zhu, J. Zou, and S. Zhang, "Instance adaptive self-training for unsupervised domain adaptation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*. Springer, 2020, pp. 415–430.

[32] Q. Lian, F. Lv, L. Duan, and B. Gong, "Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6758–6767.

[33] Y.-H. Tsai, W.-C. Hung *et al.*, "Learning to adapt structured output space for semantic segmentation," in *CVPR*, 2018, pp. 7472–7481.

[34] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *CVPR*, 2019, pp. 2517–2526.

[35] L. Tang, K. Li, C. He, Y. Zhang, and X. Li, "Source-free domain adaptive fundus image segmentation with class-balanced mean teacher," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 684–694.

[36] Z. Huai, X. Ding, Y. Li, and X. Li, "Context-aware pseudo-label refinement for source-free domain adaptive fundus image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 618–628.

[37] T. Yao, Y. Pan, C.-W. Ngo, H. Li, and T. Mei, "Semi-supervised domain adaptation with subspace learning for visual recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2015, pp. 2142–2150.

[38] L. A. Pereira and R. da Silva Torres, "Semi-supervised transfer subspace for domain adaptation," *Pattern Recognition*, vol. 75, pp. 235–249, 2018.

[39] Z. Fang, J. Lu, F. Liu, and G. Zhang, "Semi-supervised heterogeneous domain adaptation: Theory and algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 1087–1105, 2022.

[40] T. Kim and C. Kim, "Attract, perturb, and explore: Learning a feature alignment network for semi-supervised domain adaptation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 591–607.

[41] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko, "Semi-supervised domain adaptation via minimax entropy," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8050–8058.

[42] G. He, X. Liu, F. Fan, and J. You, "Classification-aware semi-supervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 964–965.

[43] Z. Yu, J. Li, Z. Du, L. Zhu, and H. T. Shen, "A comprehensive survey on source-free domain adaptation," *arXiv preprint arXiv:2302.11803*, 2023.

[44] H. Guan and M. Liu, "Domain adaptation for media: a survey," *TBME*, vol. 69, no. 3, pp. 1173–1185, 2021.

[45] F. Fleuret *et al.*, "Uncertainty reduction for model adaptation in semantic segmentation," in *CVPR*, 2021, pp. 9613–9623.

[46] C. Chen, Q. Liu, Y. Jin, Q. Dou, and P.-A. Heng, "Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling," in *MICCAI*. Springer, 2021, pp. 225–235.

[47] C. Yang, X. Guo, Z. Chen, and Y. Yuan, "Source free domain adaptation for medical image segmentation with fourier style mining," *Medical Image Analysis*, vol. 79, p. 102457, 2022.

[48] S. Hu, Z. Liao, and Y. Xia, "Prosfda: Prompt learning based source-free domain adaptation for medical image segmentation," *arXiv preprint arXiv:2211.11514*, 2022.

[49] L. G. Nyúl and J. K. Udupa, "On standardizing the mr image intensity scale," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 42, no. 6, pp. 1072–1081, 1999.

[50] K. Kushibar, S. Valverde, S. Gonzalez-Villa, J. Bernal, M. Cabezas, A. Oliver, and X. Llado, "Supervised domain adaptation for automatic sub-cortical brain structure segmentation with minimal user interaction," *Scientific reports*, vol. 9, no. 1, p. 6742, 2019.

[51] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *JAIR*, vol. 4, pp. 129–145, 1996.

[52] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *ML*. Elsevier, 1994, pp. 148–156.

[53] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden markov models for information extraction," in *ISIDA*. Springer, 2001, pp. 309–318.

[54] S. D. Jain and K. Grauman, "Active image segmentation propagation," in *CVPR*, 2016, pp. 2864–2873.

[55] S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Semisupervised svm batch mode active learning with applications to image retrieval," *TOIS*, vol. 27, no. 3, pp. 1–29, 2009.

[56] Z. Xu, K. Yu, V. Tresp, X. Xu, and J. Wang, "Representative sampling for text classification using support vector machines," in *ECIR*. Springer, 2003, pp. 393–407.

[57] S.-J. Huang, R. Jin, and Z.-H. Zhou, "Active learning by querying informative and representative examples," *NeurIPS*, vol. 23, 2010.
[58] A. Freytag, E. Rodner, and J. Denzler, "Selecting influential examples: Active learning with expected model output changes," in *ECCV*. Springer, 2014, pp. 562–577.
[59] C. Kading, A. Freytag, E. Rodner, P. Bodesheim, and J. Denzler, "Active learning and discovery of object categories in the presence of unnameable instances," in *CVPR*, 2015, pp. 4343–4352.
[60] J.-C. Su, Y.-H. Tsai, K. Sohn, B. Liu, S. Maji, and M. Chandraker, "Active adversarial domain adaptation," in *WACV*, 2020, pp. 739–748.
[61] F. Wang, Z. Han, Z. Zhang, and Y. Yin, "Active source free domain adaptation," *arXiv preprint arXiv:2205.10711*, 2022.
[62] G. E. Hinton and S. Roweis, "Stochastic neighbor embedding," *NeurIPS*, vol. 15, 2002.
[63] J. Lever, M. Krzywinski, and N. Altman, "Points of significance: model selection and overfitting," *Nature methods*, vol. 13, no. 9, pp. 703–705, 2016.
[64] A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra, "Grokking: Generalization beyond overfitting on small algorithmic datasets," *arXiv preprint arXiv:2201.02177*, 2022.
[65] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.
[66] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.
[67] G. Wang, X. Luo, R. Gu, S. Yang, Y. Qu, S. Zhai, Q. Zhao, K. Li, and S. Zhang, "Pymic: A deep learning toolkit for annotation-efficient medical image segmentation," *CMPB*, vol. 231, p. 107398, 2023.
[68] P. Xia, K. K. Fu *et al.*, "Comparison of treatment plans involving intensity-modulated radiotherapy for nasopharyngeal carcinoma," *IJROBP*, vol. 48, no. 2, pp. 329–337, 2000.
[69] W. T. Ng, J. C. Chow *et al.*, "Current radiotherapy considerations for nasopharyngeal carcinoma," *Cancers*, vol. 14, no. 23, p. 5773, 2022.
[70] Y. Guo, Q. Yang, W. Hu, Z. Zhang, J. Wang, and C. Hu, "Automatic segmentation of nasopharyngeal carcinoma on mr images: A single-institution experience," *IJROBP*, vol. 108, no. 3, p. e776, 2020.
[71] V. I. Butoi, J. J. G. Ortiz, T. Ma, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Universeg: Universal medical image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 438–21 451.
[72] S. Zhang and D. Metaxas, "On the challenges and perspectives of foundation models for medical image analysis," *Medical Image Analysis*, p. 102996, 2023.
[73] N. Saupe, K. P. Prussmann, R. Luechinger, P. Bosiger, B. Marincek, and D. Weishaupt, "Mr imaging of the wrist: comparison between 1.5-and 3-t mr imaging—preliminary experience," *Radiology*, vol. 234, no. 1, pp. 256–264, 2005.
[74] T. Ullrich, M. Quentin, C. Oelers, F. Dietzel, L. Sawicki, C. Arsov, R. Rabenalt, P. Albers, G. Antoch, D. Blondin *et al.*, "Magnetic resonance imaging of the prostate at 1.5 versus 3.0 t: A prospective comparison study of image quality," *European journal of radiology*, vol. 90, pp. 192–197, 2017.
[75] R. Chu, S. Hurwitz, S. Tauhid, and R. Bakshi, "Automated segmentation of cerebral deep gray matter from mri scans: effect of field strength on sensitivity and reliability," *BMC neurology*, vol. 17, pp. 1–10, 2017.
[76] A. S. Shirkhorshidi, S. Aghabozorgi, and T. Y. Wah, "A comparison study on similarity and dissimilarity measures in clustering continuous data," *PloS one*, vol. 10, no. 12, p. e0144059, 2015.
[77] C. Giraud, *Introduction to high-dimensional statistics*. CRC Press, 2021.