Neural Networks 165 (2023) 119–134

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

MSCDA: Multi-level semantic-guided contrast improves unsupervised domain adaptation for breast MRI segmentation in small datasets

Sheng Kuang [a], Henry C. Woodruff [a,b], Renee Granzier [c], Thiemo J.A. van Nijnatten [b,d], Marc B.I. Lobbes [b,d,e], Marjolein L. Smidt [c,d], Philippe Lambin [a,b], Siamak Mehrkanoon [f,*]

[a] The D-Lab, Department of Precision Medicine, GROW – School or Oncology and Reproduction, Maastricht University, Maastricht, The Netherlands
[b] Department of Radiology and Nuclear Medicine, Maastricht University Medical Centre+, Maastricht, The Netherlands
[c] Department of Surgery, Maastricht University Medical Centre+, Maastricht, The Netherlands
[d] GROW – School for Oncology and Reproduction, Maastricht University, Maastricht, The Netherlands
[e] Department of Medical Imaging, Zuyderland Medical Center, Sittard-Geleen, The Netherlands
[f] Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 4 January 2023
Received in revised form 9 April 2023
Accepted 9 May 2023
Available online 19 May 2023

Keywords:
Breast segmentation
Unsupervised domain adaptation
Contrastive learning

ABSTRACT

Deep learning (DL) applied to breast tissue segmentation in magnetic resonance imaging (MRI) has received increased attention in the last decade, however, the domain shift which arises from different vendors, acquisition protocols, and biological heterogeneity, remains an important but challenging obstacle on the path towards clinical implementation. In this paper, we propose a novel Multi-level Semantic-guided Contrastive Domain Adaptation (MSCDA) framework to address this issue in an unsupervised manner. Our approach incorporates self-training with contrastive learning to align feature representations between domains. In particular, we extend the contrastive loss by incorporating pixel-to-pixel, pixel-to-centroid, and centroid-to-centroid contrasts to better exploit the underlying semantic information of the image at different levels. To resolve the data imbalance problem, we utilize a category-wise cross-domain sampling strategy to sample anchors from target images and build a hybrid memory bank to store samples from source images. We have validated MSCDA with a challenging task of cross-domain breast MRI segmentation between datasets of healthy volunteers and invasive breast cancer patients. Extensive experiments show that MSCDA effectively improves the model's feature alignment capabilities between domains, outperforming state-of-the-art methods. Furthermore, the framework is shown to be label-efficient, achieving good performance with a smaller source dataset. The code is publicly available at https://github.com/ShengKuangCN/MSCDA.

1. Introduction

Breast cancer is the most commonly diagnosed cancer in women and contributes to 15% of mortality worldwide, ranking as a leading cause of death in many countries (Francies, Hull, Khanyile, & Dlamini, 2020; Sung et al., 2021). The significantly increasing mortality rates of breast cancer, especially in developing countries and low-income regions, lead to increased burdens for patients, their families, and society, highlighting the need for early detection and intervention (Azamjah, Soltan-Zadeh, & Zayeri, 2019). In the past decades, breast magnetic resonance imaging (MRI) has been recommended to supplement conventional mammography and ultrasound techniques to screen women at a high risk of breast cancer and determine the extent of breast cancer after diagnosis (Lowry et al., 2022; Sardanelli et al., 2017; Saslow et al., 2007).

A further step towards advanced MRI-based diagnosis is accurate breast segmentation. In clinical routine, whole-breast segmentation and analysis are conducted manually relying on the expertise of clinicians, which is a challenging and time-consuming process. With the advent of computer vision, atlas- and statistical-based methods with high accuracy compared to manual segmentation have been proposed (Wu, Weinstein, Conant, & Kontos, 2013). Recently, numerous deep learning (DL) approaches have been developed to further improve the performance of breast segmentation. These tools extract salient features directly from the images and automatically segment the breast boundary (Dalmış et al., 2017; Hu et al., 2018; Ivanovska et al., 2019), breast fibroglandular tissue (FGT) (Dalmış et al., 2017; Ivanovska et al., 2019) and breast lesions (Gallego-Ortiz & Martel, 2017; Negi, Raj, Nersisson, Zhuang, & Murugappan, 2020; Zhang, Saha, Zhu, & Mazurowski, 2018), which are less prone to errors and have achieved encouraging Dice similarity coefficients (DSCs) in various datasets.

[*] Corresponding author.
E-mail address: s.mehrkanoon@uu.nl (S. Mehrkanoon).

https://doi.org/10.1016/j.neunet.2023.05.014
0893-6080/

Despite the high popularity of DL approaches, there are some barriers on the path to clinical implementation. One main concern is performance degradation due to large inhomogeneities present in MRI datasets, leading to differing imaging feature distributions between training (source domain) and testing (target domain) datasets, also known as the domain shift problem. Inhomogeneities in MRI datasets primarily stem from two factors: acquisition heterogeneity and biological heterogeneity. Acquisition heterogeneity refers to variations in acquisition protocols, machine vendors, contrast-agent enhancement, and reconstruction algorithms, while biological heterogeneity encompasses differences in breast sizes and densities, menstrual cycle effects, and stage of disease progression. Additionally, factors such as patient positioning, motion artifacts, and imaging artifacts may contribute to dataset inhomogeneities. These inconsistencies within the images may lead to unstable performance of DL models, as highlighted in recent studies (Granzier et al., 2022). Although this problem could be addressed by acquiring large and varied datasets of accurately annotated target images for training, this exercise would be labor-consuming and expensive, and is further hindered by legal and ethical considerations regarding the sharing of patient data. Thus, recent published studies (Hoffman et al., 2018; Hoffman, Wang, Yu, & Darrell, 2016; Hoyer, Dai, & Van Gool, 2022) focus on developing unsupervised domain adaptation (UDA) methods to mitigate the distribution discrepancy without target labels.

Recent advances in UDA methods have enabled its application to medical images by transferring the knowledge from the source domain to the target domain (Chaitanya, Erdil, Karani, & Konukoglu, 2020; Liu et al., 2022; Perone, Ballester, Barros, & Cohen-Adad, 2019; Shanis, Gerber, Gao, & Enquobahrie, 2019). For instance, adversarial learning (e.g. Cycle-GAN) adopts the discriminator to align the distribution in the latent feature space (Dou et al., 2018; Hoffman et al., 2018) or label space (Tsai et al., 2018). Self-training (Perone et al., 2019; Tarvainen & Valpola, 2017) is another promising method which combines entropy minimization (Vu, Jain, Bucher, Cord, & Pérez, 2019), pseudo-label denoising (Zhang et al., 2021) or adversarial learning (Shanis et al., 2019). More recently, incorporating self-training with complementary contrastive learning shows remarkable performance improvement by utilizing the ground truth and pseudo labels as supervised semantic signals to guide the training (Chaitanya et al., 2020; Liu et al., 2022; Zhang, Koltun, Torr, Ranftl, & Richter, 2022).

Contrastive learning in this case explicitly computes the inter-category similarity between pixel representation pairs (Zhao, Vemulapalli, et al., 2021; Zhong et al., 2021) (refers to pixel-to-pixel (P2P) contrast) aiming to learn an invariant representation in feature space. However, it still suffers from the following two major concerns that are not taken into account: (i) P2P contrast skips the structure context of adjacent pixels, so it does not extensively exploit the semantic information present in the MRI scan. To alleviate this problem, we propose a method to integrate different levels of semantic information into a contrastive loss function. More specifically, the mean value of the pixel representations of a specific category, i.e., the centroid, should be similar to the pixels contained in the region. Likewise, centroids, regardless of whether they are from the same domain, should also be close to centroids of the same category and far away from centroids of other categories. We denote these two relations as pixel-to-centroid (P2C) and centroid-to-centroid (C2C) respectively. (ii) A common practice to perform inter-category contrast is to generate positive and negative pairs by sampling partial pixel representations in a mini-batch (Chaitanya et al., 2020). However, the imbalanced proportion between background and regions of interest (ROIs) in the breast MRIs poses a challenge to obtain adequate pairs during training. To address this problem, we build a hybrid memory bank and optimize the sampling strategy to ensure enough cross-domain positive and negative pairs especially for the highly imbalanced mini-batches. Additionally, we also explore the impact of anchors and samples from different domains on model performance.

In summary, we extend the contrastive UDA framework for breast segmentation to further mitigate the domain shift problem. To the best of our knowledge, this is the first attempt to apply contrastive UDA in breast MRI. We briefly provide the novel contributions of our work as follows:

1. To solve the domain shift problem in breast MRI, we develop a novel Multi-level Semantic-guided Contrastive Domain Adaptation (MSCDA) framework for cross-domain breast tissue segmentation.
2. To exploit the semantic information present in source labels, we propose a method that combines pixel-to-pixel, pixel-to-centroid and centroid-to-centroid contrasts into the loss function.
3. To resolve the data imbalance problem, we develop a hybrid memory bank that saves both category-wise pixel and centroid samples. We further investigate a category-wise cross-domain sampling strategy to form adequate contrastive pairs.
4. To validate the performance of the UDA framework, we replicate our experiment under multiple source datasets of different sizes. The results show robust performance and label-efficient learning ability. We further show that our framework achieves comparable performance to supervised learning.

2. Related works

2.1. Semantic segmentation

Semantic segmentation is an essential and hot topic in computer vision, achieving automatic categorization of each pixel (or voxel) into one or more categories. In recent years, convolutional neural networks (CNNs) have shown significant results in multiple fields. Fully convolutional network (FCN) (Long, Shelhamer, & Darrell, 2015), as one of the most remarkable early-stage segmentation architectures, demonstrated the pixel-level representation learning ability of CNNs. However, CNNs are still far from maturity in terms of accuracy and efficiency. Therefore, many mechanisms have been proposed to improve segmentation performance. For instance, U-Net (Ronneberger, Fischer, & Brox, 2015) introduced skip connections in an encoder–decoder design to solve the vanishing gradient problem; DeepLab v3+ (Chen, Zhu, Papandreou, Schroff, & Adam, 2018) proposed Atrous Spatial Pyramid Pooling (ASPP) to capture more context information in multi-scale receptive fields. Meanwhile, inspired by the effectiveness of residual blocks, ResNet (He, Zhang, Ren, & Sun, 2016) was adopted as the backbone in many encoder–decoder segmentation frameworks (Chen et al., 2018; He, Deng, & Qiao, 2019; Wu, Zhang, Huang, Liang, & Yu, 2019; Zhang, Dana, et al., 2018) to provide deep feature representations.

2.2. MRI-based semantic segmentation

DL techniques have also been widely adopted in various medical fields that pave the way towards more precise and automated clinical diagnosis and prognosis, as seen in oncology (Kleppe et al., 2021; Yesilkaya, Perc, & Isler, 2022), neuroimaging (Qiu et al., 2022), and cardiology (Surucu, Isler, Perc, & Kara, 2021). In MRI-based segmentation, many studies have shown that DL methods can significantly improve the accuracy and efficiency of segmenting biological tissues (Bleker et al., 2022; Zhao et al., 2020). For instance, several studies have proposed DL methods for brain MRI segmentation (Despotović, Goossens, & Philips, 2015; Ito, Nakae, Hata, Okano, & Ishii, 2019), while others have developed methods for liver (Ibtehaz & Rahman, 2020) and prostate (Hung et al., 2022) segmentation. In breast MRI, previous DL methods focus on the segmentation of contours (Dalmış et al., 2017; Piantadosi, Sansone, & Sansone, 2018; Zhang et al., 2019, 2020) and lesions (Dalmış et al., 2017; El Adoui, Mahmoudi, Larhmam, & Benjelloun, 2019; Zhang, Saha, et al., 2018). Despite the promising performance, these methods require large datasets with expert annotations, which is expensive and time-consuming.

2.3. Contrastive learning

Contrastive learning (CL) was introduced as a self-supervised learning framework, allowing the model to learn representations without labels (Chen, Fan, Girshick, & He, 2020; Chen, Kornblith, Norouzi, & Hinton, 2020; Grill et al., 2020; He, Fan, Wu, Xie, & Girshick, 2020; Oord, Li, & Vinyals, 2018). An essential step of early CL methods is to build a pretext task, such as instance discrimination (Chen, Kornblith, et al., 2020; He et al., 2020; Wu, Xiong, Yu, & Lin, 2018), to discriminate a positive pair (two augmented views of an identical image) from negative pairs (augmented view of other images). Based on this pioneering approach, many subsequent advanced mechanisms have been proposed to improve the representation learning ability. For example, Moco v1 (He et al., 2020) and v2 (Chen, Fan, et al., 2020) combined a momentum encoder with a first-in-first-out queue as a memory bank to maintain more negative samples. This results in an improved classification performance e.g., ImageNet (Deng et al., 2009) and enables training the network on normal graphics processing units (GPUs). Afterwards, the projection head (Chen, Kornblith, et al., 2020) and the prediction head (Grill et al., 2020) were introduced respectively to improve the classification accuracy on downstream tasks.

For semantic segmentation tasks, recent CL works leverage the pixel-level labels as supervised signals (Chaitanya et al., 2020; Hu, Zeng, Xu, & Shi, 2021; Wang et al., 2021; Zhao, Vemulapalli, et al., 2021; Zhong et al., 2021). The underlying idea is to group the pixel representations from the same category and to separate pixel representations from different categories. Zhao, Vemulapalli, et al. (2021) introduced a label-efficient two-stage method that pre-trained the network by using P2P contrastive loss and then fine-tuned the network using cross-entropy (CE) loss (Bishop & Nasrabadi, 2006). PC^{2}Seg (Zhong et al., 2021) improved this method in a one-stage semi-supervised learning (SSL) approach by jointly updating the network weights with pixel contrastive loss and consistency loss. ContrastiveSeg (Wang et al., 2021) combined pixel-to-region contrastive loss to explicitly leverage the context relation across images. It also proves that storing the samples from recent batches can boost segmentation tasks, especially when the training batch size is limited by the memory of device. Similar to Wang et al. (2021) and Zhong et al. (2021), the authors in Chaitanya et al. (2020) validated the effectiveness of sampling strategies on contrastive learning for multiple medical MRI segmentation tasks. Furthermore, they suggest that using a sampling strategy that involves cross-image negative sampling can lead to additional performance improvements. Although CL has shown great potential in segmentation tasks, it is important to note that its performance still remains unknown in domain adaptation problems.

2.4. Unsupervised domain adaptation

Unsupervised Domain Adaptation (UDA) is used to generalize learned knowledge from a labeled source domain to an unlabeled target domain. The key challenge of UDA is domain shift, i.e., the inconsistent data distribution across domains, which usually causes performance degradation of models. Early machine learning methods utilized different feature transformations or regularizations to overcome this problem (Kouw & Loog, 2018; Mehrkanoon, 2019; Mehrkanoon & Suykens, 2017).

A number of existing DL methods solve the domain shift problem using adversarial learning or self-training-based approaches. Adversarial learning utilizes generative adversarial networks (GANs) (Goodfellow et al., 2014) to align the distribution of the feature space (Chen, Xie, et al., 2019; Dou et al., 2018; Hoffman et al., 2016; Tzeng, Hoffman, Saenko, & Darrell, 2017) or label space (Dou et al., 2018; Tsai et al., 2018; Vu et al., 2019). In particular, CycleGAN (Kim, Cha, Kim, Lee, & Kim, 2017; Yi, Zhang, Tan, & Gong, 2017; Zhu, Park, Isola, & Efros, 2017) has been extensively explored and adopted in medical image UDA (Chen, Dou, Chen, Qin, & Heng, 2019; Guan & Liu, 2021; Jiang et al., 2018; Zhang, Miao, Mansi, & Liao, 2018) because of its ability to translate the 'style' of the source domain to the target domain in an unpaired way. While CycleGAN-based unsupervised domain adaptation (UDA) methods have shown promising results, they are known to require a large amount of data to learn effective mappings between domains, and can be prone to mode collapse, leading to limited output variations.

Self-training, frequently used in SSL, uses the predictions of the target domain as pseudo-labels and retrains the model iteratively. A typical self-training network (Tarvainen & Valpola, 2017) generates pseudo-labels from a momentum teacher network and distills knowledge to the student network by using consistency loss. The authors in Perone et al. (2019) and Perone and Cohen-Adad (2018) improved the self-training method by aligning the geometrical transformation between the student and teacher networks. DART (Shanis et al., 2019) and MT-UDA (Zhao, Xu, Li, Zeng, & Guan, 2021) combined self-training with adversarial learning in different ways, both receiving promising results. For imbalanced datasets, different denoising methods and sampling strategies have been proposed to improve the quality of pseudo-labels (Hoyer et al., 2022; Xie et al., 2022; Zhang et al., 2021).

Recent self-training approaches, such as those described in Xie et al. (2022) and Zhang et al. (2022), have followed the paradigm of Chaitanya et al. (2020) to align the features, achieved by sampling or merging contrastive features across categories. This demonstrates that the integration of CL can improve the alignment of features at the pixel level. Additionally, the use of a memory bank to expand negative samples has shown to enhance the performance in unsupervised domain adaptation tasks, while enabling training on a normal device. Inspired by the above-mentioned studies, we integrate three kinds of contrastive losses and a category-wise cross-domain sampling strategy to accomplish the UDA segmentation task for breast MRI.

3. Method

3.1. Problem definition

Source domain data and target domain data are two sets of data used in the domain adaptation problem. The source domain data $X_s = \{x_s\}_{i=1}^{M}$ have pixel-level labels whereas the target domain image data $X_t = \{x_t\}_{i=1}^{N}$ are unlabeled. We aim at developing a method that can learn from the labeled source domain and be applied to the target domain. In particular, the learned network is used to classify each pixel of the target domain image into K categories. A direct approach is to train the network in a supervised manner on the source domain and apply it directly to the target domain. However, the performance of the network

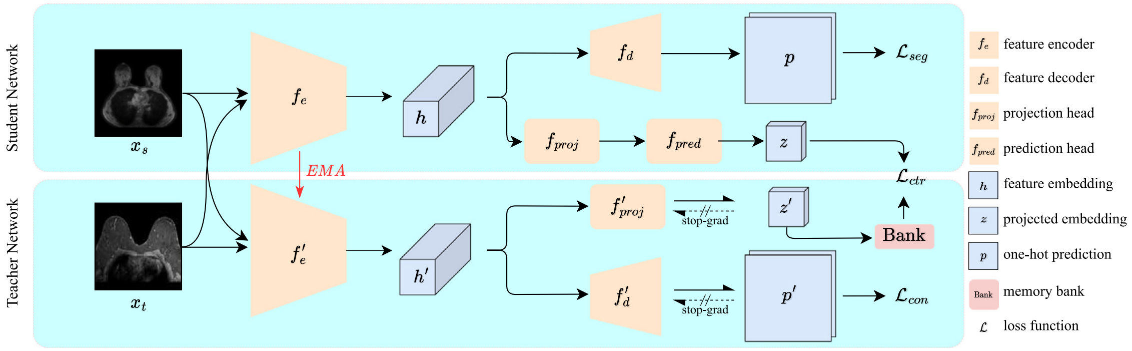


Fig. 1. The Multi-level Semantic-guided Contrastive Domain Adaptation (MSCDA) framework is an unsupervised domain adaptation method that addresses the domain shift problem in breast MRI segmentation by aligning feature representations between labeled source and unlabeled target domains. The framework takes a source image x_s and a target image x_t as inputs to the student and momentum teacher networks, respectively. Each network comprises a segmentation path and a contrast path.

The student network is trained using a supervised segmentation loss, an inter-network consistency loss, and a multi-level contrastive loss, while the teacher network updates the weights using exponential moving average (EMA). The training procedure is detailed in Sections 3.3 and 3.4.

often drops because of the aforementioned domain gap between source and target domains. To address this concern, we propose a new domain adaptation approach, named MSCDA, based on the combination of self-training and contrastive learning.

3.2. Overall framework

The proposed domain adaptation framework is depicted in Fig. 1. It consists of a student network and a momentum teacher network. The student network consists of four main components, a feature encoder f_e, a feature decoder f_d, a projection head f_{proj}, and an additional prediction head f_{pred}. These components are correspondingly mapped in the teacher network with the only exception of the last component (i.e., the prediction head). The three components in the teacher network are called f'_e, f'_d and f'_{pred}. The important notations are listed in Table 1.

In the student network, the feature encoder f_e maps the input MRI image $x \in \mathbb{R}^{H\times W\times 1}$ into a high dimension feature map $h \in \mathbb{R}^{H'\times W'\times C}$. Next, h is transferred into a segmentation probability map $p \in \mathbb{R}^{H\times W\times K}$ and a low dimension feature embedding $z \in \mathbb{R}^{H'\times W'\times D}$ through two forward passes, hereafter referred to as segmentation and contrast paths, respectively. In the first forward pass (segmentation path), the decoder f_d generates the segmentation probability map p of the input h. In the second forward pass (contrast path), the projection head f_{proj} and prediction head f_{pred} jointly reduce the feature map into a low-dimension projected feature embedding $z = f_{pred}(f_{proj}(h))$. Similar steps are conducted in the teacher network, yielding the momentum probability map p' and feature embedding z'. Finally, the probability map p and p' are used for self-training while the projected feature embeddings z and z' are used for semantic-guided contrastive learning to diminish the discrepancy between the two domains. The overall loss function is given by:

$$\mathcal{L} = \mathcal{L}_{seg} + \lambda_1\mathcal{L}_{con} + \lambda_2\mathcal{L}_{ctr}, \tag{1}$$

where $\mathcal{L}_{seg}$ is the supervised segmentation loss, $\mathcal{L}_{con}$ is the consistency loss, $\mathcal{L}_{ctr}$ is the contrastive loss, and λ_1 and λ_2 are the regularization coefficients of the corresponding losses. The summation of segmentation and consistency loss is henceforth referred to as the self-training loss. We elaborate the self-training loss in Section 3.3 and our proposed contrastive loss in Section 3.4.

Table 1
Important notations in our proposed method.

Notations	Description
x_s, x_t, y_s, $\hat{y}_s$	Source image, target image, source image ground truth and corresponding one-hot representation respectively;
p_s, p_t	Student network probability map of the source and target images respectively;
$p_{s'}$, $p_{t'}$	Teacher network probability map of the source and target images respectively;
z_t	Student network feature embedding of the target image;
$z_{s'}$	Teacher network feature embedding of the source image;
$\hat{y}_t$, $\hat{y}_{t'}$	One-hot pseudo-label of p_t and $p_{t'}$ respectively ($y = \arg\max(p)$);
v_s^k, v_t^k	Pixel feature embedding of category k of the source and target images respectively;
c_s^k, c_t^k	Centroid feature embedding of category k of the source and target images respectively;
$\mathcal{Q}_{pixel}$, $\mathcal{Q}_{centroid}$	Pixel queue and centroid queue in the memory bank.

3.3. Self-training

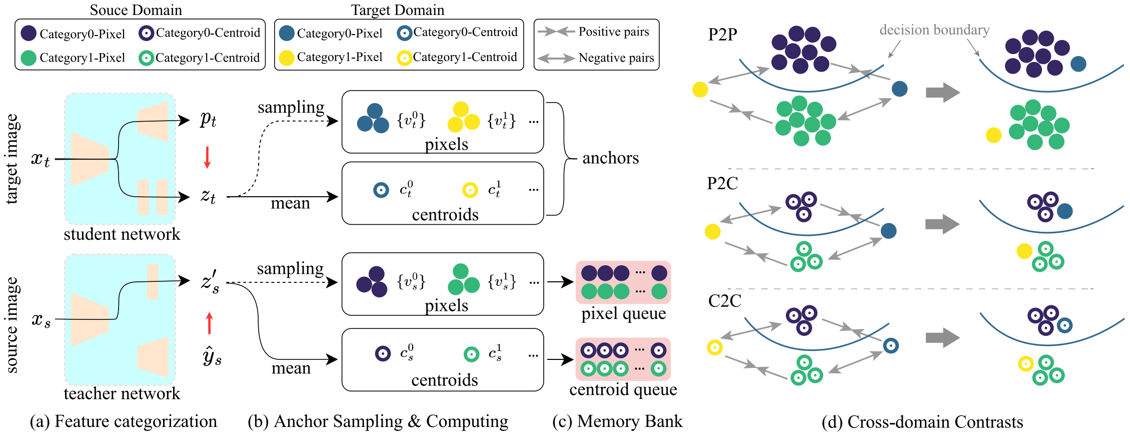
Following the self-training paradigm (Perone et al., 2019), two optimization goals were established. The first goal is to perform supervised learning on the student network from source image labels. The second goal is that the student network learns the pseudo labels generated by the teacher network to distill knowledge from target images. Only the weights in the segmentation path of both networks are updated in this phase.

3.3.1. Supervised learning

In supervised learning, we employ a hybrid segmentation loss (Isensee et al., 2018) that combines Dice loss (Sudre, Li, Vercauteren, Ourselin, & Jorge Cardoso, 2017) and CE loss, and is formulated as:

$$\mathcal{L}_{seg} = \frac{1}{2}\left[\mathcal{L}_{Dice}(p_s, \hat{y}_s) + \mathcal{L}_{ce}(p_s, \hat{y}_s)\right], \tag{2}$$

where $\hat{y}_s$ is the one-hot ground truth and p_s is the probability map of the source domain image in the student network.



(a) Feature categorization (b) Anchor Sampling & Computing (c) Memory Bank (d) Cross-domain Contrasts

Fig. 2. The data flow of our proposed multi-level contrastive loss. (a) Feature categorization: images are fed through the networks and the feature embeddings are categorized based on labels. (b) Anchor sampling and computing: A set of pixel samples is randomly selected and centroids are computed for each category. (c) Memory bank: Pixel and centroid samples are stored in category-wise first-in-first-out queues (d) Cross-domain contrasts: Pixel-to-pixel (P2P), pixel-to-centroid (P2C) and centroid-to-centroid (C2C) contrasts are computed and weighted-summed as our proposed multi-level cross-domain contrastive loss. The contrastive loss aims to bring together samples of the same category, treating them as positive pairs, while separating those from different categories, treated as negative pairs.

3.3.2. Distilling knowledge from pseudo labels

The pseudo label of the target image is generated by the segmentation path in the momentum teacher network iteratively:

$$\hat{y}'_t = \arg\max(p'_t), \tag{3}$$

where p'_t is the probability map of the target domain image in the teacher network. In order to distill knowledge from the pseudo label, an extra consistency loss is added between the two networks. In other words, the target image segmentation p_t generated by the student network is guided by the pseudo label $\hat{y}'_t$. The consistency loss is formulated as:

$$\mathcal{L}_{con} = \frac{1}{H\times W\times K}\sum_{i=1}^{H\times W}\sum_{k=0}^{K-1}\left\| p_t^{(i,k)} - \hat{y}_t^{\prime(i,k)} \right\|^2, \tag{4}$$

where i is the pixel index of the image and k is the category. Here, we update the weights of the student network by means of back propagation. However, in the teacher network, a stop-gradient operation is applied, and the network weights are updated by exponential moving average (EMA):

$$\Theta' \leftarrow \alpha\Theta' + (1-\alpha)\Theta, \tag{5}$$

where Θ and Θ' are the weights of the student network and teacher network respectively, and $\alpha \in (0, 1)$ is the momentum coefficient.

Combining data augmentation with self-training has been shown to improve the domain adaptation performance (Chen, Kornblith, et al., 2020; Tarvainen & Valpola, 2017). The student network receives strongly-augmented images, and the teacher network receives weekly-augmented images during the training process. Random resized cropping is used as the weak augmentation method, and random brightness, contrast and Gaussian blur are used as strong augmentation methods. The strongly-augmented path learns a robust feature representation from the weakly-augmented path that has less disruption.

3.4. Semantic-guided contrastive loss

In order to improve the performance of our UDA framework even further, we incorporate a multi-level semantic-guided contrast to the self-training framework. The idea is to leverage the ground truth of the source domain as supervised signals to enforce the encoder to learn a well-aligned feature representation that mitigates the domain discrepancy. A common way is to categorize the feature embedding and conduct contrastive learning using the pixels or centroids between domains. In our approach, we develop the contrastive loss at P2P, P2C and C2C levels to directly utilize multi-level semantic information to guide the feature alignment. The data flow of our proposed contrastive loss is depicted in Fig. 2.

3.4.1. Preliminaries

In unsupervised contrastive segmentation approaches, the contrast is performed using a randomly selected sample (called the anchor) v, a positive sample v^+ and n negative samples $V^- = \{v_1^-, v_2^-, \ldots, v_n^-\}$. The aim is to learn a feature representation that yields high similarity in positive pairs (v, v^+) and low similarity in negative pairs (v, v^-). Following Chen, Fan, et al. (2020), He et al. (2020) and Zhong et al. (2021), we utilize the InfoNCE as our loss function, which is given as follows:

$$\mathcal{L}_{ctr} = -\log \frac{\exp(v\cdot v^+/\tau)}{\exp(v\cdot v^+/\tau) + \sum_{i=1}^{n}\exp(v\cdot v_i^-/\tau)}, \tag{6}$$

where n is the number of negative samples per anchor, '$\cdot$' is the dot product between two samples, and τ is a temperature hyperparameter that controls the gradient penalty of hard negative samples, which is empirically set to 0.07 (He et al., 2020). Here, samples are selected from D-dimensional feature embedding followed by l_2-normalization.

3.4.2. Feature categorization

Feature categorization is a necessary step required for supervised contrastive learning in the feature space. To utilize the semantic information effectively, we categorize the feature embedding from both domains. For the source image, the feature embedding in the teacher network and its ground truth are required. Given the l_2-normalized target network feature embedding of a source image $z'_s \in \mathbb{R}^{H'\times W'\times D}$ and the one-hot ground truth $\hat{y}_s \in \mathbb{R}^{H\times W\times K}$, we first down-sample the one-hot ground truth into $\bar{y}_s \in \mathbb{R}^{H'\times W'\times K}$ to fit the embedding size, then assign the category label index $k \in \{0, K-1\}$ of $\bar{y}_s$ to each pixel of z'_s (Fig. 2(a)). Similarly, the target image embedding z_t can also be categorized using the pseudo label $\hat{y}_t$. Based on the categorized feature embedding, we further compute the category-wise mean value of pixels of the feature embedding as the centroid $C = \{c^k\}_{k=0}^{K-1}$, which is given as follows:

$$c^k = \frac{1}{\left|\mathbb{Y}^k\right|}\sum_{i=1}^{H'\times W'} \mathbb{1}\left[\bar{y}^{(i,k)} = k\right]\cdot z^i, \tag{7}$$

where $\mathbb{1}[\cdot]$ is an indicator function that returns 1 when the condition holds and 0 otherwise, z^i is the ith pixel of the feature embedding and $\bar{y}^{(i,k)}$ is the down-sampled label which belongs to the ith pixel and category k, $\mathbb{Y}^k$ is the set of labels of category k.

3.4.3. Memory bank & sampling strategy

The adequacy of negative samples plays a critical role in learning feature representations (He et al., 2020). However, the imbalanced ratio between foreground and background pixels in breast MRI segmentation tasks may result in an insufficient number of negative pairs in each batch. To tackle this issue, increasing the batch size or employing a memory bank to save samples from recent batches are ideal solutions. Nevertheless, GPU memory limitations make using a large batch size, such as 1024, impractical for typical devices. Therefore, we adopted the design presented in Wang et al. (2021) and Xie et al. (2022). Specifically, we utilized two category-wise first-in-first-out queues as a memory bank in the teacher network to preserve the pixel and centroid samples extracted from the source images. By using category-wise queues, one for foreground samples and another for background samples, we can save enough negative samples for the contrastive loss, while also ensuring a balanced distribution of samples in each queue. Therefore, we employ a strategy of uniform sampling of a fixed number of pixels from each category in the feature embedding to the pixel queue (Fig. 2(b,c)). This under-sampling approach enables the queue to maintain a sufficient number of balanced pixel samples, while avoiding redundancy. The pixel queue $\mathcal{Q}_{pixel}$ and the centroid queue $\mathcal{Q}_{centroid}$ can be represented as:

$$\mathcal{Q}_{pixel} = \{Q^k_{pixel}\}^{K-1}_{k=0}, \quad Q^k_{pixel} = \{v^k_{(s,i)}\}^{B_p}_{i=1}, \tag{8}$$

$$\mathcal{Q}_{centroid} = \{Q^k_{centroid}\}^{K-1}_{k=0}, \quad Q^k_{centroid} = \{c^k_{(s,i)}\}^{B_c}_{i=1}, \tag{9}$$

where Q^k_{pixel} is the pixel queue of category k, $v^k_{(s,i)}$ is the ith source pixel sample of category k, $Q^k_{centroid}$ is the centroid queue of category k, $c^k_{(s,i)}$ is the ith source centroid sample of category k, and B_p and B_c are the size of the queue respectively.

3.4.4. Pixel-to-pixel contrast

We perform the pixel-to-pixel (P2P) contrastive loss to align the cross-domain feature representation of the same category. To resolve this problem, we first sample m anchors from each category of the **target** feature embedding z_t in the **student** network, denoted as set V^k_t. Then, for each anchor $v^k_t \in V^k_t$ with category label k, we sample a **source** pixel of the same category from the pixel queue $\mathcal{Q}_{pixel}$ to form a positive pair (v^k_t, v^{k+}_s), and sample n **source** pixels of category $q \in \mathbb{K} \setminus \{k\}$ to form n negative pairs (v^k_t, v^{q-}_s). Based on these positive and negative pairs, the InfoNCE loss of a single **target** anchor is computed by using Eq. (6). Overall, the P2P loss is defined as:

$$\mathcal{L}^{P2P}_{ctr} = \frac{1}{\sum^{K-1}_{k=0} \left|V^k_t\right|} \sum^{K-1}_{k=0} \sum_{v^k_t \in V^k_t} \mathcal{L}_{ctr}(v^k_t, v^{k+}_s, V^{q-}_s), \tag{10}$$

where $|\cdot|$ is the number of elements in a set, and V^{q-}_s is the set of negative source pixels. Note that the number of pixels labeled as foreground categories might be less than m (or even 0) if the model predicts a few (or no) breast tissue labels in a mini-batch. Nevertheless, benefiting from the category-wise memory bank, the contrast loss can still be computed even if all pixels in a mini-batch belong to the same category.

3.4.5. Pixel-to-centroid contrast

Due to the under-sampling strategy in selecting anchors and updating the memory bank, the network may suffer from inadequate semantic knowledge and thereby be difficult to converge. This issue is further addressed by incorporating P2C and C2C contrasts to P2P contrast.

For P2C contrast, we force the pixel representation to learn a more general representation with the guidance of the centroid (Wang et al., 2021; Xie et al., 2022). Specifically, a pixel and a centroid from the same category are considered as a positive pair (v^k, c^{k+}), while a pixel and a centroid from different categories are considered as a negative pair (v^k, c^{q-}). We reuse the anchors in Section 3.4.4 and sample all positive and negative centroids from the centroid queue $\mathcal{Q}_{centroid}$. Similar to P2P loss, the P2C loss is defined as:

$$\mathcal{L}^{P2C}_{ctr} = \frac{1}{\sum^{K-1}_{k=0} \left|V^k_t\right|} \sum^{K-1}_{k=0} \sum_{v^k_t \in V_t} \mathcal{L}_{ctr}(v^k_t, c^{k+}_s, C^{q-}_s), \tag{11}$$

where C^{q-}_s is the set of negative source centroids.

3.4.6. Centroid-to-centroid contrast

For C2C contrast, the ideal situation is that the centroids from the same category are located near to one another, whereas centroids from other categories are located far apart. Unlike P2C contrast, the total number of centroids p $(BK \leq p \leq 2BK)$ is much smaller than the pixel number in a mini-batch. Besides, calculating centroids is computationally efficient. Therefore, the centroids of the whole mini-batch can be fully involved as anchors in C2C contrast. Similar to P2P and P2C contrast, the positive pairs (c^k, c^{k+}) and negative pairs (c^k, c^{q-}) are defined according to whether centroids are from the same category. Thus, the C2C loss is defined as:

$$\mathcal{L}^{C2C}_{ctr} = \frac{1}{\sum^{K-1}_{k=0} \left|C^k_t\right|} \sum^{K-1}_{k=0} \sum_{c^k_t \in C_t} \mathcal{L}_{ctr}(c^k_t, c^{k+}_s, C^{q-}_s), \tag{12}$$

where C_t is the set of target centroid anchors.

Finally, we take the weighted sum of the three above-mentioned contrasts (Fig. 2(d)) as our proposed multi-level semantic-guided contrastive loss:

$$\mathcal{L}_{ctr} = \lambda_{P2P}\mathcal{L}^{P2P}_{ctr} + \lambda_{P2C}\mathcal{L}^{P2C}_{ctr} + \lambda_{C2C}\mathcal{L}^{C2C}_{ctr}, \tag{13}$$

where λ_{P2P}, λ_{P2C} and λ_{C2C} are the regularization coefficients of the corresponding contrasts. The overall training process of our proposed MSCDA is presented in Algorithm 1.

4. Experiments

4.1. Datasets

Dataset 1. Dataset 1 consists of test-retest breast T1-weighted (T1 W) and T2-weighted (T2 W) MRI images and corresponding right-breast masks of eleven healthy female volunteers, which is described in Granzier et al. (2022). The images of each subject were collected in two separate sessions (interval<7 days), during which three 3D scans were collected. Subjects were asked to lay in the prone position and remain still in the MRI scanner while both modalities are sequentially acquired. All images were acquired with an identical 1.5T MRI scanner (Philips Ingenia, Philips Healthcare, Best, the Netherlands) using a fixed clinical breast protocol without contrast. The detailed acquisition parameters are listed in Table 2. In pre-processing, we first resize all MRI slices and corresponding masks to 256 × 256 pixels using cubic interpolation and nearest-neighbor interpolation respectively, and then normalize images with z-score transformation. In total, dataset 1 contains 14520 (11 subjects × 2 sessions × 3 scans × 220 slices) T1 W slices and 11220 (11 subjects × 2 sessions × 3 scans × 170 slices) T2 W slices.

Table 2
Dataset description and acquisition parameters of Dataset 1 (healthy volunteers) and Dataset 2 (patients with invasive breast cancer). Note the differences in acquisition parameters, imaging sequences and clinical conditions, highlighting the domain gap between the two datasets.

	Subject	Type	Scanner	Sequence	Acquisition parameters			
	number				TR (ms)	TE (ms)	PS (mm)	ST (mm)
Dataset 1	11	Healthy volunteers	Philips 1.5T (Ingenia)	T1 W	5.3	3	0.36 × 0.36	2
				T2 W	2000	223	0.79 × 0.79	2
Dataset 2	134	Patients with invasive breast cancer	Philips 1.5T (Ingenia/Intera)	DCE-T1 W	6.5–7.6	2.9–3.5	0.85 × 0.85- 0.97 × 0.97	1
				T2 W	2000	170–259	0.65 × 0.65- 0.97 × 0.97	1

Abbreviations: TR = Repetition time; TE = Echo time; PS = Pixel spacing; ST = Slice thickness; T1W = T1-weighted; T2W = T2-weighted; DCE = Dynamic contrast-enhanced; T = Tesla.

Algorithm 1: MSCDA for Breast MRI

```
Input: Source domain image $x_s$ and label $y_s$; Target
       domain image $x_t$;
Initialize the weights of the student network $\Theta_e$, $\Theta_d$ with
  pre-trained weights, $\Theta_{proj}$ and $\Theta_{pred}$ via (He, Zhang, Ren,
  & Sun, 2015). Initialize the teacher network by copying
  weights from the student network and applying
  stop-gradient; Initialize the memory bank $\mathcal{Q}_{pixel}$ and
  $\mathcal{Q}_{centroid}$ ;
for epoch = 1, $E_{max}$ do
    foreach mini-batch do
        Apply weak and strong data augmentation;
        Forward propagate weak-augmented batch in the
          student network to get $p_s$, $p_t$ and $z_t$;
        Forward propagate strong-augmented batch in the
          teacher network to get $p_t'$ and $z_s'$;
        Compute loss $\mathcal{L}_{seg}$ using $p_s$ and $y_s$ via Eq. (2);
        Compute loss $\mathcal{L}_{con}$ using $p_t$ and $p_t'$ via Eq. (4);
        Categorize the feature embedding $z_s'$ and $z_t$;
        foreach category do
            Sample pixel anchors and compute centroid
              anchors from $z_t$;
            Sample corresponding positive and negative
              pairs from $\mathcal{Q}_{pixel}$ and $\mathcal{Q}_{centroid}$;
            Update $\mathcal{Q}_{pixel}$ and $\mathcal{Q}_{centroid}$ using $z_s'$;
        end
        Compute loss $\mathcal{L}_{ctr}^{P2P}$, $\mathcal{L}_{ctr}^{P2C}$ and $\mathcal{L}_{ctr}^{C2C}$ via Eq. (10)–(12)
          respectively;
        Update the student network via Eq. (4);
        Update the teacher network by Eq. (5);
    end
end
Output: Weights of the student network $\Theta_e$ and $\Theta_d$.
```

Dataset 2. Dataset 2 consists of the images from 134 subjects with histologically confirmed invasive breast cancer imaged between 2011 and 2017 in Maastricht University Medical Center+ and collected retrospectively (Granzier et al., 2021, 2020). The images contain breast dynamic contrast-enhanced T1 W (DCE-T1 W) and T2 W MRIs and corresponding right-breast masks. Similar to Dataset 1, each subject underwent the examinations with 1.5T MRI scanners (Philips Intera and Philips Ingenia (idem)) in a prone position. In particular, DCE-T1 W images were acquired before and after the intravenous injection of gadolinium-based contrast Gadobutrol (Gadovist, Bayer Healthcare, Berlin, Germany (EU)) with a volume of 15 cc and a flow rate of 2 ml/s. The acquisition parameters are also listed in Table 2. We conduct the same image pre-processing as in Dataset 1. In total, Dataset 2 contains 21793 T2 W and 28540 T1 W slices and they are split into three folds with 45, 45 and 44 subjects for the cross-validation depicted in Section 4.2.

4.2. Experiment setup

As shown in Table 2, the subject population, machine vendor and acquisition parameters between the two datasets are heterogeneous, indicating the common domain shift problem in clinical practice. In particular, T1 W and T2 W are two different types of MRI sequences, with T1 W images typically used for observing anatomical structures, while T2 W images provide information on tissue composition. In breast MRI, T1 W images help identify the location and size of lesions, while T2 W images can detect edema or inflammation (Mann, Cho, & Moy, 2019).

We set up the experiment on both Dataset 1 and 2 to transfer the knowledge of breast segmentation from healthy women to patients. Specifically, the experiment consists of two scenarios: (1) T2W-to-T1 W: utilizing the T2 W images of Dataset 1 as the source domain and the T1 W images of Dataset 2 as the target domain; (2) T1W-to-T2 W: utilizing the DCE-T1 W images of Dataset 1 as the source domain and the T2 W images of Dataset 2 as the target domain. In each scenario, we establish three tasks with a different number of subjects in the source domain to validate the label-efficient learning ability of our framework. The three tasks contain four, eight and eleven (i.e., the whole dataset) randomly selected subjects respectively, and are denoted as S4, S8 and S11. To further verify the robustness of UDA performance, we split the target domain into three folds to perform a three-fold cross-validation. In each run of the cross-validation, two folds are used as the target domain for training and the remaining fold for testing.

4.3. Model evaluation

The DSC is used as the main evaluation metric. Additionally, we use the Jaccard Similarity Coefficient (JSC) as well as precision (PRC) and sensitivity (SEN) as auxiliary evaluation metrics. These metrics are formulated as follows:

$$\mathrm{DSC} = \frac{2 \times TP}{2 \times TP + FP + FN} \times 100\%, \tag{14}$$

$$\mathrm{JSC} = \frac{TP}{TP + FP + FN} \times 100\%, \tag{15}$$

$$\mathrm{PRC} = \frac{TP}{TP + FP} \times 100\%, \tag{16}$$

$$\mathrm{SEN} = \frac{TP}{TP + FN} \times 100\%, \tag{17}$$

where TP, FP and FN are the number of true positive, false positive and false negative pixels of the prediction respectively. Note that we show the mean value of each metric of the three-fold cross-validation.

Table 3
Evaluation results of the proposed MSCDA framework compared with source-only, supervised training, and two other UDA methods (i.e. CyCADA and SEDA). The table shows the dice similarity coefficient (DSC), Jaccard similarity coefficient (JSC), precision (PRC), and sensitivity (SEN) of the methods. The best performance in each metric is shown in bold. The results demonstrate that MSCDA outperforms the other methods in most of the evaluated metrics, highlighting its effectiveness in addressing the domain shift problem for breast MRI segmentation. Additionally, MSCDA shows high label-efficient learning ability, with a DSC that remains relatively stable across different tasks.

Method	Backbone	Task	Scenario 1: T2 W to T1W				Scenario 2: T1 W to T2W			
			DSC (%)	JSC (%)	PRC (%)	SEN (%)	DSC (%)	JSC (%)	PRC (%)	SEN (%)
Src-Only	U-Net	S11	65.8	56.7	70.7	79.8	74.3	63.7	88.9	69.3
		S8	64.0	53.9	82.7	67.8	74.2	65.1	87.1	71.9
		S4	21.0	16.0	96.5	17.4	60.7	45.2	96.3	46.0
	DeepLab v3+	S11	71.9	58.4	83.1	69.2	70.0	58.0	90.5	63.7
		S8	69.1	56.1	90.9	61.8	74.3	65.4	88.5	73.4
		S4	54.9	41.3	94.1	44.3	70.3	57.2	95.7	60.0
Supervised	U-Net	–	94.8	91.7	94.4	94.8	95.7	92.7	96.9	95.5
	DeepLab v3+	–	95.8	92.8	98.0	94.7	96.0	93.0	96.2	96.5
CyCADA	U-Net	S11	78.7	68.8	86.9	79.4	79.7	69.3	90.5	75.6
		S8	77.0	66.4	83.5	79.8	78.5	66.8	91.5	72.1
		S4	54.5	42.6	94.7	45.1	63.0	49.7	97.8	50.7
	DeepLab v3+	S11	80.0	68.0	78.6	86.0	73.8	61.3	85.8	70.6
		S8	77.2	64.5	81.1	79.2	70.3	59.0	92.1	64.5
		S4	64.0	50.4	92.4	54.1	67.6	53.8	92.8	57.4
SEDA	U-Net	S11	79.0	67.1	81.4	82.6	81.2	70.4	96.2	72.7
		S8	79.4	70.0	83.3	84.3	80.2	70.3	92.0	75.0
		S4	69.0	56.1	93.4	60.0	73.5	60.8	**98.9**	61.3
	DeepLab v3+	S11	81.7	70.6	88.6	79.2	82.5	73.7	94.0	78.5
		S8	80.3	68.4	88.0	77.9	82.4	71.4	83.9	83.2
		S4	71.4	57.9	**95.4**	60.1	75.5	62.5	98.5	63.5
MSCDA	DeepLab v3+	S11	88.6	79.9	86.5	**92.3**	83.1	71.8	88.7	**79.5**
		S8	**89.2**	**81.0**	89.3	89.9	**84.0**	**73.2**	91.7	78.8
		S4	87.2	78.0	92.4	83.6	83.4	72.5	98.0	73.8

4.4. Implementation details

4.4.1. Architecture

Encoder & decoder. We conduct our experiment by adopting DeepLab-v3+ (Chen et al., 2018) with ResNet-50 (He et al., 2016) as backbone. Benefiting from the encoder–decoder architecture, the encoder and decoder of DeepLab-v3+ are adopted in our framework. Specifically, the hidden dimension of ResNet-50 is set to $(16, 32, 64, 128)$, yielding a 512-dimension feature map.

Projection/prediction head. The projection head f_{proj} is a shallow network that contains two 1×1 convolutional layers with BatchNorm and ReLU. It projects the 512-d feature map into a 128-dimension l_2-normalized feature embedding. The prediction head f_{pred} shares the same architecture setting with f_{proj} with the exception that the f_{pred} does not change the dimension of the features.

Memory bank. The size of the pixel queue and the centroid queue of each category are set to 4096 and 1024, respectively. In each mini-batch, we randomly sample eight pixels per category of each feature embedding to the queue and discard the oldest samples. The number of pixel anchors for P2P loss is set to 32, the number of negative pairs of P2P contrast is set to 4096, which is equivalent to the size of the pixel queue, the number of negative pairs of P2C and C2C contrasts is set to 1024, which is equivalent to the size of the centroid queue. The regularization coefficients in Eqs. (1) and (13) are all set to 1 by default.

4.4.2. Training settings

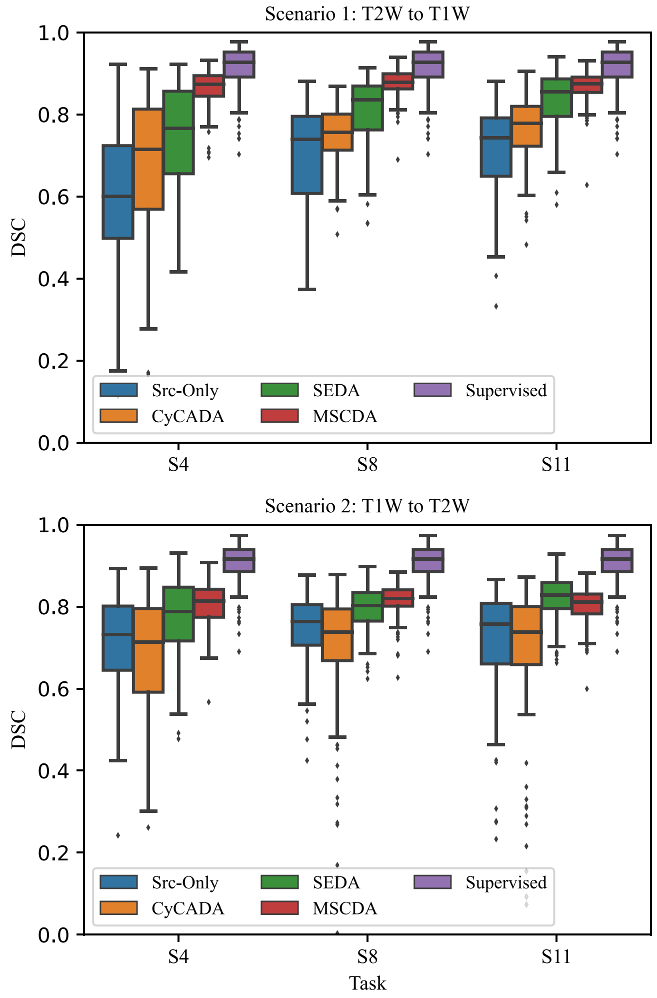
To accelerate the training procedure, we pre-train the DeepLab-v3+ on the source domain and then use the weights to initialize the encoder f_e and decoder f_d of our UDA framework. Additionally, the projection and prediction heads are initialized by He et al. (2015). The Adam (Kingma & Ba, 2014) optimizer is used for training the framework for $E_{max} = 100$ epochs with a fixed learning rate of 0.01, batch size 24. Note that only f_e and f_d participate in inference, while f_{proj}, f_{pred}, f'_e, f'_d, f'_{proj} and $\mathcal{Q}_{p/c}$ are discarded after training. All networks are implemented based on Python 3.8.8 and Pytorch 1.7.1 and are trained on an NVIDIA GeForce GTX 2080Ti GPU.

5. Results

5.1. Quantitative comparison with other start-of-art approaches

The performance of our proposed MSCDA is depicted in Table 3 and Fig. 3. We compared our proposed method with two state-of-art UDA approaches: CyCADA (Hoffman et al., 2018) using adversarial learning methods and SEDA (Perone et al., 2019) using self-training methods which are frequently used for medical images. Additionally, the two selected methods were both trained with two different domain labels, i.e. source domain labels (denoted as "Src-Only") and target domain labels (denoted as "Supervised"). In summary, we compare MSCDA to four methods and each has two different types of backbones (U-Net (Ronneberger et al., 2015) or DeepLab v3+ (Chen et al., 2018)), yielding eight combinations. Note that plain U-Net is not applicable for our method because the very small (e.g., 8×8) resolution in latent space leads to the inaccurate classification of embeddings.

The influence of domain shift on the performance of segmentation models can be quantified by comparing the DSC between the supervised and Src-Only methods. For instance, in T2W-to-T1 W scenario Task S4 with DeepLab v3+ as the backbone, the supervised method achieved a DSC of 95.8%, while the Src-Only method only reached 54.9%, resulting in a performance degradation of 40.9%. Similarly, in Task S8, and S11, Src-Only experienced a performance loss of 25.8% and 17.1%, respectively, compared to the supervised method. On the other hand, Fig. 3 also shows the performance degradation at a subject level. The medians of the supervised method show a significant DSC increase compared to


Scenario 1: T2W to T1W

Scenario 2: T1W to T2W

Fig. 3. The box plot comparison of the DSC between our proposed MSCDA and other methods. All methods are equipped with DeepLab v3+ as the backbone. The plots show the distribution of model performance at a subject level. The DSC of each subject is the mean value of all slices containing foreground pixels.

Src-Only. Meanwhile, Src-Only demonstrates a larger interquartile range (IQR) than the supervised method, indicating a wider distribution of DSC across subjects.

After applying UDA methods, MSCDA outperforms the other examined methods under the same task. More specifically, the DSC reaches over 83% in task S4 in both T2W-to-T1 W and T1W-to-T2 W scenarios (T2W-to-T1 W: 87.2%, T1W-to-T2 W: 83.4%), while the DSC of other methods are below 76% (e.g., CyCADA, T2W-to-T1 W: 64.0%, T1W-to-T2 W: 67.6%; SEDA, T2W-to-T1 W: 71.4%, T1W-to-T2 W: 75.5%). This result is supported by other evaluation metrics, such as JSC and SEN. As it can be seen in the bottom of Table 3, in both scenarios, MSCDA achieved better results in all evaluated metrics except in PRC although it reaches over 92%. For the other two tasks (S8 and S11), the proposed method in general outperforms other approaches. The box plot (see Fig. 3) also indicates that MSCDA method not only performs better but also has a smaller IQR than Src-Only and the other two methods.

From Table 3, one can observe that when comparing the performance between different tasks (i.e., S11, S8 and S4), MSCDA shows high label-efficient learning ability. More precisely, the DSC of our methods in T2W-to-T1 W scenarios only drops 2.0% from 89.2% to 87.2% while CyCACA and SEDA drop 16.0% and 10.3% respectively; The DSC of our method in T1W-to-T2 W scenario remains relatively stable across three tasks with the difference of 0.9% across tasks. Compared to our model, the performance of other methods drops significantly as the number of the source subjects decreases. Therefore, the obtained results show that our method is less sensitive to the size of source domain compared to other UDA methods. Notably, the performance of our method is very close to that of supervised learning (MSCDA: DSC = 89.2%, JSC = 81.0%, PRC = 89.3% SEN = 89.9%; Supervised: DSC = 95.8%, JSC = 92.8%, PRC = 98.0%, SEN = 94.7%) when training with the eight source subjects (task S8) in T2W-to-T1 W scenario, demonstrating the potential of contrastive representation learning and self-training framework.

Table 4
Ablation study of each proposed component on scenario 1 task S4 fold 1. A check mark indicates that a specific component is applied. The DSC is utilized to evaluate the performance and the extra points gained compared to the baseline (case 1) are listed. The best-performed combination is highlighted in bold.

Case	Self-training	Aug.	P2P	P2C	C2C	DSC (%)	Gain (%)
1	✓					50.3	
2	✓	✓				71.6	+21.3
3	✓	✓	✓			80.2	+29.9
4	✓	✓		✓		76.0	+25.7
5	✓	✓			✓	67.3	+17.0
6	✓	✓		✓	✓	79.1	+28.8
7	✓	✓	✓	✓		81.6	+31.3
8	✓	✓	✓		✓	81.5	+31.2
9	✓	✓	✓	✓	✓	**82.2**	**+31.9**

5.2. Qualitative segmentation comparison with other start-of-art approaches

To help qualitatively better understand the performance of models, we plot the segmentation results and corresponding uncertainty maps in Fig. 4. The uncertainty map reflects the confidence level of the model to each pixel, which is generated by test-time dropout (Loquercio, Segu, & Scaramuzza, 2020) with Monte Carlo simulation number equals to 20. In Fig. 4 T2W-to-T1 W scenario, the performance degradation of Src-Only is mainly manifested in a large number of under-segmented regions, and it has high uncertainty at the boundary of segmentation results and low uncertainty in under-segmented regions. Applying SEDA and CyCADA can alleviate the under-segmented regions, where the uncertainty area is reduced in SEDA while it still remains in CyCADA. MSCDA is able to generate segmentations that closely resemble the supervised model and which covered more under-segmented areas in SEDA and CyCADA. Meanwhile, the uncertainty in MSCDA occurs mainly close to the pectoral muscles, which is more difficult to segment that the breast-air boundary. In the T1W-to-T2 W scenario, however, we observed some under-segmented regions near the breast-air boundary, which is likely attributable to the substantial difference between the marginal fat and FGT tissue in T2 W images. This difference probably makes it challenging to align the feature space of fat with the source T1 W images.

5.3. Ablation study

5.3.1. Effect of loss function & augmentation

In order to investigate the contribution of augmentation and different loss function, we conduct an ablation experiment by removing/adding each component separately. We test the network on scenario 1 task S4 fold 1 with combinations of self-training, data augmentation, P2P, P2C and C2C contrast. All the networks are trained under the same experimental settings as Section 4.4. As illustrated in Table 4, adding data augmentation (see case

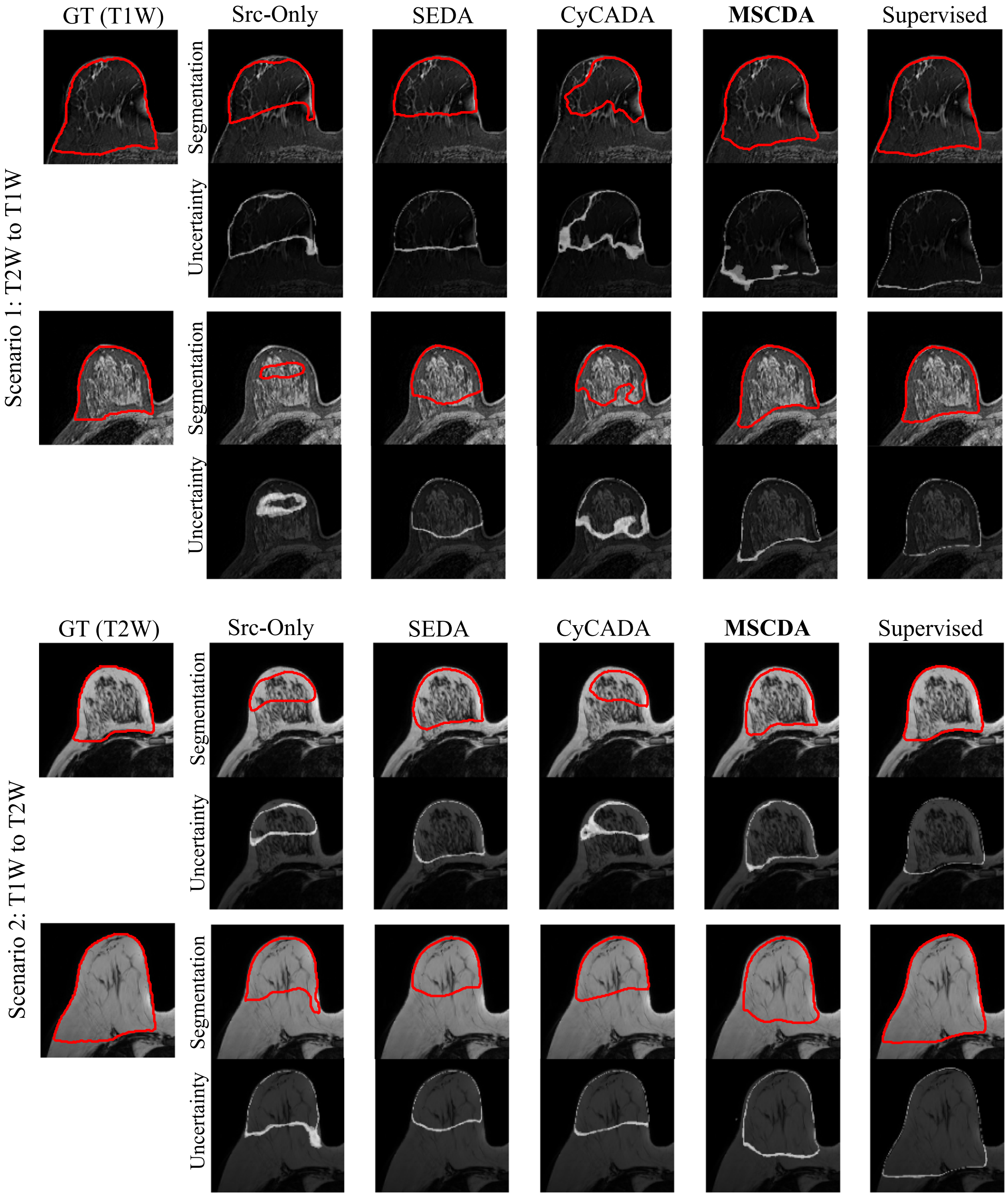


Fig. 4. Examples of segmentation results for Scenario 1/2 Task S4 using our proposed Multi-level Semantic-guided Contrastive Domain Adaptation (MSCDA) and previous methods. The leftmost subplot in each scenario shows the ground truth (GT), followed by model predictions from the Src-Only, SEDA, CyCADA, MSCDA, and supervised training, respectively. The segmentation results are visualized as red contours, and the corresponding uncertainty map is presented below each subplot. Intensities in the uncertainty map signify the degree of uncertainty, with higher values indicating greater uncertainty. All methods utilize DeepLab v3+ as the backbone.

2) to self-training can increase the DSC by 21.3% compared to plain self-training (see case 1). Combining case 2 with P2P (see case 3) or P2C (see case 4) contrast increase the DSC to 80.2% and 76.0% respectively. However, when adding C2C contrast into case 2 (see case 5), the network performance deteriorates to a DSC of 67.3%, indicating centroid-level contrastive learning does not benefit feature embeddings in our breast segmentation task. Nonetheless, this shortcoming is canceled out by adding P2P or P2C contrast, as shown in case 6 and 8. This indicates that C2C contrast is not as effective as P2P or P2C contrast in our breast segmentation task. When integrating all contrasts together (see case 9), the DSC reaches highest score of 82.2%, an increment of 31.9% compared to the simple case 1. Overall, by adding data augmentation, P2P, P2C, and C2C contrasts, MSCDA can improve the self-training framework to achieve better segmentation performance. However, we also find that not all types of contrasts are equally effective. Hence, we performed an ablation study on the regularization coefficients of three contrasts in Section 5.3.2.

5.3.2. *Effect of coefficients between contrasts*

To investigate the effectiveness of different contrast coefficients, we conduct an ablation study by varying the regularization

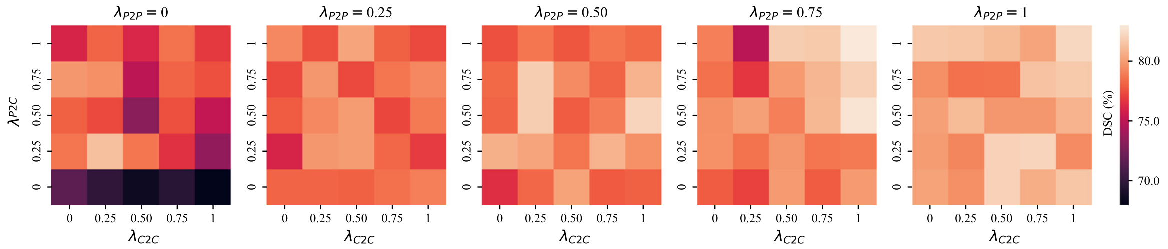


Fig. 5. Heatmaps illustrating the impact of different coefficient combinations of pixel-to-pixel (P2P), pixel-to-centroid (P2C), and centroid-to-centroid (C2C) contrasts on the segmentation performance. The ablation study involved varying each of the coefficients from 0 to 1 and utilizing DSC for evaluation.

Table 5
Ablation study of different combinations of coefficients of consistency loss and contrastive loss. λ_1 is the coefficient of the consistency loss and λ_2 is the coefficient of the contrastive loss. DSC is utilized to evaluate the performance. The best-performed combinations ($\lambda_1 = 0.5/1$, $\lambda_2 = 1$) are highlighted in bold.

DSC	$\lambda_2 = 0.2$	0.5	1.0	2.0	5.0
$\lambda_1 = 0.2$	79.2	80.9	81.6	80.1	80.5
$\lambda_1 = 0.5$	75.8	79.3	**82.2**	80.4	77.5
$\lambda_1 = 1.0$	75.8	77.9	**82.2**	77.4	78.1
$\lambda_1 = 2.0$	74.8	80.1	81.2	77.7	78.6
$\lambda_1 = 5.0$	78.8	79.0	79.4	75.6	77.4

coefficients of each contrast in Eq. (13) from 0 to 1. As shown in Fig. 5, we observed that increasing λ_{P2P} can improve model performance. However, changes in λ_{P2C} and λ_{C2C} may have varying effects depending on the value of λ_{P2P}. Specifically, when λ_{P2P} and λ_{P2C} are both 0, changes in λ_{C2C} do not significantly improve performance, which suggests that the C2C contrast may not be effective without incorporating other two contrasts. We also observe that increasing λ_{P2C} and λ_{C2C} could improve model performance only when λ_{P2P} was set to a large value (i.e., 0.75 or 1). This finding implies that P2C and C2C may be more effective when the P2P contrast is heavily weighted. We also observed several equally sub-optimal combinations when λ_{P2P} is set to 1, which indicates that there may be multiple ways to achieve optimal performance. Therefore, coefficients are set to 1 as the default in our training settings. Our result is consistent with the findings of Alonso, Sabater, Ferstl, Montesano, and Murillo (2021), which showed that increasing the weight of the P2P contrast from a low value can lead to improved performance in a similar semi-supervised setting. Moreover, our study provides additional insights into the sensitivity of the model's performance to different coefficient combinations of contrasts.

5.3.3. Effect of coefficients between consistency loss and contrastive loss

We also conduct the ablation study of the coefficients in Eq. (1) to investigate the best combination of consistency loss (λ_1) and contrastive loss (λ_2). Table 5 shows that setting λ_1 to 0.5 or 1 with λ_2 set to 1 achieves the best performance, with a DSC of 82.2%. It is worth noting that setting λ_1 to a smaller value (e.g. $\lambda_1 = 0.2$) can still result in relatively good performance, with a DSC of around 80%. However, setting λ_1 and λ_2 to larger values can lead to a decrease in performance. Overall, the study shows the importance of finding the appropriate balance between the consistency and contrastive losses in UDA tasks.

5.3.4. Effect of contrast between domains

As mentioned in Section 3.4, we compute three types of contrasts between the student and teacher networks. In particular, only the **target** feature embeddings in the **student** network are sampled as anchors, while only the **source** feature embeddings in the **teacher** network are sampled to update the memory bank.

Table 6
Ablation study of contrast between domains on scenario 1 task S4 fold 1. A check mark indicates that a specific component is applied. The DSC is utilized to evaluate the performance and the extra points gained compared to the lowest value (case 1) are listed. The best performed combination is in bold.

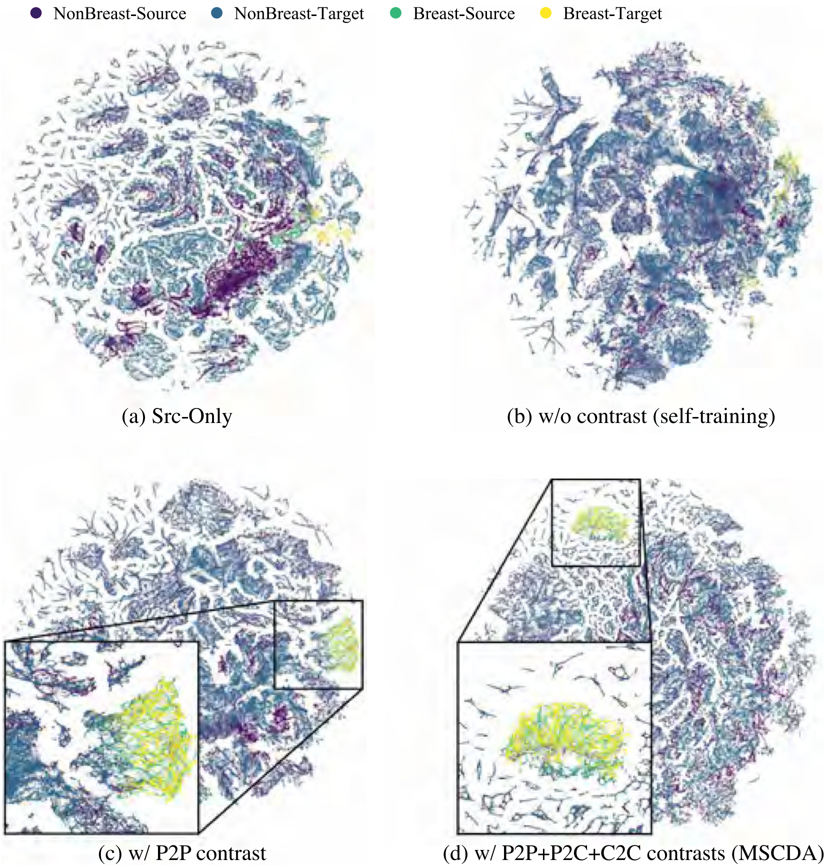
Case	Student network (anchor)		Teacher network (queue)		DSC (%)	Gain (%)
	Source	Target	Source	Target		
1	✓	✓	✓	✓	77.0	+0.5
2	✓	✓	✓		76.5	
3	✓	✓		✓	77.5	+1.0
4	✓		✓	✓	78.1	+1.6
5		✓	✓	✓	82.0	+5.5
6	✓			✓	78.7	+2.2
7		✓	✓		**82.2**	+5.7

To further elaborate our selection, we conduct an additional, complementary ablation study by selecting different domains for computing contrast. Note that all other experimental settings remained unchanged.

As shown in Table 6, we observe that the best candidate (see case 7, DSC = 82.2%) is the combination of the target samples in the student network and the source sample in the teacher network. More specifically, we adopt **source** samples from the teacher network to create the memory bank and to guide the **target** samples from the student network. As expected, when adding target samples to the memory bank (see case 5), the performance shows a minor decrease of 0.2%, indicating that the pseudo label brings uncertainty to the model. It is worth noticing that we observe 5.7% of degradation when adopting additional source samples as anchor (see case 2). It might be due to the overfitting of the model on the source domain.

5.3.5. Effect of size of memory bank/negative samples

The size of the memory bank is a critical factor in our proposed contrastive learning method since it determines the negative pairs in P2P, P2C, and C2C contrasts. To investigate its effect, we conducted an ablation study on scenario 1 task S4 fold 1, where the size of the pixel queue B_p and centroid queue B_c were varied from 512 to 8192 and from 32 to 4096, respectively. As presented in Table 7, the model's performance generally improved with an increase in the sizes of B_p and B_c. The best-performing combinations were $B_p = 4096$ and $B_c = 1024$ or $B_p = 2048$ and $B_c = 2048$, achieving a DSC score of 82.2%. However, the performance improvement reached a saturation point or declined after a certain value. This could be due to the excessive number of negative samples causing the model to suffer from collision-coverage (Ash, Goel, Krishnamurthy, & Misra, 2021). To avoid hurting the representation learning quality, Ash et al. (2021) and Awasthi, Dikkala, and Kamath (2022) suggest that an appropriate trade-off should be made in selecting the number of negative pairs. We, therefore, chose $B_p = 4096$ and $B_c = 1024$ as the default settings for our training. In conclusion, a sufficiently large memory bank is crucial for improving the model's performance,



(a) Src-Only

(b) w/o contrast (self-training)

(c) w/ P2P contrast

(d) w/ P2P+P2C+C2C contrasts (MSCDA)

Fig. 6. t-SNE visualization of the pixel representations on scenario 1 task S4. Each colored point indicates a categorized pixel representation in the high dimension feature map. Note that we only partially visualize the testing images of the target domain due to the large dataset size. All methods are equipped with DeepLab v3+ as the backbone.

Table 7
Ablation Study of the memory bank/negative samples on scenario 1 task S4 fold 1. The size of pixel queue B_p and centroid queue B_c are varied from 512 to 8192 and from 32 to 4096, respectively. The model performance is evaluated by DSC. The best-performed combinations ($B_p = 4096$, $B_c = 1024/B_p = 2048$, $B_c = 2048$) are highlighted in bold.

DSC (%)	$B_p = 512$	1024	2048	4096	8192
$B_c = 32$	76.9	78.3	80.0	80.8	81.3
$B_c = 128$	78.1	78.5	80.6	80.4	81.3
$B_c = 512$	78.3	78.9	80.3	82.1	80.0
$B_c = 1024$	78.9	79.8	81.0	**82.2**	81.3
$B_c = 2048$	78.7	79.4	**82.2**	79.7	80.4
$B_c = 4096$	79.5	78.9	78.2	80.8	78.2

but increasing its size beyond a certain limit can lead to diminishing returns due to the collision-coverage trade-off in our tasks.

5.4. Analysis of feature alignment

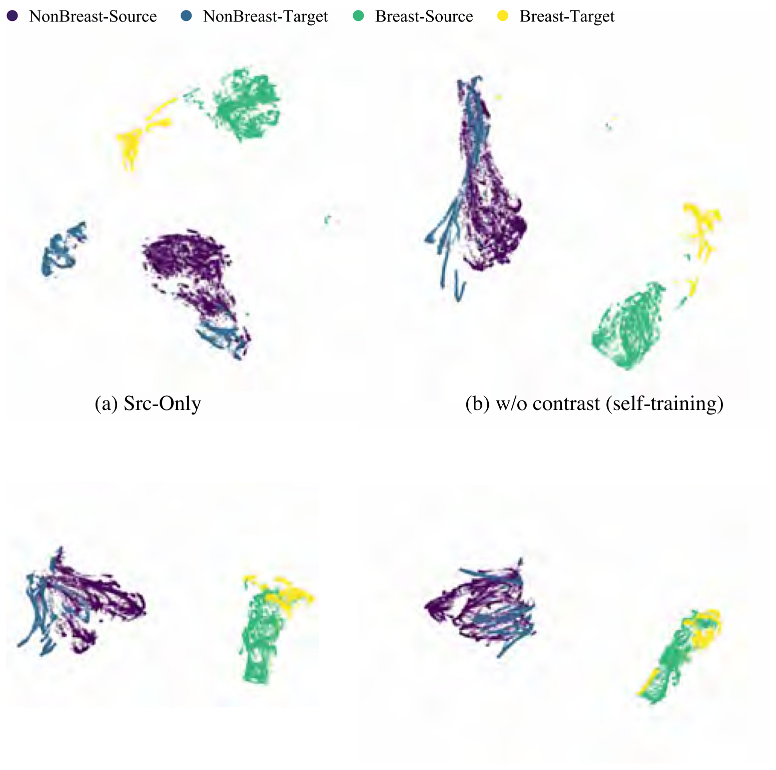
5.4.1. Visualization of feature alignment

To visualize the effect of our proposed method on domain shift, we plot the learned features from the source and target testing images with t-SNE (Van der Maaten & Hinton, 2008). The learned features are obtained by using DeepLab v3+ (Chen et al., 2018) as the backbone. At the pixel level (Fig. 6), when no domain adaptation method is applied, the breast pixels of Src-Only highly overlap with non-breast pixels (Fig. 6(a)), making them indistinguishable. Compared to Src-Only, the self-training (Fig. 6(b)) makes it possible to align part of the breast pixels between domains but fails to separate them from non-breast pixels. Incorporating P2P contrast (Fig. 6(c)) highly aligns the breast pixels; however, a number of breast pixels are contaminated by non-breast pixels which may increase the error. In contrast to the above-mentioned methods, our method nicely aligns the breast pixels and separates them from non-breast pixels.

The visualization of the centroid level in Fig. 7 further illustrates the effect of our method on the feature space. Compared to the pixel level, the uneven distribution caused by the imbalanced dataset is alleviated at the centroid level, making the visualization clearer. We can observe that the learned centroids of different categories in all methods are linearly separable. Before self-training, the centroids of the same category are completely separable by domain, as can be observed in Fig. 7(a). When self-training is applied (Fig. 7(b)), the non-breast centroids are clustered together while the breast centroids are still not aligned. The P2P contrast (Fig. 7(c)) improves the centroid alignment between domains but is still not fully overlapped. In our method (Fig. 7(d)), the centroids of the same category share a well-aligned tight representation space. In summary, the t-SNE visualization demonstrates the effect of domain shift in the feature space, an effect that can be mitigated by applying our method.

5.4.2. Quantitative analysis of feature alignment

To further quantitatively analyze the feature alignment in our proposed MSCDA, we employ the use of cluster centroid distance variation (CCD) (Luo et al., 2021), which measures the distance between the distributions of feature embeddings in the source and target domains. We first obtain feature embeddings of each domain in Section 5.4.1 separately and then calculate the CCD between the centroids of each category. Moreover, we



(a) Src-Only

(b) w/o contrast (self-training)

(c) w/ P2P contrast

(d) w/ P2P+P2C+C2C contrasts (MSCDA)

Fig. 7. t-SNE visualization of the centroid representations on scenario 1 task S4. Each colored point indicates a categorized centroid representation in the high dimension feature map. All testing images of the target domain are included in the visualization. All methods are equipped with DeepLab v3+ as the backbone.

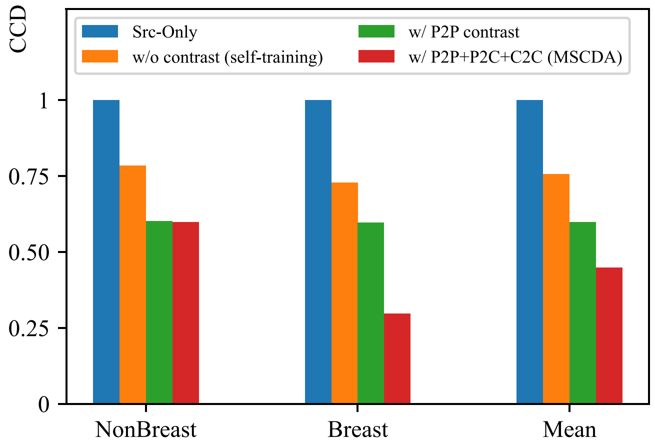


Fig. 8. Quantitative analysis of feature alignment using cluster centroid distance variation (CCD) (Luo et al., 2021) on scenario 1 task S4. CCD represents the distance between the distribution of feature embeddings in the source and the target domains. The results demonstrate that our proposed MSCDA achieves better alignment of the features compared to the other three methods, thereby yielding better discriminative power in target domain.

follow the normalization in Luo et al. (2021) so that the CCD of the baseline method Src-Only is always 1. A smaller CCD value indicates better feature alignment, while a larger CCD value indicates poorer alignment. The results depicted in Fig. 8 show that MSCDA achieves better feature alignment in both categories compared to other methods. In the 'NonBreast' category, our proposed MSCDA method exhibits a CCD of 0.599, indicating a slight improvement over the P2P contrast method (0.601). By contrast, in the foreground 'Breast' category, the CCD of MSCDA (0.297) is significantly lower than the other methods (Src-Only = 1, self-training = 0.728, P2P = 0.597), demonstrating a significant enhancement in feature alignment. These results support the hypothesis that multi-scale contrastive learning can better exploit deeper semantic information in UDA, leading to higher discrimination of the model towards target images.

6. Conclusion

In this paper, a novel multi-level semantic-guided contrastive UDA framework for breast MRI segmentation, named MSCDA, is introduced. We found that by combining self-training with multi-level contrastive loss, the semantic information can be further exploited to improve segmentation performance on the unlabeled target domain. Furthermore, we built a hybrid memory bank for sample storage and proposed a category-wise cross-domain sampling strategy to balance the contrastive pairs. The proposed model shows a robust and clinically relevant performance in a cross-sequence label-sparse scenario of breast MRI segmentation. The code of our MSCDA model is available at https://github.com/ShengKuangCN/MSCDA.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Philippe Lambin reports financial support was provided by European Research Council. Henry C. Woodruff reports financial support was provided by Dutch Cancer Society. Philippe Lambin reports a relationship with Radiomics SA that includes: consulting or advisory, speaking and lecture fees, and travel reimbursement. Philippe Lambin reports a relationship with PTTheragnostic that includes: funding grants. Philippe Lambin reports a relationship with Health Innovation Ventures that includes: funding grants. Philippe Lambin reports a relationship with BHV that includes: consulting or advisory, speaking and lecture fees, and travel reimbursement. Philippe Lambin reports a relationship with Merck that includes: consulting or advisory, speaking and lecture fees, and travel reimbursement. Philippe Lambin reports a relationship with Varian that includes: consulting or advisory, speaking and lecture fees, and travel reimbursement. Philippe Lambin reports a relationship with Elekta, ptTheragnostc, BMS, and Convert pharmaceutcals that includes: consulting or advisory, speaking and lecture fees, and travel reimbursement. Henry C. Woodruff reports a relationship with Radiomics SA that includes: equity or stocks.

Data availability

The authors do not have permission to share data.

Acknowledgments

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015 n° 694812 - Hypoximmuno), ERC-2020-PoC: 957565-AUTO.DISTINCT. Authors also acknowledge financial support from the European Union's Horizon 2020 research and innovation programme under grant agreement: ImmunoSABR n° 733008, CHAIMELEON n° 952172, EuCanImage n° 952103. This work was also partially supported by the Dutch Cancer Society (KWF Kankerbestrijding), The Netherlands, project number 14449.

References

Alonso, I., Sabater, A., Ferstl, D., Montesano, L., & Murillo, A. C. (2021). Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8219–8228).

Ash, J. T., Goel, S., Krishnamurthy, A., & Misra, D. (2021). Investigating the role of negatives in contrastive representation learning. arXiv preprint arXiv:2106.09943.

Awasthi, P., Dikkala, N., & Kamath, P. (2022). Do more negative samples necessarily hurt in contrastive learning? In *International conference on machine learning* (pp. 1101–1116). PMLR.

Azamjah, N., Soltan-Zadeh, Y., & Zayeri, F. (2019). Global trend of breast cancer mortality rate: a 25-year study. *Asian Pacific Journal of Cancer Prevention: APJCP*, *20*(7), 2015.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning, vol. 4, no. 4*. Springer.

Bleker, J., Kwee, T. C., Rouw, D., Roest, C., Borstlap, J., de Jong, I. J., et al. (2022). A deep learning masked segmentation alternative to manual segmentation in biparametric MRI prostate cancer radiomics. *European Radiology*, *32*(9), 6526–6535.

Chaitanya, K., Erdil, E., Karani, N., & Konukoglu, E. (2020). Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, *33*, 12546–12558.

Chen, C., Dou, Q., Chen, H., Qin, J., & 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, vol. 33, no. 04* (pp. 865–872).

Chen, X., Fan, H., Girshick, R., & He, K. (2020). Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607). PMLR.

Chen, C., Xie, W., Huang, W., Rong, Y., Ding, X., Huang, Y., et al. (2019). Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 627–636).

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision* (pp. 801–818).

Dalmış, M. U., Litjens, G., Holland, K., Setio, A., Mann, R., Karssemeijer, N., et al. (2017). Using deep learning to segment breast and fibroglandular tissue in MRI volumes. *Medical Physics*, *44*(2), 533–546.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). Ieee.

Despotović, I., Goossens, B., & Philips, W. (2015). MRI segmentation of the human brain: challenges, methods, and applications. *Computational and Mathematical Methods in Medicine*, *2015*.

Dou, Q., Ouyang, C., Chen, C., Chen, H., Glocker, B., Zhuang, X., et al. (2018). Pnp-adanet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation. arXiv preprint arXiv:1812.07907.

El Adoui, M., Mahmoudi, S. A., Larhmam, M. A., & Benjelloun, M. (2019). MRI breast tumor segmentation using different encoder and decoder CNN architectures. *Computers*, *8*(3), 52.

Francies, F. Z., Hull, R., Khanyile, R., & Dlamini, Z. (2020). Breast cancer in low-middle income countries: Abnormality in splicing and lack of targeted treatment options. *American Journal of Cancer Research*, *10*(5), 1568.

Gallego-Ortiz, C., & Martel, A. L. (2017). Using quantitative features extracted from T2-weighted MRI to improve breast MRI computer-aided diagnosis (CAD). *PLoS One*, *12*(11), Article e0187501.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*.

Granzier, R., Ibrahim, A., Primakov, S., Keek, S., Halilaj, I., Zwanenburg, A., et al. (2022). Test–retest data for the assessment of breast MRI radiomic feature repeatability. *Journal of Magnetic Resonance Imaging*, *56*(2), 592–604.

Granzier, R. W., Ibrahim, A., Primakov, S. P., Samiei, S., van Nijnatten, T. J., de Boer, M., et al. (2021). MRI-based radiomics analysis for the pretreatment prediction of pathologic complete tumor response to neoadjuvant systemic therapy in breast cancer patients: a multicenter study. *Cancers*, *13*(10), 2447.

Granzier, R., Verbakel, N., Ibrahim, A., Van Timmeren, J., Van Nijnatten, T., Leijenaar, R., et al. (2020). MRI-based radiomics in breast cancer: Feature robustness with respect to inter-observer segmentation variability. *Scientific Reports*, *10*(1), 1–11.

Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, *33*, 21271–21284.

Guan, H., & Liu, M. (2021). Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, *69*(3), 1173–1185.

He, J., Deng, Z., & Qiao, Y. (2019). Dynamic multi-scale filters for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3562–3572).

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9729–9738).

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026–1034).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hoffman, J., Tzeng, E., Park, T., Zhu, J. Y., Isola, P., Saenko, K., et al. (2018). Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning* (pp. 1989–1998). Pmlr.

Hoffman, J., Wang, D., Yu, F., & Darrell, T. (2016). Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649.

Hoyer, L., Dai, D., & Van Gool, L. (2022). Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9924–9935).

Hu, B., Xu, K., Zhang, Z., Chai, R., Li, S., & Zhang, L. (2018). A radiomic nomogram based on an apparent diffusion coefficient map for differential diagnosis of suspicious breast findings. *Chinese Journal of Cancer Research*, *30*(4), 432.

Hu, X., Zeng, D., Xu, X., & Shi, Y. (2021). Semi-supervised contrastive learning for label-efficient medical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 481–490). Springer.

Hung, A. L. Y., Zheng, H., Miao, Q., Raman, S. S., Terzopoulos, D., & Sung, K. (2022). CAT-Net: A cross-slice attention transformer model for prostate zonal segmentation in MRI. *IEEE Transactions on Medical Imaging*, *42*(1), 291–303.

Ibtehaz, N., & Rahman, M. S. (2020). MultiResUNet: Rethinking the U-net architecture for multimodal biomedical image segmentation. *Neural Networks*, *121*, 74–87.

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., et al. (2018). Nnu-net: Self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486.

Ito, R., Nakae, K., Hata, J., Okano, H., & Ishii, S. (2019). Semi-supervised deep learning of brain tissue segmentation. *Neural Networks*, *116*, 25–34.

Ivanovska, T., Jentschke, T. G., Daboul, A., Hegenscheid, K., Völzke, H., & Wörgötter, F. (2019). A deep learning framework for efficient analysis of breast volume and fibroglandular tissue using MR data with strong artifacts. *International Journal of Computer Assisted Radiology and Surgery*, *14*(10), 1627–1633.

Jiang, J., Hu, Y. C., Tyagi, N., Zhang, P., Rimner, A., Mageras, G. S., et al. (2018). Tumor-aware, adversarial domain adaptation from CT to MRI for lung cancer segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 777–785). Springer.

Kim, T., Cha, M., Kim, H., Lee, J. K., & Kim, J. (2017). Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning* (pp. 1857–1865). PMLR.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kleppe, A., Skrede, O. J., De Raedt, S., Liestøl, K., Kerr, D. J., & Danielsen, H. E. (2021). Designing deep learning studies in cancer diagnostics. *Nature Reviews Cancer*, *21*(3), 199–211.

Kouw, W. M., & Loog, M. (2018). An introduction to domain adaptation and transfer learning. arXiv preprint arXiv:1812.11806.

Liu, Z., Zhu, Z., Zheng, S., Liu, Y., Zhou, J., & Zhao, Y. (2022). Margin preserving self-paced contrastive learning towards domain adaptation for medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, *26*(2), 638–647.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).

Loquercio, A., Segu, M., & Scaramuzza, D. (2020). A general framework for uncertainty estimation in deep learning. *IEEE Robotics and Automation Letters*, *5*(2), 3153–3160.

Lowry, K. P., Geuzinge, H. A., Stout, N. K., Alagoz, O., Hampton, J., Kerlikowske, K., et al. (2022). Breast cancer screening strategies for women with ATM, CHEK2, and PALB2 pathogenic variants: a comparative modeling analysis. *JAMA Oncology*, *8*(4), 587–596.

Luo, Y., Liu, P., Zheng, L., Guan, T., Yu, J., & Yang, Y. (2021). Category-level adversarial adaptation for semantic segmentation using purified features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(8), 3940–3956.

Mann, R. M., Cho, N., & Moy, L. (2019). Breast MRI: state of the art. *Radiology*, *292*(3), 520–536.

Mehrkanoon, S. (2019). Cross-domain neural-kernel networks. *Pattern Recognition Letters*, *125*, 474–480.

Mehrkanoon, S., & Suykens, J. A. K. (2017). Regularized semipaired kernel CCA for domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(7), 3199–3213.

Negi, A., Raj, A. N. J., Nersisson, R., Zhuang, Z., & Murugappan, M. (2020). RDA-UNET-WGAN: an accurate breast ultrasound lesion segmentation using wasserstein generative adversarial networks. *Arabian Journal for Science and Engineering*, *45*(8), 6399–6410.

Oord, A. v. d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

Perone, C. S., Ballester, P., Barros, R. C., & Cohen-Adad, J. (2019). Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, *194*, 1–11.

Perone, C. S., & Cohen-Adad, J. (2018). Deep semi-supervised segmentation with weight-averaged consistency targets. In *Deep learning in medical image analysis and multimodal learning for clinical decision support* (pp. 12–19). Springer.

Piantadosi, G., Sansone, M., & Sansone, C. (2018). Breast segmentation in MRI via U-net deep convolutional neural networks. In *2018 24th international conference on pattern recognition* (pp. 3917–3922). IEEE.

Qiu, S., Miller, M. I., Joshi, P. S., Lee, J. C., Xue, C., Ni, Y., et al. (2022). Multimodal deep learning for Alzheimer's disease dementia assessment. *Nature Communications*, *13*(1), 3404.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Sardanelli, F., Aase, H. S., Álvarez, M., Azavedo, E., Baarslag, H. J., Balleyguier, C., et al. (2017). Position paper on screening for breast cancer by the European society of breast imaging (EUSOBI) and 30 national breast radiology bodies from Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, czech Republic, Denmark, estonia, Finland, France, Germany, Greece, Hungary, iceland, Ireland, Italy, Israel, Lithuania, Moldova, The Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Spain, Sweden, Switzerland and Turkey. *European Radiology*, *27*, 2737–2743.

Saslow, D., Boetes, C., Burke, W., Harms, S., Leach, M. O., Lehman, C. D., et al. (2007). American cancer society guidelines for breast screening with MRI as an adjunct to mammography. *CA: A Cancer Journal for Clinicians*, *57*(2), 75–89.

Shanis, Z., Gerber, S., Gao, M., & Enquobahrie, A. (2019). Intramodality domain adaptation using self ensembling and adversarial training. In *Domain adaptation and representation transfer and medical image learning with less labels and imperfect data* (pp. 28–36). Springer.

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Jorge Cardoso, M. (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support* (pp. 240–248). Springer.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209–249.

Surucu, M., Isler, Y., Perc, M., & Kara, R. (2021). Convolutional neural networks predict the onset of paroxysmal atrial fibrillation: Theory and applications. *Chaos. An Interdisciplinary Journal of Nonlinear Science*, *31*(11), Article 113119.

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, *30*.

Tsai, Y. H., Hung, W. C., Schulter, S., Sohn, K., Yang, M. H., & Chandraker, M. (2018). Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7472–7481).

Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7167–7176).

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11).

Vu, T. H., Jain, H., Bucher, M., Cord, M., & Pérez, P. (2019). Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2517–2526).

Wang, W., Zhou, T., Yu, F., Dai, J., Konukoglu, E., & Van Gool, L. (2021). Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7303–7313).

Wu, S., Weinstein, S. P., Conant, E. F., & Kontos, D. (2013). Automated fibroglandular tissue segmentation and volumetric density estimation in breast MRI using an atlas-aided fuzzy C-means method. *Medical Physics*, *40*(12), Article 122302.

Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3733–3742).

Wu, H., Zhang, J., Huang, K., Liang, K., & Yu, Y. (2019). Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. arXiv preprint arXiv:1903.11816.

Xie, B., Li, S., Li, M., Liu, C. H., Huang, G., & Wang, G. (2022). SePiCo: Semantic-guided pixel contrast for domain adaptive semantic segmentation. arXiv preprint arXiv:2204.08808.

Yesilkaya, B., Perc, M., & Isler, Y. (2022). Manifold learning methods for the diagnosis of ovarian cancer. *Journal of Computer Science*, *63*, Article 101775.

Yi, Z., Zhang, H., Tan, P., & Gong, M. (2017). Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision* (pp. 2849–2857).

Zhang, Y., Chen, J. H., Chang, K. T., Park, V. Y., Kim, M. J., Chan, S., et al. (2019). Automatic breast and fibroglandular tissue segmentation in breast MRI using deep learning by a fully-convolutional residual neural network U-net. *Academic Radiology*, *26*(11), 1526–1535.

Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., et al. (2018). Context encoding for semantic segmentation. In *The IEEE conference on computer vision and pattern recognition*.

Zhang, F., Koltun, V., Torr, P., Ranftl, R., & Richter, S. R. (2022). Unsupervised contrastive domain adaptation for semantic segmentation. arXiv preprint arXiv:2204.08399.

Zhang, Y., Miao, S., Mansi, T., & Liao, R. (2018). Task driven generative modeling for unsupervised domain adaptation: Application to x-ray image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 599–607). Springer.

Zhang, L., Mohamed, A. A., Chai, R., Guo, Y., Zheng, B., & Wu, S. (2020). Automated deep learning method for whole-breast segmentation in diffusion-weighted breast MRI. *Journal of Magnetic Resonance Imaging*, *51*(2), 635–643.

Zhang, J., Saha, A., Zhu, Z., & Mazurowski, M. A. (2018). Hierarchical convolutional neural networks for segmentation of breast tumors in MRI with application to radiogenomics. *IEEE Transactions on Medical Imaging*, *38*(2), 435–447.

Zhang, P., Zhang, B., Zhang, T., Chen, D., Wang, Y., & Wen, F. (2021). Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12414–12424).

Zhao, X., Vemulapalli, R., Mansfield, P. A., Gong, B., Green, B., Shapira, L., et al. (2021). Contrastive learning for label efficient semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10623–10633).

Zhao, X., Xie, P., Wang, M., Li, W., Pickhardt, P. J., Xia, W., et al. (2020). Deep learning–based fully automated detection and segmentation of lymph nodes on multiparametric-mri for rectal cancer: A multicentre study. *EBioMedicine*, *56*, Article 102780.

Zhao, Z., Xu, K., Li, S., Zeng, Z., & Guan, C. (2021). Mt-uda: Towards unsupervised cross-modality medical image segmentation with limited source labels. In *International conference on medical image computing and computer-assisted intervention* (pp. 293–303). Springer.

Zhong, Y., Yuan, B., Wu, H., Yuan, Z., Peng, J., & Wang, Y. X. (2021). Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7273–7282).

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