# SDC-UDA: Volumetric Unsupervised Domain Adaptation Framework for Slice-Direction Continuous Cross-Modality Medical Image Segmentation

Hyungseob Shin[1*] Hyeongyu Kim[1*] Sewon Kim[4,5] Yohan Jun[7,8] Taejoon Eo[1,6]
Dosik Hwang[1,2,3,9†,]

{[1]School of Electrical and Electronic Engineering, [2]Department of Oral and Maxillofacial Radiology, College of Dentistry, [3]Department of Radiology and Center for Clinical Imaging Data Science, College of Medicine}
@Yonsei University [4]Naver AI Lab [5]Naver Cloud [6]Probe Medical, Inc. [7]Martinos Center for Biomedical Imaging
[8]Harvard Medical School [9]Center for Healthcare Robotics, Korea Institute of Science and Technology
{whatzupsup, lion4309, dosik.hwang}@yonsei.ac.kr

## Abstract

*Recent advances in deep learning-based medical image segmentation studies achieve nearly human-level performance in fully supervised manner. However, acquiring pixel-level expert annotations is extremely expensive and laborious in medical imaging fields. Unsupervised domain adaptation (UDA) can alleviate this problem, which makes it possible to use annotated data in one imaging modality to train a network that can successfully perform segmentation on target imaging modality with no labels. In this work, we propose SDC-UDA, a simple yet effective volumetric UDA framework for **S**lice-**D**irection **C**ontinuous cross-modality medical image segmentation which combines intra- and inter-slice self-attentive image translation, uncertainty-constrained pseudo-label refinement, and volumetric self-training. Our method is distinguished from previous methods on UDA for medical image segmentation in that it can obtain continuous segmentation in the slice direction, thereby ensuring higher accuracy and potential in clinical practice. We validate SDC-UDA with multiple publicly available cross-modality medical image segmentation datasets and achieve state-of-the-art segmentation performance, not to mention the superior slice-direction continuity of prediction compared to previous studies.*

## 1. Introduction

With the surprising development of deep learning (DL), many studies are now showing remarkable performance in various applications [8,16,20]. However, when a DL model faces data from an unseen domain, performance degradation occurs [9,32]. Resolving this issue is important for the DL techniques to be applied in real world since collecting data from all domains and labeling them is very impractical and inefficient. Unsupervised Domain Adaptation (UDA) aims to alleviate this problem by adapting a model trained on source domain data to target domain, without the necessity of supervision in the target domain. Data dependency is more serious in medical image segmentation field since acquiring pixel-level expert annotation is extremely expensive and time-consuming [3,22,37].

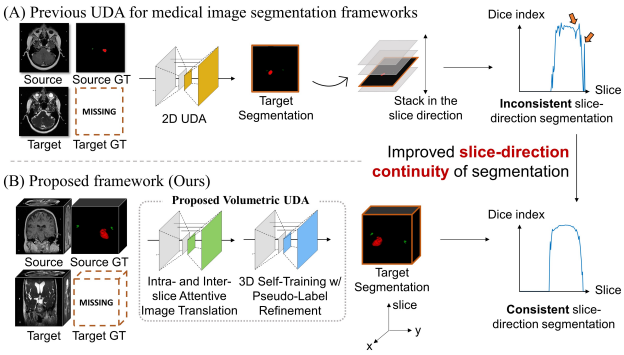


Figure 1. An illustration that describes the comparison between our proposed method with previous methods. (A) Previous UDA for medical image segmentation studies mostly utilize 2D UDA, which leads to inconsistent predictions in the slice direction when the predictions are stacked. (B) The proposed framework (**SDC-UDA**) considers volumetric information in the translation and segmentation process, respectively, which leads to improved slice-direction continuity of segmentation that is much practical for clinical use.

Previous studies on UDA in the field of cross-modality med-

[*] Equal contribution. [†] Corresponding author.

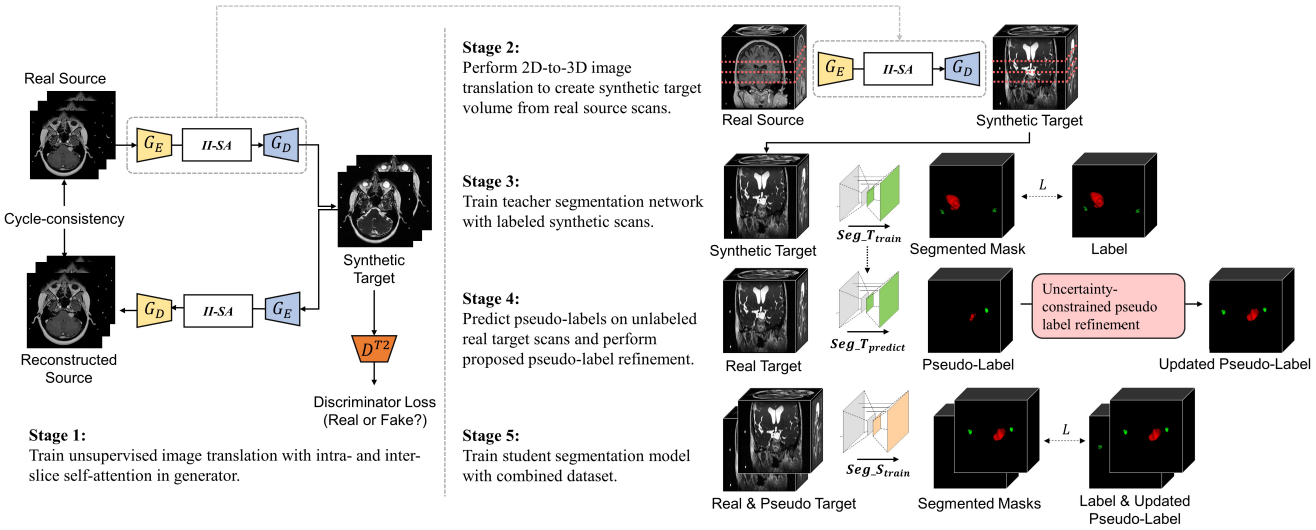


Figure 2. Overview of our volumetric UDA framework. First, source-to-target image transformation is performed via unpaired image translation with intra- and inter-slice self-attention (stage 1-2). Second, volumetric self-training is performed (stage 3-5). During self-training, uncertainty-constrained pseudo-label refinement is conducted to improve pseudo-labels, thereby maximizing the effect of self-training (stage 4). The reverse loop of image translation is omitted for ease of illustration. Detailed architecture of image translation network is described in Fig. 3. Best viewed in color and on high-resolution display. *II-SA*: Intra- and inter-slice self-attention.

ical image segmentation are normally conducted by simultaneously learning 2D image translation and target-domain segmentation, inferring the trained model on each slice of target data, and then stacking the predictions in the slice direction (Fig. 1). As a result, they can lead to inconsistent or fluctuating predictions in the slice direction that are not revealed in quantitative metrics and may interrupt accurate analysis of target structure, making it difficult to use in real world clinical practice. In contrast, SDC-UDA can achieve **improved slice-direction continuity by incorporating volumetric natures of medical imaging**, which has not been thoroughly explored in previous works. It efficiently generates synthetic target volumes with neighbor-aware image translation by utilizing intra- and inter-slice self-attention module [27]. Then, volumetric self-training is followed with uncertainty-constrained pseudo-label refinement strategy that adaptively increases the accuracy of pseudo-labels according to the target data (Fig. 2). **Preliminary version of this work has won the 1st place in an unsupervised cross-modality domain adaptation for medical image segmentation challenge** [7, 26]. We updated the framework and extended it to multiple datasets. The main contribution of our work can be summarized as follows:

- We present **SDC-UDA**, a unified **volumetric UDA framework for cross-modality medical image segmentation**.
- Intra- and inter-slice self-attention for efficient medical image translation: Proposed 2.5D translation framework with intra- and inter-slice self-attention module leads to **increased anatomy preservation and slice-direction smoothness in the synthesized volume**, enabling the synthetic volume to be used effectively in the following self-training steps.
- Volumetric self-training with uncertainty-constrained pseudo-label refinement: We propose a novel uncertainty-constrained pseudo-label refinement module that **can adaptively enhance the accuracy (i.e., sensitivity or specificity) of pseudo-labels**, thereby maximizing the performance of self-training on medical image segmentation.
- SDC-UDA was **validated on multiple public datasets with different data characteristics** for cross-modality medical image segmentation. It not only surpassed the performance of previous methods, but also showed **superior slice-direction segmentation continuity** which can provide precise analysis in clinical practice.

## 2. Related Work

### 2.1. UDA for medical image segmentation.

UDA for semantic segmentation is one of the most popular themes among UDA-related studies [12, 18, 21, 31]. Especially, UDA on medical image segmentation is very attractive [3, 6, 13, 15, 22, 37] since it can address the difficulty of obtaining expensive expert-level manual annota-

tions. Recent studies on UDA for medical image segmentation are mostly based on image adaptation in which domain translation and segmentation networks are trained end-to-end, thereby utilizing the source-to-target transformed images for training segmentation on target domain. [13, 22, 37] combined image-to-image translation and segmentation into a single network. [3] added feature adaptation with adversarial training using the segmentation output of the synthetic and real target domain images to better preserve the geometry of the structures-of-interest. Moreover, [15] added both the synthetic and real target domain images and their corresponding segmentation probability maps to adversarial training to preserve not only the geometry of target anatomies but also their appearance (i.e., intensities). [11] utilized deep symmetric networks to further align the features of two domains.

Our work differs from previous works in the following aspects: Previous works mostly train 2D-level image translation with target-domain segmentation and stack the segmentation of each target slice in the slice direction [4, 11, 15, 30]. Since the segmentation network does not consider the anatomical structure in the slice direction, this can lead to high variability of segmentation result even in adjacent slices. In contrast, SDC-UDA splits target-domain segmentation from image translation and therefore is not constrained to 2D framework. Consequently, SDC-UDA can effectively incorporate volumetric information in the translation and segmentation process, respectively. Moreover, in addition to the commonly used cross-modality medical image segmentation dataset [17, 19, 40], SDC-UDA has also been validated in the challenging task of cross-modality segmentation of small multi-class structures which are vestibular schwannoma (VS) and cochlea (i.e., CrossMoDA dataset [7, 24, 25]). Previous studies locate target organs to the center of the preprocessed images and the majority of slices contain at least one target organ, whereas VS and cochleas in CrossMoDA dataset occupy extremely small fraction of the total voxels (0.028% and 0.002% for VS and cochlea, respectively, compared to approximately 3% in average in another dataset [40]) which more reflects the real-world clinical imaging environment [4, 15].

### 2.2. Self-training on UDA for medical image segmentation.

Self-training belongs to semi-supervised learning which has emerged to improve the resources and cost put into data labeling [35, 41]. In self-training, a teacher model is first trained using only the labeled data. Next, pseudo-labels with high confidence are inferred by passing the unlabeled data on the trained model. With the labeled data and pseudo-labeled data, a larger dataset can be used to train a student model that performs better than the teacher model trained only on labeled data. Numerous attempts were made to apply self-training into semantic segmentation [23, 39, 42, 43] as pixel-level annotations are expensive. Since pseudo-labels tend to be noisy labels, previous studies attempted to refine pseudo-labels to increase accuracy, thereby providing better supervision to unlabeled data. [33] updated pseudo-labels by enhancing the sensitivity of pseudo-labels via selective voting. [29] proposed region growing on the super-pixels where the pseudo-labels serve as the initial seed points. [10] utilized uncertainty-weighted binary cross entropy loss to penalize high uncertainty region.

The proposed pseudo-label refinement strategy is effective in two aspects: 1) sensitivity and specificity-enhancing refinement can be adaptively used according to the characteristic of target data, and 2) it is a safer way to handle uncertainty maps since even the correctly segmented regions can have high uncertainty, and using uncertainty map to weight the loss function in the following training can therefore guide the model to wrong direction.

## 3. Methods

### 3.1. Unpaired image translation with intra- and inter-slice self-attention module.

Many recent works on UDA for medical image segmentation have been developed on 2D framework, without considering the volumetric nature of medical imaging (2D multi-slice sequence or 3D sequence) [3, 13, 15, 22]. Previous 2D UDA approaches split the 3D volume into 2D slices, and re-stack their translations into 3D volume afterward. Since the slices are processed individually, re-building the translated volume usually requires additional post-processing such as slice-direction interpolation, which still can not perfectly resolve problems such as slice-direction discontinuity. This remain a problem when conducting following steps such as self-training for volumetric segmentation. Although there exist frameworks such as 3D-CycleGAN which aim to conduct translation 3D volume-wise, they are rarely used due to typical drawbacks such as optimization complexity, computational burden, or the degradation of translation quality [5, 28].

To remedy both lack of volumetric consideration of 2D and optimization-efficiency issue of 3D methods, we propose a simple and effective pixel-level domain translation method for medical volume data by translating a stack of source domain images into target domain with both intra- and inter-slice self-attention module. Unlike previous 2D methods that only translated within a single slice, our approach leverages information from adjacent slices in the slice-direction. This is similar to recent advances in video processing, which exploit information both within and between frames [1, 2]. In contrast to 3D methods that require expensive computational cost, ours do not necessitate heavy

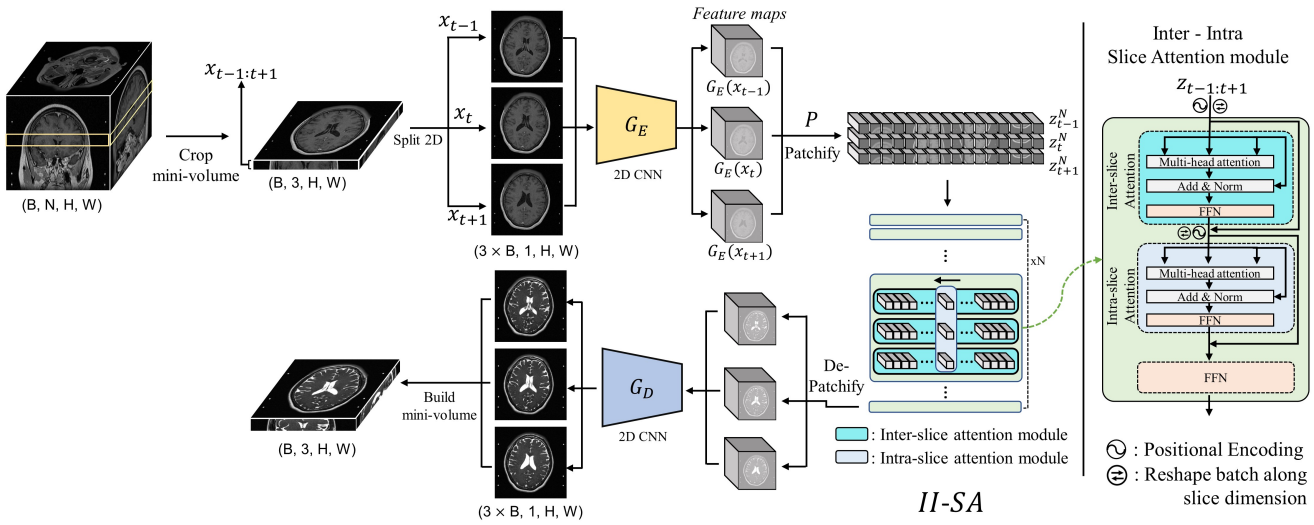


Figure 3. Illustration of the proposed 2.5D image translation with intra- and inter-slice self-attention module. Source domain volume data is cropped to a mini-volume, and fed to CNN encoder $G_E$ as a stack of batches. The weights of $G_E$ are shared during batch update. The encoded feature maps are embedded to non-overlapping patches, which become the input of intra- & inter- slice attention module $A$. Within $A$, intra- and inter-slice attention learns the volumetric information.

computation, enabling translation without intense downsampling which 3D-based translations suffer from [28].

Shown in Fig.3, proposed framework consists of sequential modules of encoder $G_E$, intra- and inter- slice self-attention ***II-SA*** and decoder $G_D$. Taking three continuous slices $x_{t-1}, x_t, x_{t+1}$ of a volume as input, they are first fed to a 2D CNN encoder $G_E$ separately. Encoded features $G_E(x_{t-1}), G_E(x_t), G_E(x_{t+1})$ are then fed to transformer-based attention module ***II-SA*** together, as feature maps from continuous frames. Before ***II-SA***, each slice is embedded into small patches $\mathbf{z}$ of size $p$, as $\mathbf{z}^n_{t-1:t+1} = P(G_E(x_{t-1:t+1}))|_{n=1:N}$, where $N = W_{G_E(x)} \cdot H_{G_E(x)}/p^2$. $W_{G_E(x)}, H_{G_E(y)}$ are each width, height of encoded feature map $G_E(x)$, respectively. In transformer-based attention module ***II-SA***, two types attention $A_{inter}$ and $A_{intra}$ exist. First, for the consideration of adjacent slices, learnable positional encoding is added to capture relative positional information within mini-volume patches. An inter-slice attention module is $A_{inter} = A(\mathbf{z}^k_{t-1}, \mathbf{z}^k_t, \mathbf{z}^k_{t+1})|_{k=1:N}$. And for the consideration of nearby pixels within a slice, an intra-slice self-attention exists, as $A_{intra} = A(\mathbf{z}^1_j, \mathbf{z}^2_j, ..., \mathbf{z}^N_j)|_{j=t-1:t+1}$. The attention module is identical for both intra- and inter-slice attention, which is

$$A = \sigma(\frac{QK^T}{\sqrt{d_K}})V \tag{1}$$

, where $Q = W_Q \cdot LN(\mathbf{z})$, $K = W_K \cdot LN(\mathbf{z})$, and $V = W_V \cdot LN(\mathbf{z})$ are query, key and value embeddings, respectively. $LN$ denotes layer-normalization, $\sigma$ denotes soft-max operation and $d_K$ denotes embedding dimension.

For maximum efficiency of translation, inter-slice attention occurs on the adjacent patches over slice-direction only for computational efficiency. After attention module, a 2D CNN decoder $G_D$ reconstructs partial volume sets in the same way as done in $G_E$, constructing a translated 2.5D mini-volume. In inference step, only the center slices of mini-volume is stacked one by one to construct a whole translated volume.

### 3.2. Volumetric self-training with uncertainty-constrained pseudo-label refinement.

Even if the source-to-target transformed images are well generated, there may still exist some distribution gap with the real target domain data, and it is difficult to completely replace it. With self-training, pseudo-labels for the unlabeled target domain data can be obtained and used to provide supervision to the target domain data, thereby mitigating the domain gap from the model's perspective. We propose to utilize volumetric self-training to close the domain gap, with a simple and novel pseudo-label refinement strategy to maximize the effect of self-training. The steps are as follows.

**3.2.1. Training segmentation with labeled synthetic scans.** With the synthetic data $\tilde{x}^t$ converted from source domain and the annotations $y^s$ on the source scans (i.e., labeled synthetic dataset), we first train a teacher segmentation network $f_{teacher}$ that minimizes the segmentation loss:

$$\mathcal{L} = \sum L_{seg}(y^s, f_{teacher}(\tilde{x}^t)) \tag{2}$$

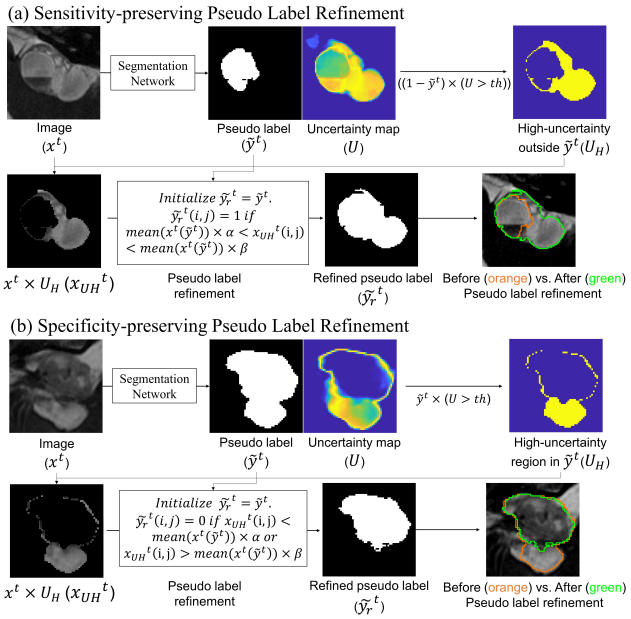


Figure 4. Illustration of the proposed uncertainty-constrained pseudo-label refinement module on vestibular schwannoma of CrossMoDA dataset. Sensitivity/specificity enhancing refinement can be performed in parallel, and either one or a combination of the two can be used according to the characteristics of each class in the dataset. Best viewed in color and on high-resolution display.

**3.2.2. Inferring pseudo-labels on unlabeled target domain data.** Once the teacher model is trained, pseudo-labels $\tilde{y}^t$ of the non-annotated real data $x^t$ can be obtained by passing real target scans to the trained segmentation model $f_{teacher}$.

$$\tilde{y}_i^t = f_{teacher}(\{x_i^t\}_{i=1}^{N_t}) \tag{3}$$

**3.2.3. Uncertainty-constrained pseudo-label refinement for sensitivity / specificity enhancement.** Since the pseudo-labels are noisy labels, they must be refined to increase accuracy and guide self-training to better direction. We devise a sensitivity and specificity enhancing pseudo-label refinement module that refines pseudo-labels based on image intensities, current pseudo-label, and high-uncertainty regions.

**Sensitivity-enhancing pseudo-label refinement.** With the inference of pseudo-labels, uncertainty (i.e., entropy) maps corresponding to each class are computed according to the following equation:

$$U = p \log p \tag{4}$$

, where $p$ is the output probability map of each class. To enhance the sensitivity of pseudo-labels, highly uncertain regions outside the range of pseudo-labels are detected. Then, if the pixel intensity in this region is within a certain range of image intensity included in the current pseudo-label, this region is included to be part of pseudo-label. The equation can be formulated as:

$$\begin{gathered}\text{Initialize } \tilde{y}_r^t = \tilde{y}^t.\ \tilde{y}_r^t(i,j) = 1 \text{ if} \\ mean\left(x^t\left(\tilde{y}^t\right)\right) \times \alpha < x_{UH}^t(i,j) < mean\left(x^t\left(\tilde{y}^t\right)\right) \times \beta\end{gathered} \tag{5}$$

where $x^t$, $\tilde{y}^t$, $\tilde{y}_r^t$, and $x_{UH}^t$ represent target domain image, pseudo-label, refined pseudo-label, and image cropped with high-uncertainty region mask, respectively. This approach is grounded on the assumption that 1) pixels that have similar intensity and 2) are close to each other in medical image are likely to belong to the same class. The workflow is described with images in Fig. 4.

**Specificity-enhancing pseudo-label refinement.** To enhance specificity of pseudo-labels, highly uncertain regions inside the range of pseudo-labels are detected. Then, if the pixel intensity in this region is outside a certain range of image intensity included in the current pseudo-label, it is excluded from the current pseudo-label.

$$\begin{gathered}\text{Initialize } \tilde{y}_r^t = \tilde{y}^t.\ \tilde{y}_r^t(i,j) = 0 \text{ if} \\ x_{UH}^t(i,j) < mean\left(x^t\left(\tilde{y}^t\right)\right) \times \alpha \text{ or} \\ x_{UH}^t(i,j) > mean\left(x^t\left(\tilde{y}^t\right)\right) \times \beta\end{gathered} \tag{6}$$

**3.2.4. Retraining segmentation with combined data.** Synthetic target scans have distribution gap with the real target scans, but they are paired with perfect annotations. On the other hand, real target scans are paired with pseudo-labels, which are yet incomplete. We incorporate both pairs to self-training, to maximize the generalization ability and minimize the performance degradation caused by difference in distributions. With the combined data of labeled synthetic target scans $(\tilde{x}^t, y^s)$ and pseudo-labeled real target scans $(x^t, \tilde{y}_r^t)$, we train a student segmentation $f_{student}$ to minimize

$$L = \sum L_{seg}(y^s, f_{student}(\tilde{x}^t)) + \sum L_{seg}(\tilde{y}_r^t, f_{student}(x^t)) \tag{7}$$

Despite not being ground truth labels, it has been reported in the previous literature that the self-training scheme and the pseudo-labels increases the performance and generalization ability of the model on unseen data by utilizing unlabeled data into training [35].

## 4. Experiments

### 4.1. Dataset

**CrossMoDA dataset.** CrossMoDA dataset [7] for vestibular schwannoma and cochlea segmentation consists of 105 labeled contrast-enhanced T1 (ceT1) MRI scans and

105 unlabeled high-resolution T2 (hrT2) MRI scans. The direction of domain adaptation was from ceT1 to hrT2 since the annotations of hrT2 scans are not publicly available. Evaluation was performed online on 32 inaccessible hrT2 scans through an official leaderboard. Please refer to the supplementary material for examples of each domain data.

**Cardiac structure segmentation dataset.** 2017 Multi-Modality Whole Heart Segmentation (MMWHS) challenge dataset [40] which consists of 20 MRI and 20 CT volumes was used for cardiac segmentation. Domain adaptation was performed from MRI to CT, with 80% of each modality used for training and 20% (i.e., 4 volumes) of randomly selected CT scans used for evaluation. Target cardiac structures for segmentation were ascending aorta (AA), left atrium blood cavity (LAC), the left ventricle blood cavity (LVC), and myocardium of left ventricle (MYO). With a fixed coronal plane resolution of 256×256, MRI and CT volumes were manually cropped to cover the 4 cardiac substructures. Please refer to the supplementary material for examples of each domain data.

### 4.2. Implementation details

**Unpaired image translation with intra- and inter-slice self-attention module.** Intra- and inter-slice attentive translation is based on 2D framework, where image slices are fed to 2D CNN $E$ as a stacked batches of size $(B \times 3, 1, H, W)$, where $H$ and $W$ each denote height and width. After $G_E$, encoded feature map is reshaped to $(B, 3, H, W)$ and fed to the attention module for the consideration of slice-direction attention. Patch size $p$ was set as 2. Transposed convolution is used to convert embedded patches back to 2D feature maps. For a stack of mini-volume generated at $G$, we implement 2D discriminator $D$ for adversarial training of domain translation. For real input to $D$, 3 consecutive slices were randomly extracted from the target volumes.

**Uncertainty-constrained pseudo-label refinement.** For each class, grid searching was conducted on how to set the threshold for masking the high-uncertainty region and also on whether to use either one, or both of the sensitivity and specificity enhancing refinement. For CrossMoDA dataset, combining sensitivity and specificity-enhancing refinement in both classes (i.e., VS and Cochlea) with uncertainty threshold of 0.3 led to the best result. For cardiac dataset, the same threshold value was used and using only specificity-enhancing refinement for all classes except one (i.e., AA) led to the best result. This is attributed to the fact that the contrast difference between adjacent substructures was very weak in cardiac CT which led to noisier pseudo-label when sensitivity-enhancing module was used. Please refer to the tables in supplementary material for ablation results. $\alpha$ and $\beta$ were set as 0.6 and 1.4, respectively.

Table 1. Ablation study on the components of the proposed method on CrossMoDA dataset. ST and PL denote self-training and pseudo-label, respectively. Best results are bolded.

| CrossMoDA (T1→T2) | | | | |
|---|---|---|---|---|
| | Dice (↑) | | ASSD (↓) | |
| Methods | VS | C | VS | C |
| Baseline | 71.5 | 72.7 | 5.91 | 1.83 |
| + intra/inter-slice attention | 80.3 | 74.5 | 1.66 | 0.67 |
| + ST w/o PL refinement | 83.2 | 76.7 | 0.58 | 0.79 |
| + ST w/ PL refinement | **84.6** | **84.9** | **0.51** | **0.14** |

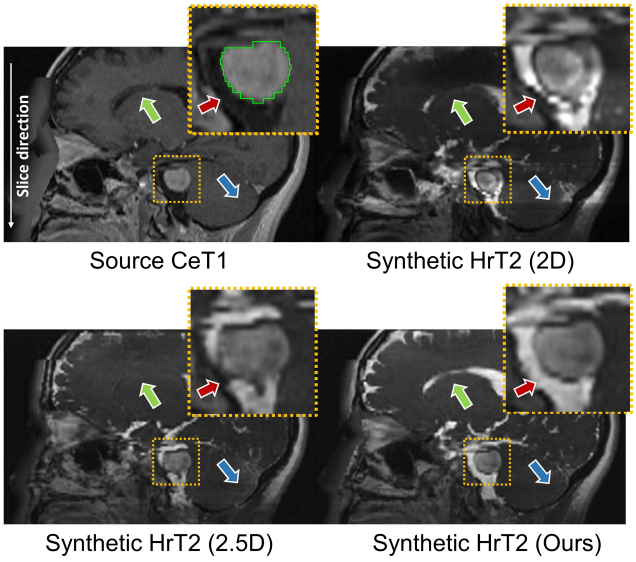


Figure 5. Representative case showing the effect of intra- and inter-slice self-attentive image translation. Translation was performed using axial slices and the resulting volume is being observed from sagittal direction. Best viewed in color and on high-resolution display.

**Volumetric self-training.** For volumetric segmentation self-training, 3D U-Net based architecture constructed from nnU-Net [14] was used. Segmentation was trained for 100 epochs using a combination of Dice and cross entropy loss with stochastic gradient descent optimizer. The initial learning rate and batch size were 1e-2 and 2, respectively. Detailed information on network architecture and training strategy can be found in the supplementary materials.

### 4.3. Results

**Effect of intra- and inter-slice self-attention module in image translation.** Fig. 5 shows a representative case that presents how the quality of the synthetic image changes depending on the design of the image translation network. Compared with the proposed method (ours), slice-direction

Table 2. Comparison of quantitative results between SDC-UDA and previous non-medical and medical UDA methods. CycleGAN, CyCADA, ADVENT, and FDA represent non-medical UDA methods while SIFA and PSIGAN represent recent medical UDA methods. Best results are bolded.

| | | **VS and Cochlea (T1 → T2)** | | | | | | **Cardiac structures (MR → CT)** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Dice (↑)** | | | **ASSD (↓)** | | | **Dice (↑)** | | | | | **ASSD (↓)** | | | | |
| | **Methods** | VS | C | Mean | VS | C | Mean | AA | LAC | LVC | MYO | Mean | AA | LAC | LVC | MYO | Mean |
| | Fully-supervised [1] | 92.5 | 87.7 | 90.1 | 0.2 | 0.1 | 0.15 | 95.9 | 92.0 | 93.0 | 88.2 | 92.3 | 1.0 | 2.5 | 1.8 | 1.7 | 1.8 |
| | w/o adaptation | 11.5 | 0.0 | 5.8 | 24.0 | NA | NA | 48.4 | 52.0 | 20.0 | 4.7 | 31.3 | 18.8 | 34.5 | 32.0 | 36.4 | 30.4 |
| Non-Medical | CycleGAN [38] | 39.7 | 0.0 | 19.8 | 10.1 | NA | NA | 58.3 | 59.8 | 61.5 | 23.1 | 50.7 | 11.4 | 10.6 | 9.2 | 14.8 | 11.5 |
| Non-Medical | CyCADA [12] | 39.2 | 0.1 | 19.7 | 10.5 | 18.1 | 14.3 | 57.2 | 62.1 | 65.0 | 47.2 | 57.9 | 8.4 | 9.0 | 8.2 | 7.9 | 8.4 |
| Non-Medical | ADVENT [34] | 8.3 | 0.0 | 4.2 | 17.1 | NA | NA | 68.2 | 71.8 | 78.8 | 50.7 | 67.4 | 9.6 | 6.6 | 4.7 | 5.1 | 6.5 |
| Non-Medical | FDA [36] | 12.5 | 0.0 | 6.3 | 13.5 | 20.5 | 17.0 | 46.5 | 68.4 | 73.8 | 55.7 | 61.1 | 10.2 | 7.6 | 4.3 | 5.0 | 6.8 |
| Medical | SIFA [4] | 47.2 | 19.7 | 33.5 | 19.1 | 3.3 | 11.2 | 76.3 | 88.1 | 74.2 | 62.4 | 75.3 | 6.1 | 3.9 | 7.1 | 4.8 | 5.5 |
| Medical | PSIGAN [15] | 57.3 | 37.2 | 47.3 | 5.5 | 2.8 | 4.1 | 67.2 | 87.0 | 80.3 | 60.8 | 73.8 | 7.5 | 4.2 | 5.5 | **4.7** | 5.5 |
| Medical | SDC-UDA (ours) | **84.6** | **84.9** | **84.8** | **0.51** | **0.14** | **0.33** | **95.8** | **91.0** | **88.6** | **66.6** | **85.5** | **1.1** | **2.8** | **3.2** | 6.3 | **3.3** |

[1] Since the ground truth for CrossMoDA Dataset is not available, fully-supervised results are referred from the dataset owner's paper [7]. These figures correspond to the upper bound and are the evaluation results from larger test set (N=137) than the one used in our study.

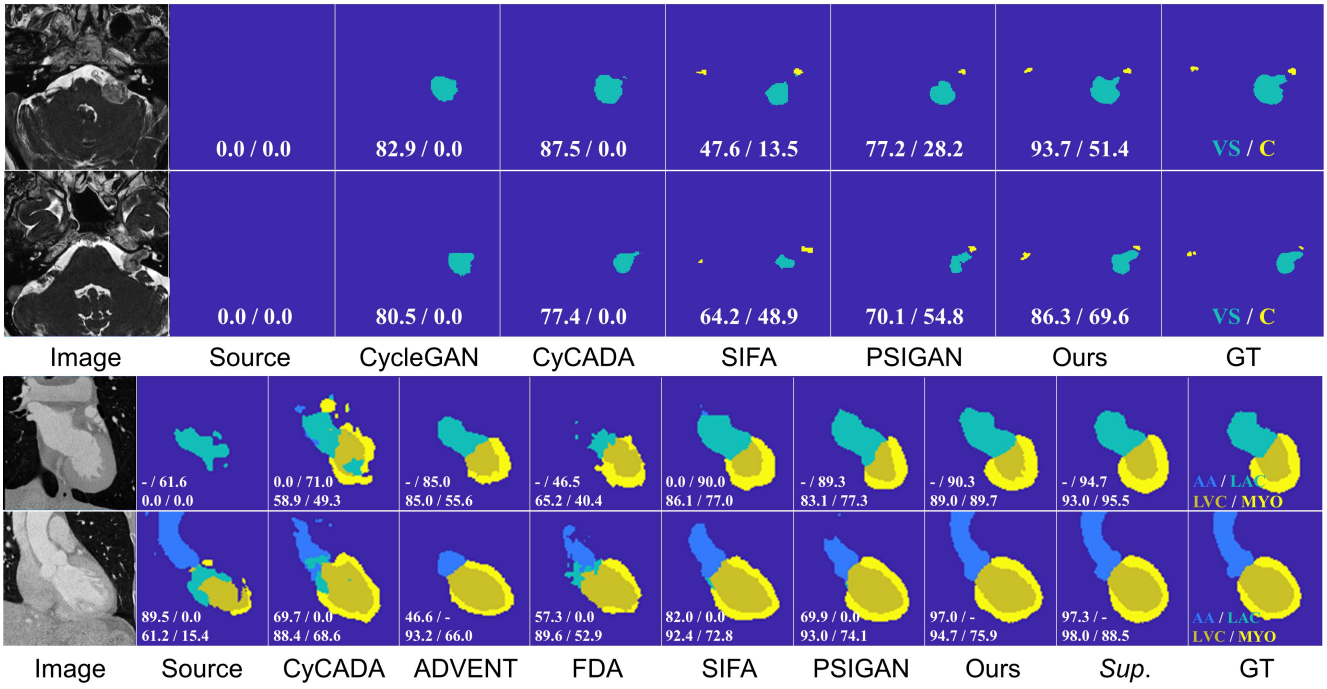


Figure 6. Comparison of qualitative results between SDC-UDA and previous non-medical and medical UDA methods. Since the ground truth of CrossMoDA dataset is not accessible, we requested to the data owner a few of the label slices and used them to make this figure. The numbers indicate Dice values of each substructure in the slice. *Sup* denotes the fully-supervised result.

inconsistency of pixel intensity was observed in the translated image due to the variation derived from slice-by-slice prediction with a 2D network (blue arrows). This was not completely overcome even with a 2.5D network that puts two adjacent slices together in the network. Also, in both 2D and 2.5D networks, it was found that the anatomy of the source image was distorted in the translated image (red and green arrows). On the other hand, it was observed that the aforementioned problems were alleviated when utilizing the intra- and inter-slice self-attention module that effectively considers neighboring anatomy. Table 1 shows the increase of final segmentation performance by incorporating intra- and inter-slice self-attention in image translation network compared to baseline that utilizes pure unpaired image translation followed by 3D segmentation network.

**Effect of uncertainty-constrained pseudo-label refinement and volumetric self-training.** Since the target domain ground truth is inaccessible for the CrossMoDA

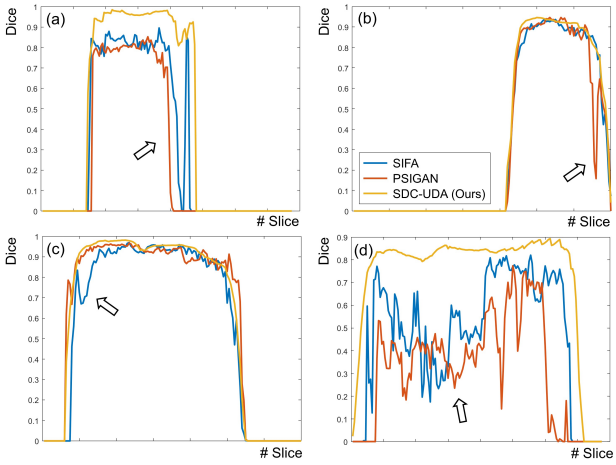


Figure 7. Representative cases demonstrating the superior slice-direction continuity of segmentation by the proposed method compared to other medical UDA methods on cardiac dataset. The plots show graphs of Dice coefficient for each slice. (a), (b), (c), and (d) denote AA, LAC, LVC, and MYO class, respectively. Best viewed in color.

dataset, the effectiveness of pseudo-label refinement was verified by comparing the performance of self-training before and after pseudo-label refinement. For cardiac datasets for which ground truth is available, it was verified by comparing dice coefficients with labels before and after pseudo-label refinement (please refer to supplementary material). Table 1 shows the improvement of target domain segmentation performance by applying the proposed uncertainty-constrained pseudo-label refinement on CrossMoDA dataset. It can be seen that volumetric self-training combined with the proposed uncertainty-constrained pseudo-label refinement increases the quantitative metrics in both VS and C by a large margin, whereas pure self-training only increase the performance in VS.

**Comparative studies.** The proposed method is compared with six popular UDA methods that are CycleGAN [38], CyCADA [12], ADVENT [34], FDA [36], SIFA [4], and PSIGAN [15]. The first four methods are from natural image field whereas the last two are UDA for medical image segmentation methods. Fully-supervised and without-adaptation results are presented to provide upper bound and lower bound, respectively. Please note that the full-supervised result of CrossMoDA dataset, for which the ground truth segmentation mask is not available, is referred from the paper of the dataset owner [7].

Table 2 and Fig. 6 shows quantitative and qualitative results of comparative studies between the proposed method and previous methods. SDC-UDA far exceeds the results of UDA for semantic segmentation studies in the non-medical field, and shows better results in multiple tasks compared to the recent studies in the medical field. In particular, there was a significant performance gap between the proposed method and comparison methods in the CrossMoDA dataset compared to the cardiac dataset. This is probably attributed to the fact that previous studies on medical UDA are dedicated to datasets in which the data are cropped around the target organs and the foreground objects occupy a large portion of the data. In contrast, SDC-UDA shows that both target structures in the CrossMoDA dataset are segmented with high accuracy.

**Slice-direction continuity of segmentation compared to previous studies.** Fig. 7 shows the excellent slice-direction continuity of segmentation by the proposed method compared to the previous studies on the MMWHS cardiac dataset. Fig. 7-(a), (b), (c), and (d) plot the slice-wise Dice values from representative cases of AA, LAC, LVC, and MYO, respectively. Our SDC-UDA shows gradual and consistent segmentation performance in the slice-direction whereas SIFA and PSIGAN, which are recently proposed medical UDA methods, show very inconsistent and sometimes sudden fluctuations of segmentation performance in the slice direction. This suggests the potential of the proposed method to be useful in the clinical practice where precise volumetric segmentation is required to analyze the patient's status with high confidence.

## 5. Conclusion

In this study, we proposed SDC-UDA, a novel volumetric UDA framework for slice-direction continuous cross-modality medical image segmentation, and validated it on multiple public datasets. SDC-UDA effectively translates medical volumes through intra- and inter-slice self-attention and better adapts to target domain via volumetric self-training enhanced by simple yet effective pseudo-label refinement strategy that utilizes uncertainty maps. Ablation studies demonstrated the effectiveness of each component and comparative studies showed the superior performance of the proposed method.

**Acknowledgements.** This research was supported by Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-TF2103-01 in constructing hardware systems and multi-modal studies. It was also supported by Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Science and ICT (2021R1A4A1031437, 2022R1A2C2008983), Artificial Intelligence Graduate School Program at Yonsei University [No. 2020-0-01361], the KIST Institutional Program (Project No.2E31051-21-204), and partially supported by the Yonsei Signature Research Cluster Program of 2022 (2022-22-0002).

## References


[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 3

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 3

[3] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 865–872, 2019. 1, 2, 3

[4] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng Ann Heng. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE transactions on medical imaging*, 39(7):2494–2505, 2020. 3, 7, 8

[5] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016. 3

[6] Reuben Dorent, Samuel Joutard, Jonathan Shapey, Sotirios Bisdas, Neil Kitchen, Robert Bradford, Shakeel Saeed, Marc Modat, Sébastien Ourselin, and Tom Vercauteren. Scribble-based domain adaptation via co-segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 479–489. Springer, 2020. 2

[7] Reuben Dorent, Aaron Kujawa, Marina Ivory, Spyridon Bakas, Nicola Rieke, Samuel Joutard, Ben Glocker, Jorge Cardoso, Marc Modat, Kayhan Batmanghelich, Arseniy Belkov, Maria Baldeon Calisto, Jae Won Choi, Benoit M. Dawant, Hexin Dong, Sergio Escalera, Yubo Fan, Lasse Hansen, Mattias P. Heinrich, Smriti Joshi, Victoriya Kashtanova, Hyeon Gyu Kim, Satoshi Kondo, Christian N. Kruse, Susana K. Lai-Yuen, Hao Li, Han Liu, Buntheng Ly, Ipek Oguz, Hyungseob Shin, Boris Shirokikh, Zixian Su, Guotai Wang, Jianghao Wu, Yanwu Xu, Kai Yao, Li Zhang, Sébastien Ourselin, Jonathan Shapey, and Tom Vercauteren. Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. *Medical Image Analysis*, 83:102628, 2023. 2, 3, 5, 7, 8

[8] Taejoon Eo, Hyungseob Shin, Yohan Jun, Taeseong Kim, and Dosik Hwang. Accelerating cartesian mri by domain-transform manifold learning in phase-encoding direction. *Medical Image Analysis*, 63:101689, 2020. 1

[9] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 1

[10] Fabian Gröger, Anne-Marie Rickmann, and Christian Wachinger. Strudel: Self-training with uncertainty dependent label refinement across domains. In *International Workshop on Machine Learning in Medical Imaging*, pages 306–316. Springer, 2021. 3

[11] Xiaoting Han, Lei Qi, Qian Yu, Ziqi Zhou, Yefeng Zheng, Yinghuan Shi, and Yang Gao. Deep symmetric adaptation network for cross-modality medical image segmentation. *IEEE transactions on medical imaging*, 41(1):121–132, 2021. 3

[12] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018. 2, 7, 8

[13] Yuankai Huo, Zhoubing Xu, Hyeonsoo Moon, Shunxing Bao, Albert Assad, Tamara K Moyo, Michael R Savona, Richard G Abramson, and Bennett A Landman. Synseg-net: Synthetic segmentation without target modality ground truth. *IEEE transactions on medical imaging*, 38(4):1016–1025, 2018. 2, 3

[14] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 6

[15] Jue Jiang, Yu-Chi Hu, Neelam Tyagi, Andreas Rimner, Nancy Lee, Joseph O Deasy, Sean Berry, and Harini Veeraraghavan. Psigan: joint probabilistic segmentation and image distribution matching for unpaired cross-modality adaptation-based mri segmentation. *IEEE Transactions on Medical Imaging*, 39(12):4071–4084, 2020. 2, 3, 7, 8

[16] Yohan Jun, Hyungseob Shin, Taejoon Eo, and Dosik Hwang. Joint deep model-based mr image and coil sensitivity reconstruction network (joint-icnet) for fast mri. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5270–5279, 2021. 1

[17] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021. 3

[18] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12975–12984, 2020. 2

[19] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, page 12, 2015. 3

[20] Jeong Ryong Lee, Sewon Kim, Inyong Park, Taejoon Eo, and Dosik Hwang. Relevance-cam: Your model already knows where to look. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14944–14953, 2021. 1

[21] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019. 2

[22] Fang Liu. Susan: segment unannotated image structure using adversarial network. *Magnetic resonance in medicine*, 81(5):3330–3345, 2019. 1, 2, 3

[23] Giuseppe Pastore, Fabio Cermelli, Yongqin Xian, Massimiliano Mancini, Zeynep Akata, and Barbara Caputo. A closer look at self-training for zero-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2693–2702, 2021. 3

[24] Jonathan Shapey, Aaron Kujawa, Reuben Dorent, Guotai Wang, Alexis Dimitriadis, Diana Grishchuk, Ian Paddick, Neil Kitchen, Robert Bradford, Shakeel R Saeed, et al. Segmentation of vestibular schwannoma from mri, an open annotated dataset and baseline algorithm. *Scientific Data*, 8(1):1–6, 2021. 3

[25] Jonathan Shapey, Guotai Wang, Reuben Dorent, Alexis Dimitriadis, Wenqi Li, Ian Paddick, Neil Kitchen, Sotirios Bisdas, Shakeel R Saeed, Sebastien Ourselin, et al. An artificial intelligence framework for automatic segmentation and volumetry of vestibular schwannomas from contrast-enhanced t1-weighted and high-resolution t2-weighted mri. *Journal of neurosurgery*, 134(1):171–179, 2019. 3

[26] Hyungseob Shin, Hyeongyu Kim, Sewon Kim, Yohan Jun, Taejoon Eo, and Dosik Hwang. Cosmos: Cross-modality unsupervised domain adaptation for 3d medical image segmentation based on target-aware domain translation and iterative self-training. *arXiv preprint arXiv:2203.16557*, 2022. 2

[27] Yejee Shin, Taejoon Eo, Hyeongseop Rha, Dong Jun Oh, Geonhui Son, Jiwoong An, You Jin Kim, Dosik Hwang, and Yun Jeong Lim. Digestive organ recognition in video capsule endoscopy based on temporal segmentation network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 136–146. Springer, 2022. 2

[28] Bin Sun, Shuangfu Jia, Xiling Jiang, and Fucang Jia. Double u-net cyclegan for 3d mr to ct image synthesis. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–8, 2022. 3, 4

[29] Bethany H Thompson, Gaetano Di Caterina, and Jeremy P Voisey. Pseudo-label refinement using superpixels for semi-supervised brain tumour segmentation. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022. 3

[30] Devavrat Tomar, Manana Lortkipanidze, Guillaume Vray, Behzad Bozorgtabar, and Jean-Philippe Thiran. Self-attentive spatial adaptive normalization for cross-modality domain adaptation. *IEEE Transactions on Medical Imaging*, 40(10):2926–2938, 2021. 3

[31] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018. 2

[32] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 1

[33] Vibashan VS, Jeya Maria Jose Valanarasu, and Vishal M Patel. Target and task specific source-free domain adaptive image segmentation. *arXiv preprint arXiv:2203.15792*, 2022. 3

[34] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019. 7, 8

[35] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 3, 5

[36] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 7, 8

[37] Zizhao Zhang, Lin Yang, and Yefeng Zheng. Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 9242–9251, 2018. 1, 2, 3

[38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 7, 8

[39] Yi Zhu, Zhongyue Zhang, Chongruo Wu, Zhi Zhang, Tong He, Hang Zhang, R Manmatha, Mu Li, and Alexander Smola. Improving semantic segmentation via self-training. *arXiv preprint arXiv:2004.14960*, 2020. 3

[40] Xiahai Zhuang and Juan Shen. Multi-scale patch and multi-modality atlases for whole heart segmentation of mri. *Medical image analysis*, 31:77–87, 2016. 3, 6

[41] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D Cubuk, and Quoc V Le. Rethinking pre-training and self-training. *arXiv preprint arXiv:2006.06882*, 2020. 3

[42] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 3

[43] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020. 3