Self-Supervised Domain Adaptive Segmentation of Breast Cancer via Test-Time Fine-Tuning

Kyungsu Lee[1], Haeyun Lee[2], Georges El Fakhri[3], Jonghye Woo[3], and Jae Youn Hwang[1](✉)

[1] Department of Electrical Engineering and Computer Science, Daegu Gyeongbuk Institute of Science and Technology, Daegu 42988, South Korea
`{ks_lee,jyhwang}@dgist.ac.kr`

[2] Production Engineering Research Team, Samsung SDI, Yongin 17084, South Korea

[3] Gordon Center for Medical Imaging, Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA

Abstract. Unsupervised domain adaptation (UDA) has become increasingly popular in imaging-based diagnosis due to the challenge of labeling a large number of datasets in target domains. Without labeled data, well-trained deep learning models in a source domain may not perform well when applied to a target domain. UDA allows for the use of large-scale datasets from various domains for model deployment, but it can face difficulties in performing adaptive feature extraction when dealing with unlabeled data in an unseen target domain. To address this, we propose an advanced test-time fine-tuning UDA framework designed to better utilize the latent features of datasets in the unseen target domain by fine-tuning the model itself during diagnosis. Our proposed framework is based on an auto-encoder-based network architecture that fine-tunes the model itself. This allows our framework to learn knowledge specific to the unseen target domain during the fine-tuning phase. In order to further optimize our framework for the unseen target domain, we introduce a re-initialization module that injects randomness into network parameters. This helps the framework to converge to a local minimum that is better-suited for the target domain, allowing for improved performance in domain adaptation tasks. To evaluate our framework, we carried out experiments on UDA segmentation tasks using breast cancer datasets acquired from multiple domains. Our experimental results demonstrated that our framework achieved state-of-the-art performance, outperforming other competing UDA models, in segmenting breast cancer on ultrasound images from an unseen domain, which supports its clinical potential for improving breast cancer diagnosis.

Keywords: Unsupervised Domain Adaptation · Test-Time Tuning · Breast Cancer · Segmentation · Ultrasound Imaging

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43907-0_52.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14220, pp. 539–550, 2023.
https://doi.org/10.1007/978-3-031-43907-0_52

1 Introduction

In recent years, deep learning (DL) methods have demonstrated remarkable performance in detecting and localizing tumors on ultrasound images [2,27]. Compared with conventional image processing methods, DL methods provide an accurate feature extraction capability on ultrasound images, despite their low resolution and noise disturbance, leading to superior segmentation accuracy [2,5,14]. However, there are some limitations in developing a DL model in a source domain and deploying it in an unseen target domain. The primary limitation is that DL models require a large number of training samples to achieve accurate predictions [8,24]. Yet, acquiring large training datasets and their corresponding labels, especially from a cohort of patients, can be costly or even infeasible, which poses a significant challenge in developing a DL model with high performance [7]. Second, even when large-scale datasets are available through collaborative research from multiple sites, DL models trained on such datasets may yield sub-optimal solutions due to domain gaps caused by differences in images acquired from different sites [20]. Third, due to the small number of datasets from each domain, the images for each individual domain may not capture representative features, limiting the ability of DL models to generalize across domains [3].

Domain adaptation (DA) has been extensively studied to alleviate the aforementioned limitations, the goal of which is to reduce the domain gap caused by the diversity of datasets from different domains [12,20,26,29,33]. Example solutions include transfer learning- and style transfer-based methods. Nonetheless, unlike natural images, generating labels can be a challenging task, making it difficult to apply general DA methods; thus bridging domain gaps by DA methods remains limited [26,33]. This is due to sensitive privacy issues in patients' data, particularly in collaborative research, which restricts access to labels from different domains. As a result, conventional DA methods cannot be easily applied [10]. More recently, unsupervised domain adaptation (UDA) has been introduced to address this issue [16,33], aiming to generate semi-predictions (pseudo-labels) in target domains first, followed by producing accurate predictions using the pseudo-labels. One critical limitation of pseudo-label-based UDA is the possibility of error accumulation due to mispredicted pseudo-labels. This can lead to significant degradation of the performance of DL models, as errors can compound and become more pronounced over time [17,25].

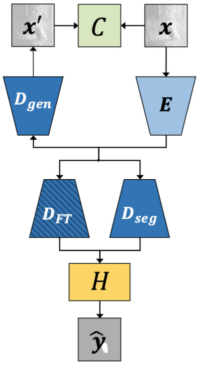
To alleviate the problem of pseudo-label-based UDA, in this work, we propose an advanced UDA framework based on self-supervised DA with a test-time fine-tuning network. Test-time adaptation methods have been developed [4,11,13,23] to improve the learning of knowledge in target domains. The distinctive feature of our test-time self-supervised DA is that it enables the DL network (i) to learn knowledge about the features of target domains by fine-tuning the network itself during the test-time phase, rather than generating pseudo-labels and then (ii) to provide precise predictions on images in target domains, by using the fine-tuned network. Specifically, we adopt self-supervised learning and verify the model via thorough mathematical analysis. Our framework was tested on the task of breast cancer segmentation in ultrasound images, but it could also be applied to other lesion segmentation tasks.

To summarize, our contributions are three-fold:

- We design a self-supervised DA framework that includes a parameter search method and provide a mathematical justification for it. With our framework, we are able to identify the best-performing parameters that result in improved performance in DA tasks.
- Our framework is effective at preserving privacy, since it carries out DA using only pre-trained network parameters, without transferring any patient data.
- We applied our framework to the task of segmenting breast cancer from ultrasound imaging data, demonstrating its superior performance over competing UDA methods.

Our results indicate that our framework is effective in improving the accuracy of breast cancer segmentation from ultrasound images, which could have potential implications for improving the diagnosis and treatment of breast cancer.

2 Methodology



```
Algorithm 1: Test-Time Fine-Tuning Scheme
Input:  $E$, $H$, $C$, and $D_{\text{gen}} = D_{\text{seg}}$
def Training_on_Source:
        Sample batches of $(s, \bar{s}) \sim \mathcal{S}$
        Update $E$ and $D_{\text{seg}}$ via $\mathcal{L}_{\text{BCE}}((H \circ D_{\text{seg}} \circ E)(s), \bar{s}))$
        Update $E$ and $D_{\text{gen}}$ via $\mathcal{L}_{\text{GAN}}((D_{\text{gen}} \circ E)(s), s)$
        return $E^{\mathcal{S}}$ and $D^{\mathcal{S}}_{\text{seg}} = D^{\mathcal{S}}_{\text{gen}}$
End
def Fine_Tuning_on_Target:
        Sample batches of $(t, ?) \sim \mathcal{T}$
        Update $D^{\mathcal{S}}_{\text{gen}}$ via $\mathcal{L}_{\text{GAN}}(D^{\mathcal{S}}_{\text{gen}}(E(t)), t)$, then $D^{\mathcal{S}\to\mathcal{T}}_{\text{gen}}$
        Share parameters from $D^{\mathcal{S}\to\mathcal{T}}_{\text{gen}}$ to $D_{\text{FT}}$
        return $D_{\text{FT}} = D^{\mathcal{S}\to\mathcal{T}}_{\text{seg}}$
End
def Prediction_on_Target:
        Sample batches of $(t, ?) \sim \mathcal{T}$
        $\hat{t} = \Big(H \circ (D^{\mathcal{S}}_{\text{seg}} \oplus D_{\text{FT}}) \circ E\Big)(t))$
        return $\hat{y}$
End
Output: Predictions $(\hat{y})$ on $\mathcal{T}$
```

Fig. 1. Architecture of our TTFT network (**Left**) and its pipeline (**Right**).

2.1 Test-Time Fine-Tuning (TTFT) Network and Its Pipeline

Network Architecture. Our proposed TTFT network is based on self-supervised DA [31], which is a part of UDA and can be seen as multi-task learning, involving both the main and pretext tasks, as shown in Fig. 1. In the main task, an encoder (E), a decoder for segmentation (D_{seg}), and a segmentation header (H) are included. The main task is the segmentation task, $(H \circ D_{\text{seg}} \circ E)(x)$. In predicting segmentation labels in the target domain ($\mathcal{T}$), D_{FT} is also involved in the main task, and the final prediction after the fine-tuning is provided by $\big(H \circ (D_{\text{seg}} \oplus D_{\text{FT}}) \circ E\big)(x)$, where $\oplus$ is the concatenation operation. In the pretext task, E, a decoder for a generator, D_{gen}, and a discriminator, C, are involved. The pretext task aims to generate synthetic images, $(D_{\text{gen}} \circ E)(t)$. Note that D_{gen} and D_{seg} share the same parameters to enable knowledge transfer. However, since the headers of image reconstruction and generating segmentation mask are different (different output), a new header incorporating D_{FT} and D_{seg} is devised and leverages the outputs of two decoders. Besides, $D_{gen} = D_{FT}$ is fine-tuned during the fine-tuning step, and the D_{FT} learns the knowledge of the input domain via image reconstruction. Two distinct knowledge (information) from D_{seg} and D_{FT} enable the network to utilize target domain knowledge and predict precise predictions.

Pre-training in Source Domain. The model M is first trained in $\mathcal{S}$ in a supervised manner with $(s, \bar{s}) \sim \mathcal{S}$ in both main and pretext tasks as below:

$$\Theta_{\mathcal{S}}^{m}, \Theta_{\mathcal{S}}^{p} = \underset{\theta_{\mathcal{S}}^{m}, \theta_{\mathcal{S}}^{p}}{\operatorname{argmin}} \sum_{s} \Big\{ \mathcal{L}_{\text{BCE}}\big((H \circ D_{\text{seg}} \circ E)(s), \bar{s}\big) + \mathcal{L}_{\text{GAN}}\big((D_{\text{gen}} \circ E)(s), s\big) \Big\}, \tag{1}$$

where $\mathcal{L}_{\text{BCE}}$ and $\mathcal{L}_{\text{GAN}}$ represent the loss functions for binary cross-entropy and generative adversarial network [6], respectively. $\Theta_{\mathcal{S}}^{m}$ includes $E^{\mathcal{S}}$, $D_{seg}^{\mathcal{S}}$, and $H^{\mathcal{S}}$, while $\Theta_{\mathcal{S}}^{p}$ includes $E^{\mathcal{S}}$, $D_{gen}^{\mathcal{S}}$, and $C^{\mathcal{S}}$. Additionally, $D_{\text{seg}}^{\mathcal{S}} = D_{\text{gen}}^{\mathcal{S}}$.

Fine-Tuning in Target Domain. Since the pre-trained model is likely to produce imprecise predictions in $\mathcal{T}$, the model should learn domain knowledge about $\mathcal{T}$. To this end, in the pretext task, for self-supervised learning, the model is fine-tuned in $\mathcal{T}$ to generate synthetic images identical to the input images as below:

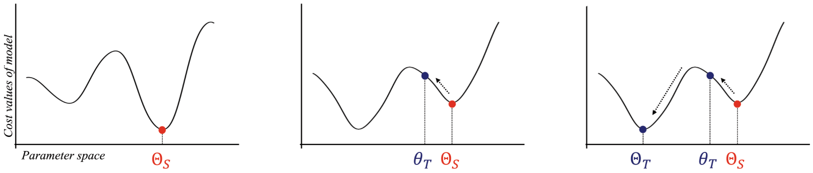
$$\Theta_{\mathcal{T}}^{p} = \underset{\theta_{\mathcal{T}}^{p}}{\operatorname{argmin}} \sum_{t} \mathcal{L}_{\text{GAN}}\big((D_{\text{gen}}^{\mathcal{S}} \circ E^{\mathcal{S}})(t), t\big) \;\; \Rightarrow \;\; \Theta_{\mathcal{T}}^{p} \supseteq E^{\mathcal{S}} \cup D_{\text{gen}}^{\mathcal{S} \to \mathcal{T}}, \tag{2}$$

where only D_{gen} is fine-tuned to achieve memory efficiency and to decrease the fine-tuning time, and $D_{\text{gen}}^{\mathcal{S}}$ is fine-tuned as $D_{\text{gen}}^{\mathcal{S} \to \mathcal{T}}$. Then, $D_{\text{gen}}^{\mathcal{S} \to \mathcal{T}}$ is transferred to D_{FT}, and knowledge distillation via self-supervised learning is realized. Hence, the precise predictions in $\mathcal{T}$ could be provided by $\big(H \circ (D_{\text{seg}}^{\mathcal{S}} \oplus D_{\text{FT}}^{\mathcal{T}}) \circ E\big)(x)$.

Benefits of Our Dual-Pipeline. Due to the symmetric property of mutual information in information entropy ($\mathbb{H}$), we have $I(X;Y) = H(X) + H(Y) - H(X,Y)$. As a result, the predictions made by the fine-tuned network in the target domain ($\mathcal{T}$) lead to reduced entropy, as shown below:

$$\mathbb{H}\big((H \circ (D_{\text{seg}}^{\mathcal{S}} \oplus D_{\text{FT}}^{\mathcal{T}}) \circ E)(t), \bar{t}\big) \leq \; \mathbb{H}\big((H \circ D_{\text{seg}}^{\mathcal{S}} \circ E)(t), \bar{t}\big) + \mathbb{H}\big((H \circ D_{\text{FT}}^{\mathcal{T}} \circ E)(t), \bar{t}\big). \tag{3}$$

Since $D_{\text{seg}}^{\mathcal{S}}$ is fully optimized for $\mathcal{S}$ in a supervised manner, it guarantees a baseline segmentation performance. Furthermore, since $D_{\text{FT}}^{\mathcal{T}}$ is fine-tuned in $\mathcal{T}$



(a) Source domain (b) Target domain (c) Target domain

Fig. 2. Illustration of the local minimum of the source (a) and target (b) domains and parameter fluctuation (c)

using knowledge distillation, it can provide domain-specific information for $\mathcal{T}$. As a result, the predictions made by the fine-tuned model in $\mathcal{T}$ are jointly constrained by the expectations of $D^{\mathcal{S}}_{\text{seg}}$ and $D^{\mathcal{T}}_{\text{FT}}$. This enables the final model to provide precise predictions in $\mathcal{T}$ by taking into account both the source domain and target domain information.

2.2 Parameter Fluctuation: Parameter Randomization Method

Since the loss function and its values can vary based on the distribution of inputs, and different domains can have different distributions, the local minimum identified in the source domain ($\mathcal{S}$) cannot be considered as the same local minimum in $\mathcal{T}$, as illustrated in Fig. 2. The y-axis of Fig. 2 indicates $\frac{1}{|\mathcal{X}|}\sum_x \mathcal{L}(M(x;\theta),\bar{x})$, and the local minimum is different in $\mathcal{S}$ and $\mathcal{T}$ as $\Theta_{\mathcal{S}}$ in Fig. 2a and $\Theta_{\mathcal{T}}$ in Fig. 2c, respectively. A longer fine-tuning time is required to re-position $\Theta_{\mathcal{S}}$ to $\Theta_{\mathcal{T}}$ as in Fig. 2c than to re-position $\theta_{\mathcal{T}}$ to $\Theta_{\mathcal{T}}$. Therefore, efficient fine-tuning is necessary to re-position the local minimum in Fig. 2b and this process is known as parameter fluctuation. Note that the parameter fluctuation is followed by the fine-tuning step.

Suppose C_i be the i^{th} convolution operator in D_{seg} with weight w_i, then $C_i(x) = w_i \cdot x$. Since $D^{\mathcal{S}}_{\text{seg}}$ provides the baseline segmentation performance, $D^{\mathcal{T}}_{FT}$ should provide similar feature maps to achieve the baseline performance. To this end, the mid-feature maps generated should be similar, i.e., $\forall_i C_i(F_i) \approx C'_i(F'_i)$, where C'_i represents the convolution in $D^{\mathcal{T}}_{FT}$, F_i represents i^{th} feature map, and $F_0 = E(x)$. Suppose $\forall_i |C_i(F_i) - C'_i(F'_i)| < \epsilon_i \ll 1$, such that $\forall_i F_i \approx F'_i$ by mathematical induction. Therefore, the sum of errors $(\sum |C_i(F_i) - C'_i(F'_i)|)$ is approximated by $\sum |w_i F_0 - w'_i F_0|$ *iff* $\forall_i F_i \approx F'_i$, which can be expressed as:

$$\sum |w_i F_0 - w'_i F_0| < \epsilon \ll 1 \Leftarrow \sum |w_i F_0 - w'_i F_0| \approx 0 \Leftrightarrow \sum |w_i - w'_i| = 0. \tag{4}$$

Here, we denote $w_i - w'_i = f_i$ as the fluctuation vector in the vector space, and the condition $\sum f_i = 0$ indicates that the sum of the fluctuation vectors should be zero under the condition of $|f_i| < r \ll 1$. Hence, we achieve the condition for the parameter fluctuation that the centers of parameters of $\Theta_{\mathcal{S}}$ and $\theta_{\mathcal{T}}$ should be the same in the vector space, and the length of the fluctuation vector should be less than a certain small threshold ($0 < r \ll 1$). Therefore, the parameter fluctuation aims to add random vectors of which length is less than $0 < r \ll 1$ on the parameters of $\Theta_{\mathcal{S}}$, and the sum of vectors should be zero. To summarize, the parameter fluctuation aims to add randomness on $\Theta_{\mathcal{S}}$ as follows:

$$\theta_{\mathcal{T}} = \{w_i + f_i | \ w_i \in \Theta_{\mathcal{S}}, \ \sum f_i = 0, \ \ 0 < |f_i| < r \ll 1\}. \tag{5}$$

3 Experiments

3.1 Experimental Set-Ups

To evaluate the segmentation performance of our TTFT framework, we used three different ultrasound databases: BUS [32], BUSI [1], and BUV [18], which are considered to be different domains. All three databases contain ultrasound imaging data and segmentation masks for breast cancer, with the masks labeled as 0 (background) and 1 (lesion) using a one-hot encoding. The BUS database consists of 163 images along with corresponding labels. The BUSI database contains 780 images, with 133 images belonging to the NORMAL class and having labels containing only 0 values. The BUV database originally consists of ultrasound videos, providing a total of 21,702 frames. While the database also provides labels for the detection task, we processed these labels as segmentation masks using a region growing method [15].

We employed different deep-learning models for evaluation. Specifically, U-Net [22] and FusionNet [21] were employed as our baseline models, since U-Net is a widely used basic model for segmentation, and FusionNet contains advanced residual modules, compared with U-Net. *Ours I* and *Ours II* were based on U-Net and FusionNet as the baseline network, respectively. Additionally, MIB-Net [28], which is a state-of-the-art model for breast cancer segmentation using ultrasound images, was employed for comparison. Furthermore, CBST [33] and CT-Net [16] were employed as the comparison models for UDA methods. As the evaluation metrics, dice coefficient (D. Coef), PRAUC, which is an area under a precision-recall curve, and cohen kappa (κ) were employed [30]. Our experimental set-ups included: (i) individual databases were used to assess the baseline segmentation performance (Appendix); (ii) the domain adaptive segmentation performance was assessed using the three databases, where two databases were regarded as the source domain, and the remaining database was regarded as the target domain; and (iii) the ablation study was carried out to evaluate the proposed network architecture along with the randomized re-initialization method.

3.2 Comparison Analysis

Since all compared DL models show similar D. Coef, only UDA performance is comparable as a control in our experiments. In this experiment, two databases were used for training, and the remaining database was used for testing. For instance, *BUS* in Fig. 3 illustrates the *BUS* database was used for testing, and

Dataset	Model	κ	PRAUC	D. Ceof (95% CI)
BUS	U-Net	0.462	0.588	0.638 (0.628-0.648)
	FusionNet	0.504	0.620	0.664 (0.652-0.675)
	MIB-Net	0.554	0.653	0.695 (0.686-0.703)
	CBST	0.724	0.817	0.805 (0.791-0.820)
	CT-Net	0.734	0.819	0.812 (0.802-0.823)
	Ours I	0.790	0.864	0.850 (0.836-0.864)
	Ours II	0.813	0.876	0.866 (0.854-0.878)
BUSI	U-Net	0.250	0.190	0.338 (0.327-0.350)
	FusionNet	0.289	0.213	0.370 (0.356-0.385)
	MIB-Net	0.329	0.237	0.403 (0.392-0.415)
	CBST	0.565	0.424	0.606 (0.573-0.640)
	CT-Net	0.597	0.456	0.634 (0.608-0.661)
	Ours I	0.711	0.591	0.735 (0.692-0.778)
	Ours II	0.747	0.638	0.768 (0.728-0.808)
BUV	U-Net	0.225	0.189	0.329 (0.316-0.341)
	FusionNet	0.260	0.207	0.356 (0.341-0.372)
	MIB-Net	0.302	0.231	0.391 (0.379-0.402)
	CBST	0.536	0.411	0.585 (0.550-0.620)
	CT-Net	0.551	0.424	0.598 (0.576-0.620)
	Ours I	0.671	0.566	0.701 (0.655-0.748)
	Ours II	0.712	0.619	0.737 (0.699-0.776)

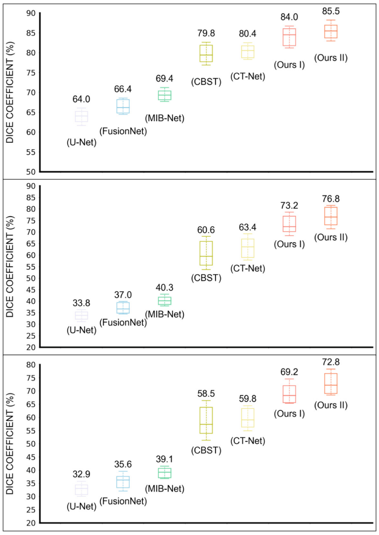


Fig. 3. Comparison analysis of our framework and comparison models: performance comparison table (Left) and Box-and-Whisker plot (Right).

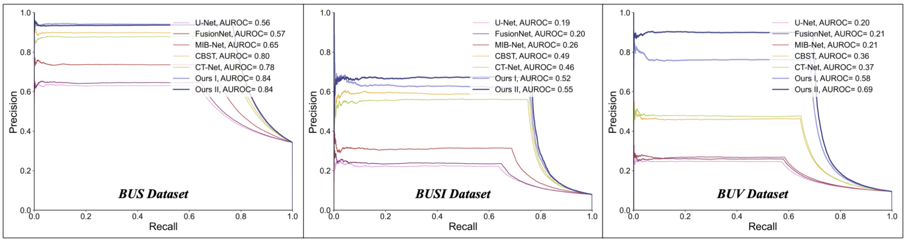


Fig. 4. Precision-Recall curves by ours and comparison models on each database. Area under the precision-recall curve (PR-AUC) values were reported.

the other two databases of *BUSI* and *BUV* were used for training. Figs. 3 and 4 show quantitative results, and Fig. 5 shows the sample segmentation results. Unlike the experiment using the individual database, U-Net, FusionNet, and MIB-Net showed significantly inferior scores due to domain gaps. In contrast, UDA methods of CBST and CT-Net showed superior scores, compared with others, and the scores were not strongly reduced, compared with the experiment with the single database. Note that, our TTFT framework achieved the best performance compared with other DL models. Additionally, *Ours II*, based on FusionNet, showed the best scores, potentially due to the advanced residual connection module. Furthermore, as illustrated in Fig 4, our framework provides superior precision scores in a long range of (0, 0.7), indicating that our frameworks estimated unnecessary mispredictions but precise predictions on cancer.

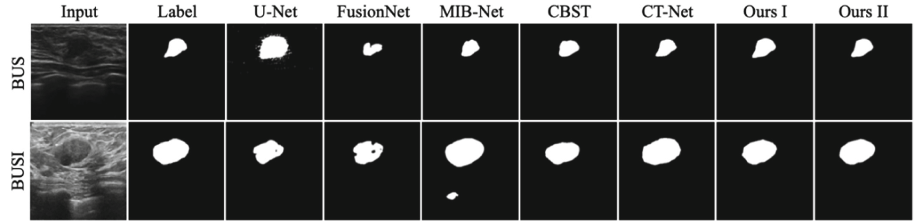


Fig. 5. Segmentation results by ours and comparison models on each database.

$E(t)$	$D^{\mathcal{S}}\text{seg}$	$D^{\text{fl}}_{\text{seg}}$	$D^{\mathcal{S}\to\mathcal{T}}_{\text{seg}}$
$D^{\mathcal{S}}_{\text{seg}}$	-	9.96 (3.26)	10.72 (3.02)
$D^{\mathcal{T}}_{\text{seg}}$	10.50 (3.87)	8.73 (2.95)	5.12 (0.87)

* Values are style loss of $D^{row}_{\text{seg}}(E(t))$ and $D^{col}_{\text{seg}}(E(t))$.

** The lower style loss, the more similar.

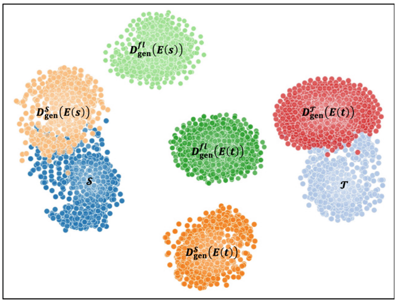


Fig. 6. Illustration of feature maps: style loss comparison (Left) and a T-SNE plot of generated images by different decoders (Right)

3.3 Ablation Study

In order to assess the effectiveness of each of the proposed modules, including the parameter fluctuation and fine-tuning methods, the ablation study was carried out. Since our framework contains three types of decoders, including $D^{\mathcal{S}}_{\text{seg}}$, D^{fl}_{seg}, and $D^{\mathcal{S}\to\mathcal{T}}_{\text{seg}}$ for the fine-tuning, we mainly targeted those decoders in our ablation study. Table 1 illustrates the quantitative results by different types of decoders. The higher D. coef value (+3.4%) of Pre-train + PF than that of Pre-train + Random Init and Pre-train + Offset confirms the effectiveness of the parameter fluctuation in the UDA performance. Additionally, the higher score (+11%) of Fine-tuning than Pre-train shows an outstanding UDA performance of the fine-tuning pipeline. Furthermore, the simultaneous utilization of the dual pipeline with $D^{\mathcal{S}}_{\text{seg}}$ and $D^{\mathcal{S}\to\mathcal{T}}_{\text{seg}}$ is justified by the scores of Pre-train + Fine-tuning. Using dual-pipeline and parameter fluctuation yielded the best performance. However, the utilization of ensemble pipelines of multiple fine-tuning modules was inefficient, since negligible performance improvements (+0.002) were observed, despite the heavy memory utilization.

Furthermore, Fig. 6 shows the effectiveness of the parameter fluctuation and fine-tuning methods. We first compared the similarity of feature-maps by decoders, including $D^{\mathcal{S}}_{\text{seg}}$, D^{fl}_{seg}, and $D^{\mathcal{S}\to\mathcal{T}}_{\text{seg}}$, with $D^{\mathcal{S}}_{\text{seg}}$ and $D^{\mathcal{T}}_{\text{seg}}$, which was fully optimized decoder in $\mathcal{T}$. Here, a style loss [9] was employed to measure the similarity of feature maps. Our framework was fine-tuned as $D^{\mathcal{S}}_{\text{seg}} \to D^{fl}_{\text{seg}} \to D^{\mathcal{S}\to\mathcal{T}}_{\text{seg}}$

Table 1. Dice coefficients by different versions of our TTFT framework. Random Init is D_{FT} is randomly initialized, and Offset indicates D_{FT} is initialized with the value of D_{seg} added by the offset value.

D. Coef (95% CI)	BUS	BUSI	BUV
Pre-train	0.664 (0.653–0.675)	0.664 (0.653–0.675)	0.664 (0.653–0.675)
Fine-tuning	0.774 (0.763–0.785)	0.774 (0.763–0.785)	0.774 (0.763–0.785)
Pre-train + Random Init	0.663 (0.653–0.673)	0.663 (0.653–0.673)	0.663 (0.653–0.673)
Pre-train + Offset	0.676 (0.668–0.684)	0.676 (0.668–0.684)	0.676 (0.668–0.684)
Pre-train + PF	0.697 (0.686–0.707)	0.697 (0.686–0.707)	0.697 (0.686–0.707)
Pre-train + Fine-tuning	0.799 (0.789–0.809)	0.799 (0.789–0.809)	0.799 (0.789–0.809)
Pre-train + PF + Fine-tuning	0.855 (0.844–0.866)	0.855 (0.844–0.866)	0.855 (0.844–0.866)
Pre-train + PF + N Fine-tuning	0.857 (0.842–0.872)	0.857 (0.842–0.872)	0.857 (0.842–0.872)

along which the similarity with $D^{\mathcal{T}}_{\text{seg}}$ of those decoders were increasing, and the feature-maps by $D^{\mathcal{S}\rightarrow\mathcal{T}}_{\text{seg}}$ were similar to those of $D^{\mathcal{T}}_{\text{seg}}$, compared with $D^{\mathcal{S}}_{\text{seg}}$, indicating UDA was successfully performed. Additionally, the generated images by decoders, including $D^{\mathcal{S}}_{\text{seg}}$, $D^{\text{fl}}_{\text{seg}}$, and $D^{\mathcal{S}\rightarrow\mathcal{T}}_{\text{seg}}$ in $\mathcal{S}$ and $\mathcal{T}$ are plotted with T-SNE, where the short distance represents the similar features [19]. The generated images became similar to $\mathcal{T}$ in order of $D^{\mathcal{S}}_{\text{seg}}$, $D^{\text{fl}}_{\text{seg}}$, and $D^{\mathcal{S}\rightarrow\mathcal{T}}_{\text{seg}}$, which confirmed the effectiveness of the fine-tuning method in terms of knowledge distillation. Additionally, the parameters were successfully re-positioned from the local minimum in $\mathcal{S}$ by parameter fluctuation, which was confirmed by the distances from $\mathcal{S}$ to $D^{\mathcal{S}}_{\text{gen}}$ and D^{fl}_{gen}.

4 Discussion and Conclusion

In this work, we proposed a DL-based segmentation framework for multi-domain breast cancer segmentation on ultrasound images. Due to the low resolution of ultrasound images, manual segmentation of breast cancer is challenging even for expert clinicians, resulting in a sparse number of labeled data. To address this issue, we introduced a novel self-supervised DA network for breast cancer segmentation in ultrasound images. In particular, we proposed a test-time fine-tuning network to learn domain-specific knowledge via knowledge distillation by self-supervised learning. Since UDA is susceptible to error accumulation due to imprecise pseudo-labels, which can lead to degraded performance, we employed a self-supervised learning-based pretext task. Specifically, we utilized an auto-encoder-based network architecture to generate synthetic images that matched the input images. Moreover, we introduced a randomized re-initialization module that injects randomness into network parameters to reposition the network from the local minimum in the source domain to a local minimum that is better suited for the target domain. This approach enabled our framework to efficiently fine-tune the network in the target domain and achieve better segmentation performance. Experimental results, carried out with three ultrasound databases from different domains, demonstrated the superior segmentation performance of our framework over other competing methods. Additionally, our framework is well-suited to a scenario in which access to source domain data is limited, due to data privacy protocols. It is worth noting that we used vanilla U-Net [22] and FusionNet [21] as baseline models to evaluate the basic performance of our TTFT framework. However, the use of more advanced baseline models could lead to even better segmentation performance, which is a subject for our future work. Moreover, our proposed framework is not limited to breast cancer segmentation on ultrasound images acquired from different domains. It can also be applied to other disease groups or imaging modalities such as MRI or CT.

Acknowledgements. This work was partially supported by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: 1711174564, RS-2022-00141185). Also, this work was partially supported by the Technology Innovation Program(20014214) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea).

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data Brief **28**, 104863 (2020)
2. Badawy, S.M., Mohamed, A.E.N.A., Hefnawy, A.A., Zidan, H.E., GadAllah, M.T., El-Banby, G.M.: Automatic semantic segmentation of breast tumors in ultrasound images based on combining fuzzy logic and deep learning-a feasibility study. PLoS ONE **16**(5), e0251899 (2021)
3. Barbato, F., Toldo, M., Michieli, U., Zanuttigh, P.: Latent space regularization for unsupervised domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2835–2845 (2021)
4. Bateson, M., Kervadec, H., Dolz, J., Lombaert, H., Ayed, I.B.: Source-free domain adaptation for image segmentation. Med. Image Anal. **82**, 102617 (2022)
5. van Beers, F., Lindström, A., Okafor, E., Wiering, M.A.: Deep neural networks with intersection over union loss for binary image segmentation. In: ICPRAM, pp. 438–445 (2019)
6. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
7. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Trans. Biomed. Eng. **69**(3), 1173–1185 (2021)
8. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
9. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
10. Kaissis, G.A., Makowski, M.R., Rückert, D., Braren, R.F.: Secure, privacy-preserving and federated machine learning in medical imaging. Nat. Mach. Intell. **2**(6), 305–311 (2020)
11. Karani, N., Erdil, E., Chaitanya, K., Konukoglu, E.: Test-time adaptable neural networks for robust medical image segmentation. Med. Image Anal. **68**, 101907 (2021)
12. Kouw, W.M., Loog, M.: An introduction to domain adaptation and transfer learning. arXiv preprint arXiv:1812.11806 (2018)

13. Kundu, J.N., Kulkarni, A., Singh, A., Jampani, V., Babu, R.V.: Generalize then adapt: source-free domain adaptive semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7046–7056 (2021)
14. Lee, H., Park, J., Hwang, J.Y.: Channel attention module with multi-scale grid average pooling for breast cancer segmentation in an ultrasound image. Ferroelectrics, and Frequency Control, IEEE Transactions on Ultrasonics (2020)
15. Lee, M.H., Kim, J.Y., Lee, K., Choi, C.H., Hwang, J.Y.: Wide-field 3D ultrasound imaging platform with a semi-automatic 3D segmentation algorithm for quantitative analysis of rotator cuff tears. IEEE Access **8**, 65472–65487 (2020)
16. Lee, S., Hyun, J., Seong, H., Kim, E.: Unsupervised domain adaptation for semantic segmentation by content transfer. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 8306–8315 (2021)
17. Liang, J., He, R., Sun, Z., Tan, T.: Exploring uncertainty in pseudo-label guided unsupervised domain adaptation. Pattern Recogn. **96**, 106996 (2019)
18. Lin, Z., Lin, J., Zhu, L., Fu, H., Qin, J., Wang, L.: A new dataset and a baseline model for breast lesion detection in ultrasound videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 614–623. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_59
19. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11), 2579–2605 (2008)
20. Nam, H., Lee, H., Park, J., Yoon, W., Yoo, D.: Reducing domain gap by reducing style bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8690–8699 (2021)
21. Quan, T.M., Hildebrand, D.G., Jeong, W.K.: FusionNet: a deep fully residual convolutional neural network for image segmentation in connectomics. arXiv preprint arXiv:1612.05360 (2016)
22. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
23. Roy, S., Trapp, M., Pilzer, A., Kannala, J., Sebe, N., Ricci, E., Solin, A.: Uncertainty-guided source-free domain adaptation. In: Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV. pp. 537–555. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19806-9_31
24. Ruder, S.: An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747 (2016)
25. Sun, Y., Tzeng, E., Darrell, T., Efros, A.A.: Unsupervised domain adaptation through self-supervision. arXiv preprint arXiv:1909.11825 (2019)
26. Toldo, M., Maracani, A., Michieli, U., Zanuttigh, P.: Unsupervised domain adaptation in semantic segmentation: a review. Technologies **8**(2), 35 (2020)
27. Vakanski, A., Xian, M., Freer, P.E.: Attention-enriched deep learning model for breast tumor segmentation in ultrasound images. Ultrasound Med. Biol. **46**(10), 2819–2833 (2020)
28. Wang, J., et al.: Information bottleneck-based interpretable multitask network for breast cancer classification and segmentation. Med. Image Anal., 102687 (2022)
29. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7201–7211 (2022)

30. Wang, Y., Yao, Y.: Breast lesion detection using an anchor-free network from ultrasound images with segmentation-based enhancement. Sci. Rep. **12**(1), 1–12 (2022)
31. Xu, J., Xiao, L., López, A.M.: Self-supervised domain adaptation for computer vision tasks. IEEE Access **7**, 156694–156706 (2019)
32. Yap, M.H., et al.: Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J. Biomed. Health Inform. **22**(4), 1218–1226 (2017)
33. Zou, Y., Yu, Z., Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 289–305 (2018)