A Unified Degradation-Robust Approach to SSL and UDA for 3D Medical Images

Suruchi Kumari, Pravendra Singh

Indian Institute of Technology Roorkee

Abstract

Medical image segmentation often faces the dual challenges of limited annotations and domain shifts, further complicated by degraded images in practical scenarios. Traditional methods tend to underperform when these issues occur simultaneously, as they are typically designed for specific tasks. To address this, we propose a unified framework that effectively handles limited annotations and domain shifts while also managing both clean and degraded images during inference. Overcoming these challenges requires focusing on three critical aspects: First, the model must be robust to various noise conditions. Second, it should excel at capturing domain-invariant features. Third, it should effectively utilize unlabeled data. We propose three major components in our approach to tackle these challenges. First, the Wavelet-based Cross-Component Exchange (WCCE) swaps high-frequency wavelet components between labeled and unlabeled images to enhance robustness. Second, we employ a diffusion VNet architecture with a reweighting mechanism to capture domain-invariant features. Finally, we utilize Cross-Decoder Pseudo (CDP) training to effectively leverage unlabeled data. Evaluations on three publicly available medical datasets and across four types of degraded image scenarios demonstrate that our method outperforms state-of-the-art (SOTA) techniques, consistently delivering superior performance across varying image qualities. Our approach not only addresses annotation scarcity and domain shift but also effectively manages noisy and blurred conditions, setting a new benchmark in medical image segmentation.

Introduction

Labeling medical images is a time-consuming and costly process, especially for volumetric data. In contrast, unlabeled data is more readily available and cost-effective to obtain. Therefore, semi-supervised learning (SSL) is highly desirable for tasks such as segmentation, which require per-pixel annotation (Chen et al. 2022; Shen et al. 2023). In SSL, limited labeled data is leveraged to explore large amounts of unlabeled data. Various SSL techniques have been proposed for semi-supervised medical image segmentation (SSMIS) (Jiao et al. 2023). However, these methods often assume that both labeled and unlabeled data come from the same distribution. In practice, medical images are frequently acquired from different medical centers using various scanning devices, leading to significant domain shifts (Kumari and Singh 2023a). As a result, SSMIS methods have limited applicability in real-world scenarios.

To address the challenges of domain shift, researchers have increasingly focused on unsupervised domain adaptation (UDA) methods. UDA addresses scenarios where a domain gap exists between labeled and unlabeled data (Liu et al. 2022). Previously, both SSL and UDA problems were addressed independently. However, since both frameworks involve learning from both labeled and unlabeled data, it is logical to develop a unified approach that effectively manages both scenarios. Existing methods for SSL do not perform well for UDA, and vice versa (Wang and Li 2024). Creating an architecture that performs well in both contexts is a significant challenge. Recently, Wang et al. (Wang and Li 2024) introduced a generic framework for volumetric SSL that simultaneously tackles SSL, and UDA tasks using an aggregating and decoupling approach. However, as shown in Table 5, this method underperforms when applied to degraded medical images. Additionally, although they propose decoupling the flows of labeled and unlabeled data to reduce overfitting, we demonstrate that training with a coupled flow of labeled and unlabeled data in a cross-pseudo supervision framework is more beneficial, as evidenced in Tables 1, 2, 3 and 4.

Medical images in the real world are often affected by various types of noise and artifacts, such as motion blur from patient movement, electronic noise from imaging equipment, and inconsistencies in image acquisition protocols. These factors can degrade image quality, making it challenging to extract accurate diagnostic information (Sagheer and George 2020; Ravishankar et al. 2017). Consequently, even if deep learning models are not explicitly trained on these degraded images, they should be robust enough to handle such imperfections. Addressing UDA, SSL, and degraded images at inference together is crucial. This will ensure that models can generalize well across different clinical settings and imaging conditions, leading to more practical, reliable, and effective medical image analysis, and ultimately improving patient care. The goal of this work is to develop an approach that performs effectively in both semi-supervised learning and unsupervised domain adaptation settings while

handling both clean and degraded images. To address these challenges, the model must be designed to tolerate various noise conditions, excel at capturing domain-invariant features, and effectively utilize unlabeled data.

We propose a novel approach with three major components to address the challenges mentioned above. First, we introduce the Wavelet-based Cross-Component Exchange (WCCE) technique, which involves exchanging high-frequency wavelet components between labeled and unlabeled images. The modified unlabeled images, now enriched with details from the labeled data, are then used for training. WCCE offers two main advantages: it enhances the model's ability to generalize by leveraging the combined frequency details of both labeled and unlabeled data, and it increases the model's robustness by training on images with varied high-frequency details, making it more resilient to real-world scenarios (i.e., noisy and blurred conditions). Second, we utilize Diffusion VNet (D-VNet) architecture with a re-weighting mechanism (RW-DVNet). This architecture extracts domain-invariant features by building a shared knowledge base from multiple domains. The re-weighting mechanism adjusts the backbone and skip-connection features in the D-VNet architecture, enhancing the quality of the sampled features (Si et al. 2024). Third, we propose the Cross-Decoder Pseudo (CDP) Training. It contains three decoders: one RW-DVNet decoder and two additional vanilla VNet decoders, enabling cross-pseudo supervision for unlabeled data. Specifically, the combined pseudo-label generated by the RW-DVNet decoder and one vanilla VNet decoder is used to guide the training of the third vanilla VNet decoder, and this process is mirrored for the other pair. This cross-supervision encourages the different decoders to learn complementary features and correct each other's mistakes, enhancing the overall performance on unlabeled data.

Our work makes significant contributions in addressing both semi-supervised learning (SSL) and unsupervised domain adaptation (UDA) challenges. We introduce a novel unified framework that effectively handles challenges such as limited annotations, domain shifts, and degraded images at inference time simultaneously using the Wavelet-based Cross-Component Exchange (WCCE), Diffusion VNet architecture with a re-weighting mechanism (RW-DVNet), and Cross-Decoder Pseudo (CDP) training. Through empirical demonstrations, we showcase the superior performance of our approach not only on clean images (idealistic scenarios) but also on degraded images (practical scenarios), proving its effectiveness across various image qualities. Our approach outperforms state-of-the-art methods and is validated on three publicly available real-world medical datasets. Additionally, we conduct experiments on four different types of degraded image scenarios to demonstrate the efficacy of our approach in handling different imaging conditions.

Related Work

Semi-supervised Medical Image Segmentation. Medical image segmentation requires per-pixel annotations, which are both costly and time-consuming to obtain. SSL has emerged as a valuable alternative to fully supervised learning, as it can explore large amounts of unlabeled data with supervision from a limited amount of labeled data. Pseudo-labeling (Lee et al. 2013) is a key technique that achieves this by generating pseudo-labels for unlabeled data based on predictions from a model trained on labeled data. Existing approaches (Yu et al. 2019; Bai et al. 2023; You et al. 2022) have achieved significant success but struggle in practical scenarios where there is a domain shift between labeled and unlabeled data. Recently, Wang et al. (Wang and Li 2024) introduced a generic framework that concurrently addresses SSL and UDA tasks through an aggregating and decoupling approach. Despite its effectiveness, the performance of this method diminishes when dealing with degraded images.

Unsupervised Domain Adaptation. Domain adaptation reduces the domain gap by leveraging both source and target domains, where labels are available for both domains. Unsupervised domain adaptation (UDA) presents a more challenging scenario as the target domain lacks labels (Kumari and Singh 2023b). In medical imaging, extensive research has been dedicated to UDA. Various methods focus on feature alignment, which adjusts feature representations to match between source and target domains (Ganin and Lempitsky 2015; Yu et al. 2022), and image translation, which converts images from one domain to another to bridge the domain gap (Zhu et al. 2017; Chen et al. 2019; Yang et al. 2019). Generative Adversarial Networks (GANs) are widely adopted for both purposes (Ganin and Lempitsky 2015; Zhu et al. 2017). Some methods utilize pseudo-labeling techniques to enhance the adaptation process by generating and using pseudo-labels for unlabeled target domain data (Wu et al. 2021a; Cho et al. 2022). Additionally, some research explores self-supervised tasks to improve domain alignment (Sun et al. 2019).

Noise in Medical Images. Deep neural networks can produce misclassifications and errors when exposed to noise (Szegedy et al. 2013). In the real world, medical images are often subject to various types of noise and artifacts that occur naturally during image acquisition, transmission, or storage. There has been extensive research dedicated to denoising medical images using both traditional and deep learning methods (El-Shafai et al. 2024). However, denoising images affects their spatial and temporal characteristics, often leading to a loss of significant details and altering image contrast (Dhar et al. 2023). Consequently, there is an urgent need for deep learning models that can perform effectively on both clean and noisy images, even if they are not explicitly trained on these variations. While many researchers have focused on defending against adversarial attacks, deliberate perturbations designed to mislead models (Dong et al. 2023), our focus here is on the naturally occurring noises in medical images.

Methodology

Preliminaries

In semi-supervised medical image segmentation, we work with labeled data $D_l = \{(x_i^l, y_i)\}_{i=1}^N$ and unlabeled data $D_u = \{x_i^u\}_{i=1}^M$. Here, N represents the number of labeled

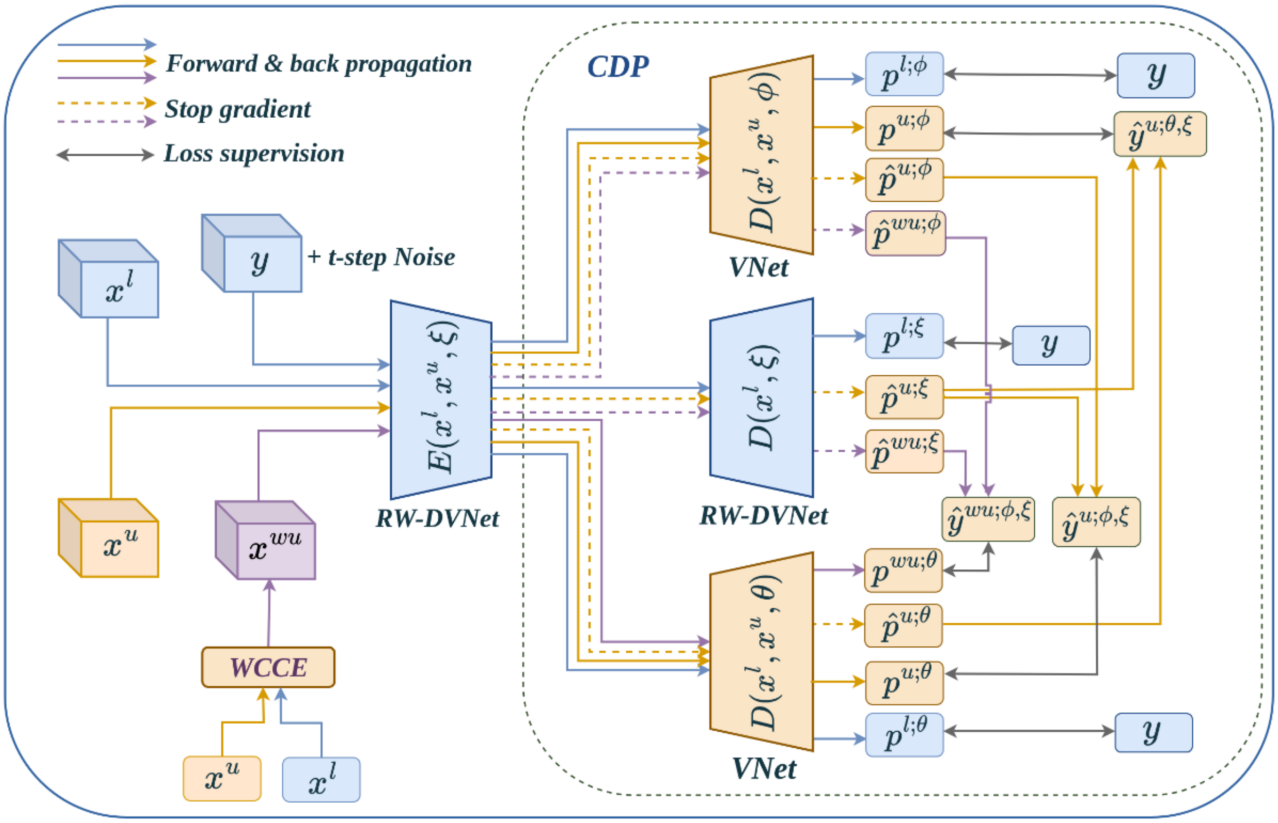

Figure 1: Overview of the proposed framework: The RW-DVNet encoder extracts domain-invariant features from both labeled and unlabeled data. Cross-Decoder Pseudo (CDP) training is utilized for pseudo-supervision across decoders.

images, and M represents the number of unlabeled images. Each image x_i belongs to $\mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{D}}$, while each label y_i belongs to $\mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{D}\times C}$, where C denotes the number of classes. A batch of input data B comprises both labeled x^l and unlabeled x^u data, denoted as B^l and B^u, respectively.

Re-weighting based Diffusion VNet architecture

We utilize a diffusion model to extract domain-invariant features. First, we convert the ground truth y of a labeled image into a one-hot format y_0. During the forward process, successive steps of Gaussian noise are added to y_0, with each step introducing a small amount of noise, progressively transforming y_0 into a noisy version, as shown below:

$$y_t = \sqrt{\bar{\alpha}_t}\, y_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0,1) \tag{1}$$

Further, y_t is concatenated with the original input x^l, as concat$([y_t, x^l])$, and passed to the diffusion encoder along with the time step t. This generates time-step-embedded multi-scale features $MS_j^{l;\xi} \in \mathbb{R}^{j\times F\times \frac{H}{2^j}\times\frac{W}{2^j}\times\frac{D}{2^j}}$, where j denotes the stage and F represents the basic feature size.

In Diffusion VNet (D-VNet) (Wang and Li 2024), the skip features are combined with the backbone features at each stage of the decoding process. The backbone features are primarily responsible for denoising, while the skip connections introduce high-frequency details into the decoder module (Si et al. 2024). Based on this insight, we employ a re-weighting strategy that dynamically balances the contribution of backbone and skip-connection features in the D-VNet architecture. Two scaling factors, b_v for the backbone feature map x_v and s_v for the skip feature map h_v, are introduced to modulate these features effectively, where v is the v-th block in the D-VNet decoder. Instead of using fixed scaling factors b_v and s_v directly, these values are used to calculate dynamic scaling factors α_v and β_v, respectively. We have calculated the value of α_v and β_v in the same way as described in (Si et al. 2024). These dynamic scaling factors are then used to generate the updated backbone and skip features, as described below:

$$x'_{v,k} = \begin{cases} x_{v,k}\cdot\alpha_v & \text{if } k < \frac{\#channel}{2} \\ x_{v,k} & \text{otherwise} \end{cases} \tag{2}$$

$$h'_v = \mathrm{IFFT}\left(\mathrm{FFT}(h_v)\odot\beta_v\right) \tag{3}$$

Where $\#channel$ denotes the total number of channels in x_v and k refers to the index of the channel in the backbone feature map x_v. FFT$(\cdot)$ and IFFT$(\cdot)$ are the Fourier transform and inverse Fourier transform, respectively. The symbol $\odot$ denotes element-wise multiplication. Finally, Both the re-weighted backbone features x'_v and skip features h'_v are then added in place of the original features x_v and h_v in

decoder $D(x^l, \xi)$. This re-weighting strategy enhances denoising while preserving critical high-frequency details.

Further, the time-step-embedded multi-scale features $MS_j^{l;\xi}$ is used by the decoder $D(x^l, \xi)$ to generate the clear label y_0. The objective loss function is formulated as follows:

$$L_{\text{deno}} = \frac{1}{B^l} \sum_{i=1}^{B^l} \mathcal{L}_{\text{DiceCE}}(p_i^{l;\xi}, y_i) \tag{4}$$

Where $\mathcal{L}_{\text{DiceCE}}(x, y) = \frac{1}{2}\left(\mathcal{L}_{\text{dice}}(x, y) + \mathcal{L}_{\text{CE}}(x, y)\right)$ represent the conventional cross-entropy and Dice loss, respectively.

Cross-Decoder Pseudo Training

This section outlines our approach to integrating labeled and unlabeled data through a Cross-Decoder Pseudo (CDP) training strategy. As mentioned before, a batch of input data B comprises both labeled x^l and unlabeled x^u data, denoted as B^l and B^u, respectively. To implement CDP, we split the batch and pass half of B^l and B^u to decoder $D(x^l, x^u, \phi)$ and the other half to decoder $D(x^l, x^u, \theta)$, instead of passing the entire batch of labeled and unlabeled images to both decoders. By doing this, we introduce diversity in both decoders, as they learn from different subsets of data (see Figure 1).

To learn from labeled data, the supervised loss for decoder $D(x^l, x^u, \phi)$ and decoder $D(x^l, x^u, \theta)$ is defined as follows:

$$\mathcal{L}_{\text{sup}_1} = \frac{1}{B^l/2} \sum_{i=1}^{B^l/2} \mathcal{L}_{\text{DiceCE}}(p_i^{l;\phi}, y_i) \tag{5}$$

$$\mathcal{L}_{\text{sup}_2} = \frac{1}{B^l/2} \sum_{i=B^l/2+1}^{B^l} \mathcal{L}_{\text{DiceCE}}(p_i^{l;\theta}, y_i) \tag{6}$$

The combined total supervised loss for two vanilla VNet decoders are:

$$\mathcal{L}_{\text{sup}} = (\mathcal{L}_{\text{sup}_1} + \mathcal{L}_{\text{sup}_2})/2 \tag{7}$$

To learn from unlabeled data, we need to generate cross-decoder pseudo-labels. To achieve this, we first convert the decoder predictions into softmax probabilities. Specifically, the prediction $\hat{p}^{u;\xi}$ is obtained by iterating the diffusion model $(E(x^l, x^u; \xi) + D(x^l; \xi))$ t times using the Denoising Diffusion Implicit Models (DDIM) method (Song, Meng, and Ermon 2020). This prediction $\hat{p}^{u;\xi}$ is then transformed into probabilities using the Gumbel-Softmax technique, with an additional smoothing step applied to enhance $\hat{p}^{u;\xi}$. Similarly, the predictions $\hat{p}^{u;\phi}$ and $\hat{p}^{u;\theta}$ from the decoders $D(x^l, x^u; \phi)$ and $D(x^l, x^u; \theta)$, respectively, are converted into softmax probabilities using the Softmax function.

The pseudo-label for the decoder $D(x^l, x^u; \phi)$ is defined as

$$\hat{y}^{u;\theta,\xi} = \text{argmax}\left(\hat{p}^{u;\xi} + \hat{p}^{u;\theta}\right).$$

Similarly, for the decoder $D(x^l, x^u; \theta)$, the pseudo-label is

$$\hat{y}^{u;\phi,\xi} = \text{argmax}\left(\hat{p}^{u;\xi} + \hat{p}^{u;\phi}\right).$$

Here, argmax returns the index of the maximum value. After obtaining the pseudo-labels, the unsupervised loss for decoder $D(x^l, x^u, \phi)$ and $D(x^l, x^u, \theta)$ are defined as follows:

$$\mathcal{L}_{\text{unsup}_1} = \frac{1}{B^u/2} \sum_{i=1}^{B^u/2} \mathcal{L}_{\text{DiceCE}}(p_i^{u;\phi}, \hat{y}_i^{u;\theta,\xi}) \tag{8}$$

$$\mathcal{L}_{\text{unsup}_2} = \frac{1}{B^u/2} \sum_{i=B^u/2+1}^{B^u} \mathcal{L}_{\text{DiceCE}}(p_i^{u;\theta}, \hat{y}_i^{u;\phi,\xi}) \tag{9}$$

The combined total unsupervised loss for two vanilla VNet decoders are:

$$\mathcal{L}_{\text{unsup}} = (\mathcal{L}_{\text{unsup}_1} + \mathcal{L}_{\text{unsup}_2})/2 \tag{10}$$

This cross decoder pseudo training strategy encourages different decoders to learn complementary features and correct each other's mistakes.

Wavelet-Based Cross-Component Exchange

High-Low Frequency Decomposition To enhance the model's robustness across varying image conditions, we employ wavelet transforms. Specifically, we first convert the input images into the wavelet space using discrete wavelet transformations (DWT) with Haar wavelets. This process allows us to separate an image into its low-frequency components, representing broad structural information, and high-frequency components, capturing detailed variations. For a 3D image, we obtain eight distinct frequency components, which can be expressed as:

- One low-pass component: LLL
- Seven high-pass components: LLH, LHL, LHH, HLL, HLH, HHL, HHH

Where the low-pass component captures the overall approximation of the image, while the high-pass components capture variations along different axes. For example, the high-pass component LLH captures variations along the Z-axis. We obtain the wavelet components/coefficients for both labeled and unlabeled images.

High-Low Frequency Interaction After the decomposition of the image into higher and lower frequency components, we utilize the High-Low Frequency Interaction (HLFI) function, which takes the higher frequency components of the labeled image and the lower frequency component of the unlabeled image to form the wavelet-enhanced unlabeled image x^{wu}, as described below:

$$\begin{aligned} x^{wu} = \text{HLFI}\big(&LLL^u,\ LLH^l,\ LHL^l,\ LHH^l,\\ &HLL^l,\ HLH^l,\ HHL^l,\ HHH^l\big) \end{aligned} \tag{11}$$

In the above equation, the superscript u denotes that the frequency components are from the unlabeled image, while the superscript l indicates that the frequency components are from the labeled image. Finally, x^{wu} is converted back to the spatial domain using the inverse discrete wavelet transform (IDWT) (see Figure 2). To incorporate x^{wu} into our training process, we generate pseudo-labels for x^{wu} in the same

manner as for the original unlabeled images. The final loss is given by:

$$\mathcal{L}_{\text{unsup}_{wu}} = \frac{1}{B^u} \sum_{i=1}^{B^u} \mathcal{L}_{\text{DiceCE}}(p_i^{wu;\theta}, \hat{y}_i^{wu;\phi,\xi}) \tag{12}$$

By utilizing $\mathcal{L}_{\text{unsup}_{wu}}$ in our model training, the model learns to handle images where high-frequency noise patterns are altered, making it less sensitive to noise present in real world medical images.

Final Loss To effectively utilize the domain agnostic features, we also apply a knowledge distillation strategy:

$\theta = w_{\text{ema}} \times \theta + (1 - w_{\text{ema}}) \times \frac{(\xi+\phi)}{2}$

$\phi = w_{\text{ema}} \times \phi + (1 - w_{\text{ema}}) \times \frac{(\xi+\theta)}{2}$

Where $w_{\text{ema}} = 0.99$.

The final loss for model training is:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{deno}} + \mathcal{L}_{\text{sup}} + \beta(\mathcal{L}_{\text{unsup}} + \mathcal{L}_{\text{unsup}_{wu}}) \tag{13}$$

At the testing stage, predictions are generated using the diffusion encoder $E(x^l, x^u; \xi)$ and the decoder $D(x^l, x^u; \theta)$. The value of β is set empirically for different datasets (details are provided in supplementary material[1]), and we use an epoch-dependent Gaussian ramp-up strategy, as described in (Lin et al. 2022), to gradually increase the influence of the unsupervised loss.

Experiments

Datasets and Experimental Details

We have performed experiments on two semi-supervised benchmark datasets, Left Atrium (Xiong et al. 2021) and Pancreas (Clark et al. 2013), and one domain adaptation dataset, MMWHS (Zhuang and Shen 2016), with two settings: CT to MR and MR to CT. The widely-used Left Atrium (LA) MRI segmentation dataset contains 80 images for training and 20 images for testing. The Pancreas dataset includes 82 CT images with annotated pancreas regions for segmentation tasks. It consists of 62 images for training and 20 images for testing. The Multi-Modality Whole Heart Segmentation (MMWHS) dataset includes paired CT and MRI images of the same anatomical regions, with 20 images in each modality. This dataset is used to evaluate the effectiveness of domain adaptation techniques for improving segmentation performance when transferring models between different imaging modalities. The Dice coefficient and the average surface distance (ASD) are used as standard metrics for medical image segmentation. For the LA and Pancreas datasets, we also utilize the Jaccard index and HD95 metrics, as referenced in (Wu et al. 2021a).

Implementation Details The proposed framework is implemented in PyTorch, running on a single NVIDIA A5000 GPU. Network parameters are optimized using SGD with Nesterov momentum set at 0.9. A "poly" decay strategy is used, as described in (Isensee et al. 2021). For further details on implementation, including data preprocessing, learning rates, and batch sizes, please refer to the supplementary material[1].The RW-DVNet module requires four parameters: two backbone factors b_1 and b_2 and two skip factors s_1 and s_2. Based on experiments, we set $b_1 = 1.1$, $b_2 = 1.2$, $s_1 = 0.6$, and $s_2 = 0.4$ for the MMWHS dataset. For the LA and Pancreas datasets, we used $b_1 = 1.3$, $b_2 = 1.4$, $s_1 = 0.9$, and $s_2 = 0.2$.

[1]The supplementary material is included in the arXiv version of this paper.

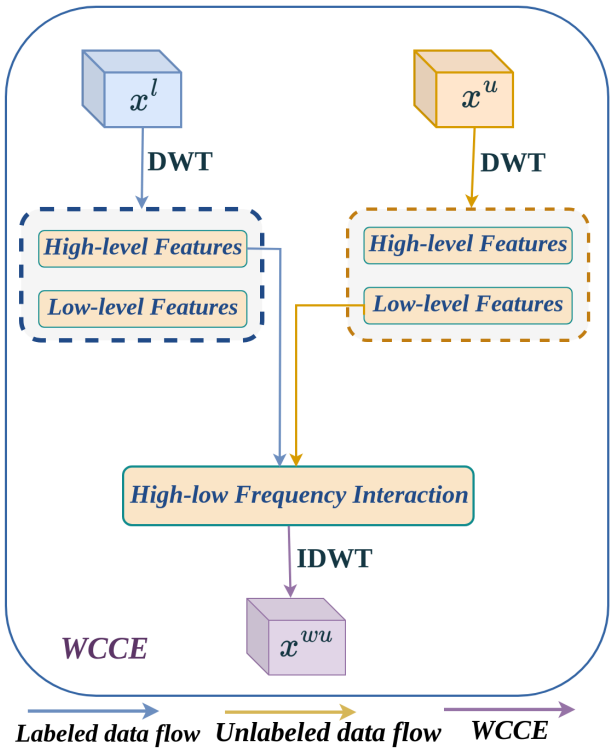

Figure 2: Wavelet-Based Cross-Component Exchange (WCCE) involves swapping high-frequency wavelet components between labeled and unlabeled images to enhance the model's robustness.

Comparison with SOTA Methods for SSL and UDA

To comprehensively evaluate our approach in addressing SSL, we compare it with several state-of-the-art (SOTA) methods, as shown in Tables 1 and 2. Our method consistently demonstrates superior performance across various SSL datasets. Similarly, for UDA, we compare our method against SOTA approaches. As shown in Tables 3 and 4, our approach outperforms existing methods. Notably, on the MMWHS dataset in the CT-to-MR setting, our method achieves a substantial improvement, with a 12.02 percentage point increase in the Dice score (see Table 4).

Comparative Experiments on Degraded Images

To assess the practical impact of our method, we apply various degradation techniques commonly occurring in medical imaging. We then compare our method with the recent

Method	Scans used		Metrics			
	Labeled	Unlabeled	Dice↑	Jaccard↑	95HD↓	ASD↓
VNet	8(10%)	0	82.74	71.72	13.35	3.26
UA-MT (Yu et al. 2019)	8(10%)	72(90%)	87.79	78.39	8.68	2.12
URPC (Luo et al. 2022)			86.92	77.03	11.13	2.28
DTC (Luo et al. 2021)			87.51	78.17	8.23	2.36
SASSNet (Li, Zhang, and He 2020)			87.54	78.05	9.84	2.59
MC-Net (Wu et al. 2021b)			87.62	78.25	10.03	1.82
LMISA-3D (Jafari et al. 2022)			86.06	76.53	12.99	2.41
SS-Net (Wu et al. 2022b)			88.55	79.62	7.49	1.90
Simcvd (You et al. 2022)			89.03	80.34	8.34	2.59
BCP (Bai et al. 2023)			89.62	81.31	6.81	1.76
MLRPL (Su et al. 2024)			89.86	81.68	6.91	1.85
Genericssl (Wang and Li 2024)			90.31	82.40	5.55	1.64
Ours			91.07	83.67	4.96	1.65

Table 1: Comparison of the proposed approach with other approaches on the LA dataset using 10 percent labeled data.

Method	Scans used		Metrics			
	Labeled	Unlabeled	Dice↑	Jaccard↑	95HD↓	ASD↓
VNet	12(20%)	0	71.52	57.68	18.12	5.41
UA-MT (Yu et al. 2019)	12(20%)	50(80%)	76.10	62.62	10.84	2.43
SS-Net (Wu et al. 2022b)			76.20	63.00	10.65	2.68
DTC (Luo et al. 2021)			76.27	62.82	8.70	2.20
SASSNet (Li, Zhang, and He 2020)			76.39	63.17	11.06	1.42
URPC (Luo et al. 2022)			80.02	67.30	8.51	1.98
MC-Net+ (Wu et al. 2022a)			79.37	66.83	8.52	1.72
SC-SSL (Miao et al. 2023b)			80.76	68.17	6.79	1.73
MCCauSSL (Miao et al. 2023a)			80.92	68.26	8.11	1.53
Genericssl (Wang and Li 2024)			79.41	66.20	8.62	1.98
MLRPL (Su et al. 2024)			81.53	69.35	6.81	1.33
Ours			82.81	71.19	5.77	1.3

Table 2: Comparison of the proposed approach with other approaches on pancreas dataset using 20 $\%$ labeled data.

Method	Dice ↑					ASD ↓
	AA	LAC	LVC	MYO	Avg	Avg
PnP-AdaNet (Dou et al. 2019)	74.0	68.9	61.9	50.8	63.9	12.8
AdaOutput (Tsai et al. 2018)	65.2	76.6	54.4	43.6	59.9	9.6
CycleGAN (Zhu et al. 2017)	73.8	75.7	52.3	28.7	57.6	10.8
CyCADA (Hoffman et al. 2018)	72.9	77.0	62.4	45.3	64.4	9.4
SIFA (Chen et al. 2019)	81.3	79.5	73.8	61.6	74.1	7.0
DSFN (Zou, Zhu, and Yan 2020)	84.7	76.9	79.1	62.4	75.8	N/A
DSAN (Han et al. 2021)	79.9	84.8	82.8	66.5	78.5	5.9
LMISA-3D (Jafari et al. 2022)	84.5	82.8	88.6	70.1	81.5	2.3
Genericssl (Wang and Li 2024)	93.2	89.5	91.7	86.2	90.1	1.7
Ours	85.4	92.9	91.0	95.1	91.1	1.6

Table 3: Comparison of the proposed method with other approaches on the MMWHS dataset for the MR to CT setting.

SOTA method (Genericssl) in handling these degraded images. As shown in Table 5, Genericssl exhibit significant performance degradation if we do not use any pre-processing (wo/p) technique to improve the quality of degraded images. For a fair comparison, we also examine the performance of our method against Genericssl when combined with the pre-processing (w/p) technique to improve the quality of degraded images, as shown in Table 5. Specifically, the degraded test images are first pre-processed using standard noise/blur filtering techniques before being input into the models. Please note that we are not using any pre-processing technique in our approach and are still performing significantly better than the SOTA approach (see Table 5). Complete details of the pre-processing steps are provided in the supplementary material[1].

Method	Dice ↑					ASD ↓
	AA	LAC	LVC	MYO	Avg	Avg
PnP-AdaNet (Dou et al. 2019)	43.7	68.9	61.9	50.8	63.9	8.9
AdaOutput (Tsai et al. 2018)	60.8	39.8	71.5	35.5	51.9	5.7
CycleGAN (Zhu et al. 2017)	64.3	30.7	65.0	43.0	50.7	6.6
CyCADA (Hoffman et al. 2018)	60.5	44.0	77.6	47.9	57.5	7.9
SIFA (Chen et al. 2019)	65.3	62.3	78.9	47.3	63.4	5.7
DSAN (Han et al. 2021)	71.3	66.2	76.2	52.1	66.5	5.4
LMISA-3D (Jafari et al. 2022)	60.7	72.4	86.2	64.1	70.8	3.6
SS-Net (Wu et al. 2022b)	62.1	58.4	68.9	51.4	60.2	5.9
BCP (Bai et al. 2023)	63.6	63.7	70.9	58.0	64.1	4.5
Genericssl (Wang and Li 2024)	62.8	87.4	61.3	74.1	71.4	7.9
Ours	76.2	84.9	91.0	81.6	83.42	2.23

Table 4: Comparison of the proposed method with other approaches on the MMWHS dataset for the CT to MR setting.

Gaussian Noise In medical images, Gaussian noise typically results from electronic interference in the imaging equipment, such as CT or MRI scanners. This noise appears as random variations in pixel values, following a Gaussian distribution. It can reduce the clarity of fine anatomical details and affect the accuracy of diagnostic interpretations. To check the robustness of our model against gaussian noise, we introduce noise with standard deviations (σ) of 0.1 and 0.2 into the test images. Our method shows minimal performance drop under noise compared to SOTA method, maintaining results close to clean images (Table 5).

Motion Blur In medical imaging, motion blur is caused by patient movement or instability during scanning, leading to blurred images that can impact the clarity of anatomical structures and diagnostic accuracy. To assess our model's robustness against motion blur, we apply motion blur with kernel sizes $k = 5$ and $k = 10$, which average pixel values along a horizontal line, varying the strength and direction of the blur. The results demonstrate that our method outperforms the SOTA method in handling motion blur.

Rician Noise Rician noise is a prevalent type of noise in medical imaging, particularly in MRI scans. To simulate Rician noise in an image, Gaussian noise is first added separately to the real and imaginary parts of the image signal, each drawn from a normal distribution with a specific standard deviation (σ). The Rician noise is then computed by taking the square root of the sum of the squared real and imaginary components. To evaluate the robustness of our model against Rician noise, we introduce Rician noise with sigma values of 0.05 and 0.1 into the test images. The results show that while preprocessing can help SOTA method mitigate Rician noise, our method consistently achieves better resilience across imaging modalities and noise levels wo/p, as shown in Table 5.

Poisson Noise Poisson noise occurs in medical images due to the random nature of photon detection during the image acquisition process, particularly in low-light or low-dose imaging such as X-rays and CT scans. This noise is signal-dependent, meaning it increases with the intensity of the signal, and can affect the clarity of the images. To simulate this condition, we scaled the input image to introduce intensity variations and then applied Poisson noise to mimic

Dataset	Method	Original image	Gaussain Noise		Motion Blur		Rician Noise		Poisson Noise
			σ =0.1	σ = 0.2	k = 5	k = 10	σ = 0.05	σ = 0.1	Scale = 60
MR to CT	Genericssl (wo/p)	90.1	89.72	88.57	89.01	86.67	87.72	85.1	87.41
	Genericssl (w/p)	90.1	89.81	89.01	89.52	88.16	89.26	88.29	88.4
	Ours (wo/p)	91.1	91.02	90.82	90.83	89.51	90.92	90.2	90.31
CT to MR	Genericssl (wo/p)	71.4	70.55	68.76	68.51	64.85	70.34	65.65	N/A
	Genericssl (w/p)	71.4	70.69	69.13	69.4	67.27	70.91	68.43	N/A
	Ours (wo/p)	83.4	83.37	83.06	82.92	81.74	83.33	82.31	N/A
LA	Genericssl (wo/p)	90.31	89.81	88.27	89.53	86.0	89.75	87.12	N/A
	Genericssl (w/p)	90.1	89.87	88.61	89.7	86.8	89.86	87.92	N/A
	Ours (wo/p)	91.15	91.01	90.7	90.85	88.21	91.12	91.0	N/A
Pancreas	MLRPL (wo/p)	81.53	80.96	79.5	80.64	78.23	80.54	78.65	80.65
	MLRPL (w/p)	81.53	81.23	80.21	81.01	79.26	80.91	79.76	81.02
	Ours (wo/p)	82.81	82.66	82.33	82.24	81.2	82.5	81.54	82.03

Table 5: Comparison of our method with SOTA methods on degraded images using Dice (↑) score, evaluated on both preprocessed (w/p) and non-preprocessed (wo/p) images. Poisson noise is not applicable to MR images.

D-VNet	VNet2	CDP	RW-DVNet	WCCE	Dice ↑	ASD ↓
✓					60.6	16.1
✓	✓				65.4	9.0
✓	✓	✓			76.1	6.4
✓	✓	✓	✓		78.8	4.16
✓	✓	✓	✓	✓	**83.4**	**2.23**

Table 6: Ablation study of different components used in our approach on the MMWHS dataset in the CT to MR setting.

the effects of low-dose CT imaging. However, adding Poisson noise to reconstructed images does not fully replicate the spatial characteristics of real CT noise, which originates at the sinogram level. This represents a limitation, as the simulated noise is not entirely representative of actual CT noise. Our approach exhibits robust performance even under Poisson noise conditions, whereas Genericssl struggle to achieve comparable results, even with the use of preprocessing techniques.

Ablation Study

Contribution of Different Components Our method incorporates three major components: RW-DVNet, CDP, and WCCE. Table 6 presents the ablation study, illustrating the impact of these components on the overall performance. We began by applying self-training to the diffusion VNet (D-VNet) architecture (baseline model), which achieved a Dice score of 60.6%. To enhance performance, we introduced an additional vanilla VNet decoder (VNet2) and implemented cross supervision between the diffusion decoder and the new vanilla decoder. This modification led to a significant improvement in the Dice score to 65.4%. Further enhancement was achieved by adding a second vanilla decoder and fully utilizing our proposed CDP, which resulted in a notable increase in the Dice score to 76.1%. Incorporating the reweighting strategy within the D-VNet decoder added an additional 2.7% improvement, raising the Dice score to 78.8%. Finally, integrating the WCCE module provided a substantial boost, culminating in a final Dice score of 83.4% and an ASD of 2.23. This progression clearly demonstrates the cumulative benefits of each component in our approach.

Impact of WCCE for Degraded Images WCCE is crucial for enhancing the robustness of our method against degraded images. By incorporating WCCE, the model learns to manage images with modified high-frequency noise patterns, thus reducing its sensitivity to degraded images during inference time. For this experiment, we conduct an ablation study where WCCE is removed from our approach, and we evaluate the results on degraded images. The ablation study presented in Table 7 clearly demonstrates the significant impact of the WCCE module in handling degraded images.

Method	Gaussian Noise	Motion Blur	Rician Noise
	σ = 0.2	k = 10	σ = 0.1
w/o WCCE	81.88	79.02	78.65
w/ WCCE	83.06	81.74	82.31

Table 7: Ablation study on the impact of WCCE on degraded images in the MMWHS dataset, with the CT to MR setting. The first row shows the dice score without (w/o) using WCCE in our approach, and the second row shows the results with (w/) WCCE in our approach.

Conclusion

This work introduces a novel unified architecture designed to tackle the challenges of limited annotations, domain shifts, and degraded 3D images in medical image segmentation. Our approach integrates three key components: the Wavelet-Based Cross-Component Exchange (WCCE) module, which enhances model robustness by swapping high-frequency wavelet components between labeled and unlabeled images; the Diffusion VNet architecture with a reweighting mechanism (RW-DVNet), complemented by two vanilla decoders to extract domain-invariant features; and Cross-Decoder Pseudo (CDP) Training, which fosters cross-pseudo supervision for unlabeled data, encouraging different decoders to learn complementary features and correct each other's mistakes. Through extensive experimentation, we validate the effectiveness of our approach against various state-of-the-art techniques, demonstrating superior performance in both clean and noisy or blurry environments. This work highlights its significance by addressing the complexities of real-world medical imaging applications, moving beyond idealized experimental settings.

References

Bai, Y.; Chen, D.; Li, Q.; Shen, W.; and Wang, Y. 2023. Bidirectional Copy-Paste for Semi-Supervised Medical Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11514–11524.

Chen, C.; Dou, Q.; Chen, H.; Qin, J.; and Heng, P.-A. 2019. 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, 865–872.

Chen, X.; Wang, X.; Zhang, K.; Fung, K.-M.; Thai, T. C.; Moore, K.; Mannel, R. S.; Liu, H.; Zheng, B.; and Qiu, Y. 2022. Recent advances and clinical applications of deep learning in medical image analysis. *Medical image analysis*, 79: 102444.

Cho, H.; Nishimura, K.; Watanabe, K.; and Bise, R. 2022. Effective pseudo-labeling based on heatmap for unsupervised domain adaptation in cell detection. *Medical Image Analysis*, 79: 102436.

Clark, K.; Vendt, B.; Smith, K.; Freymann, J.; Kirby, J.; Koppel, P.; Moore, S.; Phillips, S.; Maffitt, D.; Pringle, M.; et al. 2013. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of digital imaging*, 26: 1045–1057.

Dhar, T.; Dey, N.; Borra, S.; and Sherratt, R. S. 2023. Challenges of deep learning in medical image analysis—improving explainability and trust. *IEEE Transactions on Technology and Society*, 4(1): 68–75.

Dong, J.; Chen, J.; Xie, X.; Lai, J.; and Chen, H. 2023. Adversarial attack and defense for medical image analysis: Methods and applications. *arXiv preprint arXiv:2303.14133*.

Dou, Q.; Ouyang, C.; Chen, C.; Chen, H.; Glocker, B.; Zhuang, X.; and Heng, P.-A. 2019. Pnp-adanet: Plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation. *IEEE Access*, 7: 99065–99076.

El-Shafai, W.; El-Nabi, S. A.; Ali, A. M.; El-Rabaie, E.-S. M.; and Abd El-Samie, F. E. 2024. Traditional and deep-learning-based denoising methods for medical images. *Multimedia Tools and Applications*, 83(17): 52061–52088.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 1180–1189. PMLR.

Han, X.; Qi, L.; Yu, Q.; Zhou, Z.; Zheng, Y.; Shi, Y.; and Gao, Y. 2021. Deep symmetric adaptation network for cross-modality medical image segmentation. *IEEE transactions on medical imaging*, 41(1): 121–132.

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

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211.

Jafari, M.; Francis, S.; Garibaldi, J. M.; and Chen, X. 2022. LMISA: A lightweight multi-modality image segmentation network via domain adaptation using gradient magnitude and shape constraint. *Medical Image Analysis*, 81: 102536.

Jiao, R.; Zhang, Y.; Ding, L.; Xue, B.; Zhang, J.; Cai, R.; and Jin, C. 2023. Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. *Computers in Biology and Medicine*, 107840.

Kumari, S.; and Singh, P. 2023a. Data efficient deep learning for medical image analysis: A survey. *arXiv preprint arXiv:2310.06557*.

Kumari, S.; and Singh, P. 2023b. Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives. *Computers in Biology and Medicine*, 107912.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 896. Atlanta.

Li, S.; Zhang, C.; and He, X. 2020. Shape-aware semi-supervised 3D semantic segmentation for medical images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, 552–561. Springer.

Lin, Y.; Yao, H.; Li, Z.; Zheng, G.; and Li, X. 2022. Calibrating label distribution for class-imbalanced barely-supervised knee segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 109–118. Springer.

Liu, X.; Yoo, C.; Xing, F.; Oh, H.; El Fakhri, G.; Kang, J.-W.; Woo, J.; et al. 2022. Deep unsupervised domain adaptation: A review of recent advances and perspectives. *APSIPA Transactions on Signal and Information Processing*, 11(1).

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

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

Miao, J.; Chen, C.; Liu, F.; Wei, H.; and Heng, P.-A. 2023a. Caussl: Causality-inspired semi-supervised learning for medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21426–21437.

Miao, J.; Zhou, S.-P.; Zhou, G.-Q.; Wang, K.-N.; Yang, M.; Zhou, S.; and Chen, Y. 2023b. SC-SSL: Self-correcting Collaborative and Contrastive Co-training Model for Semi-Supervised Medical Image Segmentation. *IEEE Transactions on Medical Imaging*.

Ravishankar, A.; Anusha, S.; Akshatha, H.; Raj, A.; Jahnavi, S.; and Madhura, J. 2017. A survey on noise reduction

techniques in medical images. In *2017 international conference of electronics, communication and aerospace technology (ICECA)*, volume 1, 385–389. IEEE.

Sagheer, S. V. M.; and George, S. N. 2020. A review on medical image denoising algorithms. *Biomedical signal processing and control*, 61: 102036.

Shen, W.; Peng, Z.; Wang, X.; Wang, H.; Cen, J.; Jiang, D.; Xie, L.; Yang, X.; and Tian, Q. 2023. A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction. *IEEE transactions on pattern analysis and machine intelligence*, 45(8): 9284–9305.

Si, C.; Huang, Z.; Jiang, Y.; and Liu, Z. 2024. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4733–4743.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Su, J.; Luo, Z.; Lian, S.; Lin, D.; and Li, S. 2024. Mutual learning with reliable pseudo label for semi-supervised medical image segmentation. *Medical Image Analysis*, 103111.

Sun, Y.; Tzeng, E.; Darrell, T.; and Efros, A. A. 2019. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tsai, Y.-H.; Hung, W.-C.; Schulter, S.; Sohn, K.; Yang, M.-H.; and Chandraker, M. 2018. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 7472–7481.

Wang, H.; and Li, X. 2024. Towards generic semi-supervised framework for volumetric medical image segmentation. *Advances in Neural Information Processing Systems*, 36.

Wu, S.; Chen, C.; Xiong, Z.; Chen, X.; and Sun, X. 2021a. Uncertainty-aware label rectification for domain adaptive mitochondria segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, 191–200. Springer.

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

Wu, Y.; Wu, Z.; Wu, Q.; Ge, Z.; and Cai, J. 2022b. Exploring smoothness and class-separation for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 34–43. Springer.

Wu, Y.; Xu, M.; Ge, Z.; Cai, J.; and Zhang, L. 2021b. Semi-supervised left atrium segmentation with mutual consistency training. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, 297–306. Springer.

Xiong, Z.; Xia, Q.; Hu, Z.; Huang, N.; Bian, C.; Zheng, Y.; Vesal, S.; Ravikumar, N.; Maier, A.; Yang, X.; et al. 2021. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical image analysis*, 67: 101832.

Yang, J.; Dvornek, N. C.; Zhang, F.; Chapiro, J.; Lin, M.; and Duncan, J. S. 2019. Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, 255–263. Springer.

You, C.; Zhou, Y.; Zhao, R.; Staib, L.; and Duncan, J. S. 2022. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(9): 2228–2237.

Yu, L.; Wang, S.; Li, X.; Fu, C.-W.; and Heng, P.-A. 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, 605–613. Springer.

Yu, M.; Guan, H.; Fang, Y.; Yue, L.; and Liu, M. 2022. Domain-Prior-Induced Structural MRI Adaptation for Clinical Progression Prediction of Subjective Cognitive Decline. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 24–33. Springer.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Zhuang, X.; and Shen, J. 2016. Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. *Medical image analysis*, 31: 77–87.

Zou, D.; Zhu, Q.; and Yan, P. 2020. Unsupervised Domain Adaptation with Dual-Scheme Fusion Network for Medical Image Segmentation. In *IJCAI*, 3291–3298.