Machine Vision and Applications (2025) 36:27
https://doi.org/10.1007/s00138-024-01615-2

RESEARCH

Unsupervised domain adaptation by cross-domain consistency learning for CT body composition

Shahzad Ali[1] · Yu Rim Lee[2] · Soo Young Park[2] · Won Young Tak[2] · Soon Ki Jung[1]

Received: 22 April 2024 / Revised: 17 August 2024 / Accepted: 30 September 2024 / Published online: 3 January 2025
© The Author(s) 2024

Abstract

Computed tomography (CT) scans of the abdomen have become the gold standard for assessing body composition (BC). Accurate estimation of skeletal muscle and adipose tissues from CT scan slices is crucial for diagnosis and prognosis. Much research in abdominal image analysis focuses on the third lumbar vertebra (L3) due to its stability and ease of labeling compared to other lumbar vertebrae. This study leverages labeled L3 slices (source domain) to predict unlabeled slices from thoracic T1 to sacrum S5 region (target domain). We proposed a Twin Encoder–Decoder Network (TED-Net) with distinct weight initialization employing Cross-domain Consistency Learning (CDCL) for joint training across the domains. This strategy extends the network's knowledge by enforcing consistency between predictions from two segmentation networks. The training objective includes supervised loss terms for the source domain and unsupervised loss terms for the target domain. This results in increases of 6.68%, 3.31%, and 4.40% in Precision, Dice Similarity Coefficient, and Intersection over Union, respectively, indicating significant improvement in performance on the target domain, suggesting that domain-invariant feature learning through cross-domain consistency learning enhances a network's adaptability over unlabeled domains.

Keywords Body composition · Unsupervised domain adaptation · CT segmentation · Skeletal muscle · Adipose tissues

1 Introduction

Computed tomography (CT) is widely used in medical imaging, providing detailed insights into the internal anatomical structures of the human body. Due to its extensive clinical use, much CT data is available for various research and analytical purposes. However, acquiring labeled CT data can be costly, leading to exploring alternative approaches in the literature. Semi-supervised learning (SSL) is one such approach that uses both labeled and unlabeled data, while unsupervised learning (UL) does not use any labeled data. Another approach, transfer learning (TL), involves leveraging data and models from similar domains.

Deep learning methods often assume that data are independent and identically distributed (IID), meaning that training and testing datasets are independent and have similar characteristics. However, variations in imaging devices, patient demographics, and study cohorts across medical centers can lead to significant differences between training and testing datasets. This phenomenon, known as domain shift, presents a challenge for real-world clinical applications. The related but distinct nature of these domains hinders the effectiveness of traditional supervised learning approaches in the target domain due to violating the IID assumption.

The source domain comprises labeled slices from the third lumbar vertebra (L3), while the target domain consists of unlabeled slices from the thoracic, lumbar, and sacrum vertebrae. The primary objective of this approach is to ensure that predictions are invariant across two lightweight encoder–decoder networks that are initialized with different weights. To ensure robustness against domain shifts, we enforce consistency between the predictions obtained from these encoder–decoder networks. Additional training signals extracted from the unlabeled data are combined with the supervised signals received from the labeled data to enhance the feature representations. To keep the total number of

✉ Soon Ki Jung
skjung@knu.ac.kr

Shahzad Ali
shahzadali@knu.ac.kr

[1] School of Computer Science and Engineering, Kyungpook National University, Daegu 41566, South Korea

[2] Department of Internal Medicine, College of Medicine, Kyungpook National University, Kyungpook National University Hospital, Daegu 41944, South Korea

trainable network parameters within reasonable limits, we utilize a twin lightweight encoder–decoder network. Moreover, during inference, we only use one encoder–decoder, which helps to reduce the computational overhead. It has been observed in a recent study [1] that network consistency tends to decrease when the training data is contaminated with noisy labels, leading to network overfitting. We propose using Jensen–Shannon divergence loss to address this issue and regularize the network. This loss function promotes consistency across multiple distributions spanning different domains, enhancing the network's resilience to noise. The proposed method is both simple and efficient, with the added advantage of being flexible in its capacity to integrate labeled and unlabeled CT data seamlessly within the framework of unsupervised domain adaptation. We conducted experiments on an in-house abdominal CT dataset to validate our approach, demonstrating its competitive performance across different domains. In this paper, we outline our contributions as follows:

- To the best of our knowledge, this work represents the first application of unsupervised domain adaptation (UDA) through Cross-Domain Consistency Learning (CDCL) specifically for analyzing body composition. It promotes the network's ability to learn by focusing on the consistency of its generated predictions.
- We propose a Twin Encoder–Decoder Network (TED-Net) for body composition analysis. TED-Net comprises two lightweight encoder–decoder networks, each initialized with distinct weights and designed specifically from a memory optimization perspective.
- We conduct joint training of TED-Net across diverse domains by utilizing unsupervised training signals derived from unlabeled CT data. This strategy enables a comprehensive and robust training process, effectively leveraging cross-domain information.
- The effectiveness of our method is demonstrated through extensive experimentation, including comparative analysis with supervised and transfer learning approaches. Unlike the previous studies, our evaluation includes complete CT volumes ranging from the thoracic (T1) to the sacrum region (S5) for a comprehensive body composition assessment.

The preliminary version of this work was first introduced at the International Symposium on Visual Computing (ISVC) 2023 [2]. The current version significantly expands upon the conference paper. Firstly, we offer a thorough review of related works to provide a broader context for our research. Secondly, we reevaluate our method using a larger dataset. Thirdly, we present detailed results and analysis and the application of CDCL from the source to the target domain. We specifically extend our evaluation to cover cases where CDCL is employed from the target to the source domain, and we compare performance between intra- and inter-domain scenarios. Additionally, we conduct ablation studies to assess performance with and without image augmentations and estimate labeled data's sufficiency.

2 Related works

2.1 CT body composition

Body composition (BC) quantifies body skeletal muscle and adipose (body fat) tissues and stands as a crucial modifiable risk factor influencing the clinical outcomes of chronic diseases, malignancies, and health risks associated with obesity [3–5]. Other clinical applications of BC include studying several types of cancers, cardiovascular, liver, and kidney diseases, inflammatory bowel disease, illnesses, and contrast dose adjustment [6]. The assessment of BC in research employs various methods such as densitometry, isotopic dilution, bioelectrical impedance, whole-body counting, neutron activation, dual-energy X-ray absorptiometry (DXA), computed tomography (CT) scans, magnetic resonance imaging (MRI), and spectroscopy [7, 8]. Among these methods, cross-sectional CT scans and MRI are widely regarded as the gold standards for BC analysis [9, 10]. Numerous studies have demonstrated a robust correlation between CT and MRI for BC assessment [11, 12]. However, CT scans are often preferred for their superior visualization of bones and dense tissues, faster imaging time, widespread availability, and cost-effectiveness. BC estimation based on a 2D slice from a vertebral location strongly correlates with the volumetric BC [13–15].

Intermuscular adipose tissue (IMAT), a minority adipose class, often results in limited segmentation accuracy. Existing research frequently overlooks IMAT entirely or partially during training, opting to derive it later through postprocessing techniques. AutoMATiCA [16] employs four U-Nets, each dedicated to a single adipose class, trained on 898 slices from a single L3 vertebral location. This method uses post-processing to remove misclassified pixels and combine predictions from all four networks. BodySegAI [17] trained a U-Net on 2989 slices collected from five different vertebral locations. Although IMAT was not explicitly used for training, they later applied thresholding on the predicted muscle class to infer the IMAT class. Comp2Comp [18] trained a U-Net using 400 slices from a single L3 vertebral location. While they included all adipose classes in training, post-processing was employed to fill holes and reassign predicted muscle pixels to IMAT by adjusting respective Hounsfield values.

Our proposed method addresses such limitations by achieving precise IMAT segmentation without training large

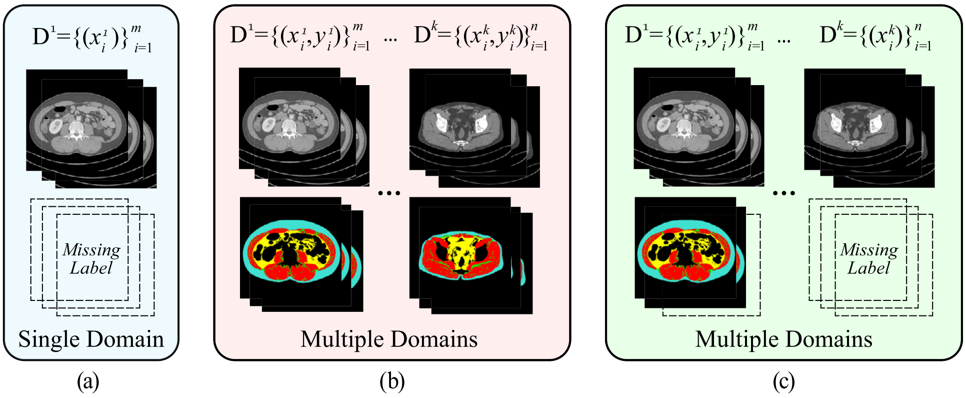

$D^1=\{(x_i^1)\}_{i=1}^m$
Missing Label
Single Domain
(a)
$D^1=\{(x_i^1,y_i^1)\}_{i=1}^m$ ... $D^k=\{(x_i^k,y_i^k)\}_{i=1}^n$
...
Multiple Domains
(b)
$D^1=\{(x_i^1,y_i^1)\}_{i=1}^m$ ... $D^k=\{(x_i^k)\}_{i=1}^n$
...
Missing Label
Multiple Domains
(c)

Fig. 1 Illustration of similar tasks, i.e., **a** Unsupervised learning, **b** Domain adaptation, and **c** Unsupervised domain adaptation. The labeled and unlabeled data is shown as (x_i^k, y_i^k) and (x_i^k) for $k = \{1, 2, \dots, k\}$ domains

networks and requiring post-processing steps. Furthermore, in contrast to prior studies that sample slices from predetermined vertebral positions, our network's training and evaluation encompassed the entirety of CT volumes (specifically, including 22 vertebral locations) to assess BC comprehensively.

2.2 Unsupervised domain adaptation

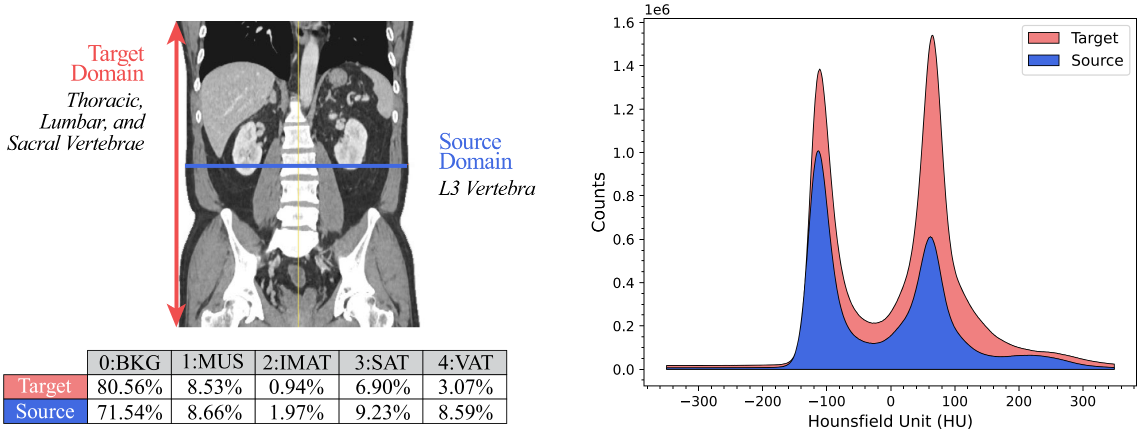
A domain refers to a probability distribution from which samples are drawn, and it can be classified as either a source or target domain. *Domain Adaptation* (DA) is a particular type of *Transfer Learning* (TL) that leverages labeled data from one or more related source domains to adapt to a new task in a target domain (see Fig. 1b). In comparison, *Unsupervised Domain Adaptation* (UDA) utilizes labeled data from the source domain and unlabeled data from the target domain (see Fig. 1c). It aims to enhance the performance of a deep learning network on an unlabeled target domain data [19–21] by minimizing the distribution gap from the related source domains. UDA has garnered substantial interest within the medical research community due to the scarcity of labeled data, often attributed to the demanding nature of expert knowledge and labor-intensive manual labeling. In this context, an inexpensive unlabeled dataset is considered the target domain, while a much smaller labeled dataset serves as a source domain. UDA methods can be categorized into three main approaches: *feature-level adaptation*, *image-* or *pixel-level adaptation*, or a combination of both. Feature-level adaptation methods aim at domain alignment by learning domain-invariant features [22–24], often employing domain adversarial training [25–27]. On the other hand, image-level adaptation methods minimize domain disparities by transforming the source domain images to match the style of the target domain [28–30]. Lastly, combining feature- and image-level adaptation techniques aims to simultaneously match the target domain style and reduce the domain gap [31–33].

Pseudo-labeling is a widely utilized UDA method for leveraging unlabeled data. It assigns pseudo labels to unlabeled target domain data using a pre-trained network on labeled source domain data. However, domain shifts introduce noise into these pseudo labels, and substantial research efforts are directed toward generating accurate and reliable pseudo labels to mitigate this issue [34–36]. The current UDA methodology predominantly depends on the accessibility of source domain data. However, there has been a notable shift towards Source-Free Unsupervised Domain Adaptation (SFUDA), which allows a pre-trained model on a labeled source domain to adapt to unlabeled target data. This approach is particularly advantageous in scenarios where the source data cannot be shared due to privacy concerns, storage and transmission costs, and potential computational overhead [37, 38].

2.3 Consistency learning

Semi-supervised learning methods leverage both fully and partially labeled data to improve network performance. Among these methods, consistency learning, or regularization, is widely employed. These approaches impose constraints on network predictions (or pseudo labels) to remain consistent under various types of perturbations that can be classified as *input-level* or *image-level*, *feature-level*, and *network-level*.

Input-level perturbations encompass both weak or strong data augmentations [39–41] and encourage the network to generate consistent predictions given different views of the same input. The assumption is that the various views of an input should map similarly in latent space; however, a recent work [42] treated augmented images from original images differently by increasing the distance between their features while assigning them the same semantic labels. Feature-level perturbations reduce the risk of overfitting and encourage generalization. Such perturbations involve techniques like dropout [43] that forces layers to co-adapt and rectify errors from prior layers, random noise injection [44]



	0:BKG	1:MUS	2:IMAT	3:SAT	4:VAT
Target	80.56%	8.53%	0.94%	6.90%	3.07%
Source	71.54%	8.66%	1.97%	9.23%	8.59%

(a)

(b)

Fig. 2 Overview of the source and target domains. **a** The coronal view of a male patient's CT scan, highlighting slices from the thoracic, lumbar, and sacral vertebrae ($T1, T2, ..., T12, L1, L2, ..., S5$) and the third lumbar vertebra $L3$ with their respective class distributions. **b** Histograms for the source and target domains are provided. Both domains exhibit two major peaks, delineated by contrasting heights, located approximately at − 110 Hounsfield Units (HU) and 65 HU. The discrepancy in peak magnitude underscores a notable domain shift between the source and target domains

to the encoder's output to map a data point differently in latent space, masking [45, 46] to help network pay more attention to less discriminative input regions, and virtual adversarial training [47] that measures local distribution smoothness of the network's output resulting from the small changes in its input. In addition to predicting binary segmentation maps [48], proposed predicting signed distance maps to capture geometric shape information in their uncertainty-guided mutual consistency framework. Network-level perturbations include techniques such as stochastic depth [49], shared encoder with slightly different decoders [50], identical networks with weight perturbations [51, 52], and mean teacher network [53] which maintain an exponential moving average of weights. Likewise [54], proposes complementary consistency between the pseudo-labels generated from a primary network and two auxiliary networks, which all have identical but independent encoders. Furthermore, recent work combines perturbations at multiple levels for performance gains like [55] proposed UniMatch to broaden the perturbations space by unifying image-level and feature-level perturbations. Han et al. [56] provides a taxonomy and comprehensive review of existing methods in their study.

This study employs network-level perturbation to investigate the efficacy of a twin encoder–decoder segmentation network with perturbed weights in enhancing cross-domain consistency and aiding network adaptation to the target domain.

3 Proposed method

The segmentation network demonstrates a notable decline in performance when exposed to out-of-distribution (OOD) data, slices explicitly from positions beyond the L3 vertebra. Relying solely on training from a single vertebral level leads to considerable prediction variability, rendering it an imperfect representation of overall BC. Figure 2a delineates the thoracic to sacral region and the L3 slice position within the coronal view of an abdominal CT scan acquired from a male patient. The distribution of classes in both domains is also tabulated. This study aims to procure reliable predictions for unlabeled slices of thoracic to sacral vertebrae utilizing only labeled slices of the L3 vertebra.

A grayscale CT image is generated by reconstructing radiation's absorption/attenuation coefficient within tissues. The Hounsfield Unit (HU) is derived via a linear transformation of the measured attenuation coefficient. Figure 2b depicts histograms illustrating the distribution of HU values obtained for both domains. Each domain exhibits two distinct peaks: one corresponding to adipose and the other to muscle mass. These peaks demonstrate a close alignment, albeit with differing magnitudes. For the source domain, the highest peak represents adipose tissues, encompassing the IMAT, SAT, and VAT classes. Adjacent to this is the second-highest peak corresponding to the muscle (MUS) class. In contrast to the source domain, the target domain is characterized by a predominant peak for the MUS class. We propose an unsupervised domain adaptation method to address the disparities between these domains, commonly called the domain shift problem. This approach enables a network to capture rich feature representations from unlabeled data.

3.1 Problem formulation

Let $\mathbb{D}^s = \{(x_i^s, y_i^s)\}_{i=1}^m$ represent the source domain with labeled data, and $\mathbb{D}^t = \{x_i^t\}_{i=1}^n$ the target domain consisting unlabeled data. Here, m and n are the given total numbers of slices in both domains. Each $x^* \in \mathbb{R}^{H\times W}$ is a 2D CT slice and when $x^s, y^s \in \mathbb{R}^{H\times W\times C}$ is a corresponding label, where C is the total number of classes. In the context of UDA, the goal is to improve the network's generalization ability by leveraging both domains $\mathbb{D}^s$ and $\mathbb{D}^t$ for the semantic segmentation task. The objective is to learn a function f that can effectively map input CT slices to the label space, i.e., $f : x^* \rightarrow y^*$. This function operates on slices drawn from any of the k source domains $\mathbb{D}^s = \{D^1, \ldots, D^k\}$ or the target domain. The amount of unlabeled data in the target domain can be significantly larger than the labeled data in the source domains as indicated by $|\mathbb{D}^t| \gg |\mathbb{D}^s|$.

3.1.1 Cross-domain consistency learning (CDCL)

To obtain additional training signal from the unlabeled data in $\mathbb{D}^t$, we ensure consistency between the outputs of two parallel encoder–decoder networks, namely $f(x^*;\theta_1)$ and $f(x^*;\theta_2)$, where θ_* represents the set of learning parameters. Each encoder–decoder network produces a prediction map for a given input. For labeled input slices x_m^s, we first compute soft dice coefficient score loss L_{dsc} and mean absolute error loss L_{mae} as the supervised loss terms:

$$L_{dsc}^s = \frac{1}{m}\sum_{i=1}^{m}\left(1 - \frac{2y_{1i}^s \cdot \tilde{y}_{1i}^s}{y_{1i}^s + \tilde{y}_{1i}^s}\right), \tag{1}$$

$$L_{mae}^s = \frac{1}{m}\sum_{i=1}^{m}|y_{1i}^s - \tilde{y}_{1i}^s|, \tag{2}$$

where y_{1i}^s is the target label and $\tilde{y}_{1i}^s$ is the predicted label obtained from $f(x;\theta_1)$ for x_1^s and m indicates total source domain slices. In the absence of target labels y_n^t corresponding to the inputs x_n^t, the labels predicted by both encoder–decoder networks are turned into the pseudo target labels, and unsupervised loss terms are calculated. This means by substituting $y_1^s \approx \tilde{y}_2^t$ and $y_2^s \approx \tilde{y}_1^t$ [52], we can calculate unsupervised loss terms L_{mae}^t and L_{jsd}^t.

Jensen–Shannon Divergence (JSD) has demonstrated remarkable effectiveness in learning feature representations from noisy labels, combating underfitting, and encouraging consistency regularization [1]. It is a symmetric and smoothed version of the Kullback–Leibler (KL) divergence, also known as relative entropy. For given two distributions $\tilde{y}_1^s$ and y_1^s, the divergence is defined as:

$$KLD(\tilde{y}_1^s \parallel y_1^s) = \sum_{i=1}^{m} y_{1i}^s \cdot (\log y_{1i}^s - \log \tilde{y}_{1i}^s). \tag{3}$$

In contrast to the divergence, JSD measures the similarity between these distributions as follows:

$$L_{jsd}^s(\tilde{y}_1^s \parallel y_1^s) = \frac{1}{2}KLD(\tilde{y}_1^s \parallel m) + \frac{1}{2}KLD(y_1^s \parallel m), \tag{4}$$

where $m = \frac{1}{2}(\tilde{y}_1^s + y_1^s)$ is the midpoint distribution. L_{jsd}^s ranges between [0, 1], where zero means the two distributions are identical, and one means they are completely dissimilar.

Dynamic loss weighing via task-dependent uncertainty, denoted as L_{cdc}, combines all loss terms calculated for the source or target domains, as specified by Eqs. (1), (2) and (4), as:

$$L_{cdc} = \sum_{L_\tau \in \mathbb{T}} \lambda_\tau L_\tau^*, \tag{5}$$

where $\mathbb{T}$ is a set of tasks or loss functions, and λ is a set of hyper-parameters denoting the weights assigned to each loss term. We empirically combined the five loss terms as:

$$L_{cdc} = \lambda_1 \cdot L_{dsc}^s + \lambda_2 \cdot (L_{jsd}^s + L_{mae}^s) + \lambda_3 \cdot (L_{jsd}^t + L_{mae}^t). \tag{6}$$

Generally, values for λ_* in the above equation are set as constants, which is not optimal. Instead, a dynamic weighting scheme is needed to exploit task-specific uncertainties that vary across $\mathbb{T}$. We adopt an idea similar to that first proposed by [57] and further refined by [58], which enforces positive regularization values. Under the assumption that a loss or a subset of loss functions is a task, task-dependent uncertainty turns the aggregated loss function in Eq. (6) into a learnable combined loss:

$$L_{cdc} = \sum_{L_\tau \in \mathbb{T}} \frac{1}{2\lambda_\tau^2} L_\tau^* + ln(1 + \lambda_\tau^2). \tag{7}$$

We initialize the weights λ_* to 0.5 at the beginning of network training. For the UDA of source to target domain, the learned values for these weights were $\lambda_1 = 0.3981$, $\lambda_2 = 0.2323$, and $\lambda_3 = 0.1547$.

3.2 Twin encoder–decoder network (TED-Net)

A twin encoder–decoder network (TED-Net), drawing inspiration from the architectural principles of U-Net [59] and ResNet [60], was implemented to constitute a pair of networks, namely $f(\theta_1)$ and $f(\theta_2)$ (Fig. 3a). Each network undergoes initialization with distinct weights to ensure

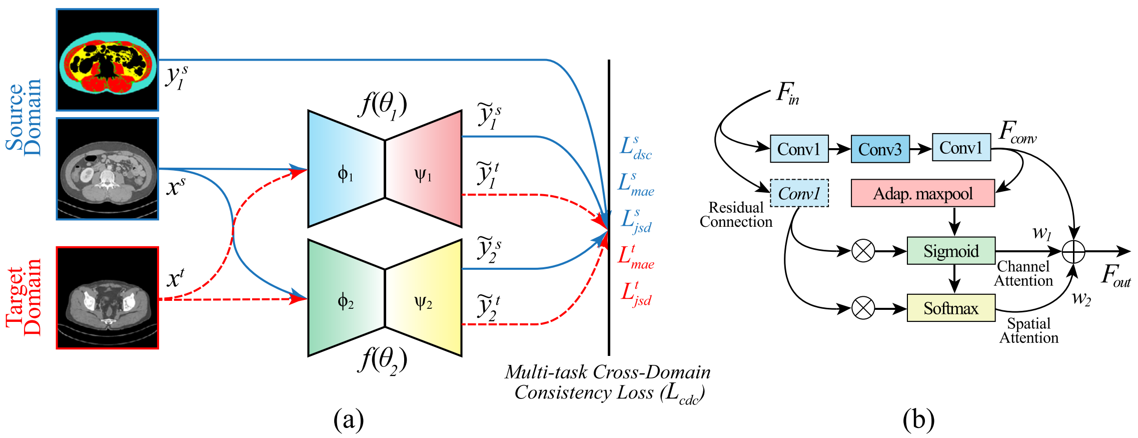


Fig. 3 Overview of the proposed TED-Net architecture. **a** It consists of twin encoder–decoder networks, denoted by $f(\theta_1)$ and $f(\theta_2)$, wherein supervised loss terms are calculated for the source domain $\mathbb{D}^s$ since label y_1^s is given. For target domain $\mathbb{D}^t$, unsupervised loss terms are calculated from predicted labels $\tilde{y}_1^t$ and $\tilde{y}_2^t$ used as the pseudo labels. **b** *Res-Conv*, the building block of TED-Net, with spatial and channel attention. The residual connection is indicated by a dashed box and is activated when the number of input and output channels differs

they begin learning trajectories from varied starting points. Instead of employing an explicit weight-sharing mechanism, adjustments to the learning trajectories are made through the cross-domain consistency loss (L_{cdc}) function as specified in Eq. (7).

In contrast to commonly available deeper or wider alternatives, opting for a network with fewer trainable parameters was motivated to facilitate training two analogous networks while preserving the original image resolution and managing memory constraints effectively. Moreover, the efficacy of this lightweight autoencoder has been previously demonstrated in the context of segmenting foot ulcer wounds [61]. The proposed CDCL method necessitates each encoder–decoder to make one feed-forward pass before initiating a single backpropagation step during training. In this context, TED-Net, the lightweight semantic segmentation network, is optimal for realizing such a training strategy.

The TED-Net encoders (ϕ_*) consist of 11 Residual-Convolution (Res-Conv) blocks and 5 MaxPooling layers. Conversely, both decoders (ψ_*) comprise 4 Res-Conv blocks and 5 Convolutional Transpose (ConvTranspose) layers. The structure of the Res-Conv block is illustrated in 3b. It comprises three convolution layers, each followed by batch normalization and GELU activation function, aiming to generate convolutional features denoted as F_{conv}. This block incorporates spatial and channel attention applied to the input F_{in} and max pooled convolutional features, respectively. These attentions are combined elementwise to generate the block output, F_{out}. Notably, the weights for these attention mechanisms are optimized through network training. The final prediction is derived through a 1×1 convolutional layer, succeeded by a softmax layer. The class label of each pixel is determined based on the highest probability value exceeding 0.5.

3.3 Preprocessing

Following [16], the initial step involves the application of thresholding to the intensity values of a raw DICOM slice. This process aims to confine the values within predefined Hounsfield Units (HU) ranges: $[-1024, 1024]$, $[-29, 150]$, and $[-190, -30]$ for the CT, MUS, SAT/VAT channels, respectively. The outcome of this thresholding operation manifests in distinct features highlighted within each channel. Subsequently, the resulting thresholded values collectively constitute a three-channel input network, as depicted in Fig. 4. The preprocessing phase is conducted offline, mitigating any overhead during network training.

4 Experiments

4.1 Experimental setup

4.1.1 Datsets

The proposed TED-Net has undergone comprehensive training and evaluation procedures utilizing two datasets comprising two-dimensional axial CT image slices. These datasets, obtained from Kyungpook National University Hospital (KNUH) in Daegu, South Korea, consist of 1004 and 424

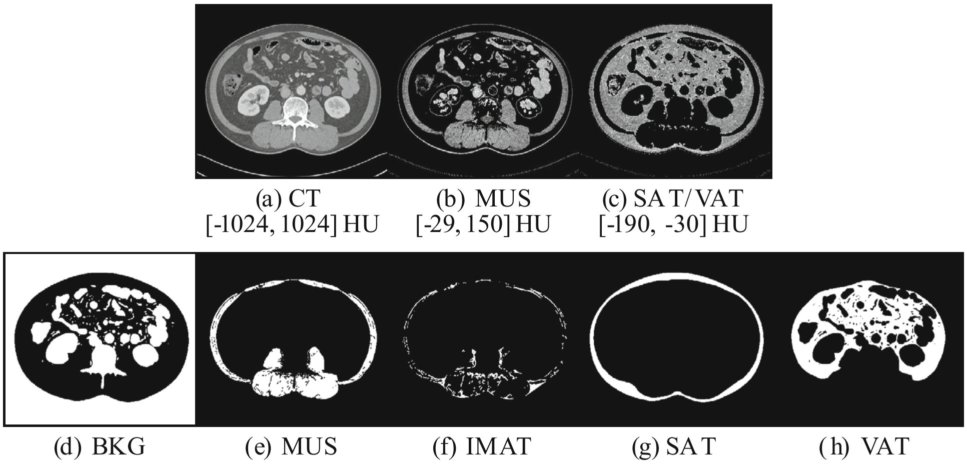


Fig. 4 Input and output of the TED-Net. **a**–**c** Three-channel input is formed by windowing raw CT pixels within specific Hounsfield Unit (HU) value ranges. **d**–**h** The network output comprises five channels representing background (BKG), skeletal muscle (MUS), intermuscular adipose tissue (IMAT), subcutaneous adipose tissue (SAT), and visceral adipose tissue (VAT), respectively

Table 1 Datasets used for this study

Domain	Dataset	Total	Train	Test
Source	L3	1004	804	200
Target	T1S5	424	338	86

retrospectively acquired slices corresponding to patients' CT scans conducted between 2020 and 2021[1].

The Third Lumbar Vertebra (L3) dataset contains labeled slices specifically positioned at the L3 level (i.e., single vertebral location). All slices were subsequently extracted and semi-automatically labeled [62, Fig. 1] by a radiologist, who identified five distinct classes within each slice. This labeling process was conducted under the supervision of experienced radiologists to ensure its accuracy and reliability. For this study, the labeled data primarily served as the source domain and was randomly divided into training and validation sets, consisting of 804 and 200 slices, respectively (refer to Table 1).

The Thoracic to Sacral Vertebrae(T1S5) dataset encompasses unlabeled slices sourced from a total of 22 vertebral locations spanning from the first thoracic vertebra (T1) to the fifth sacral vertebra (S5) along the vertebral column. This dataset has 424 labeled slices resulting from the aforementioned meticulous labeling process. During the network training phase, 338 of the 424 slices from this T1S5 dataset were employed for feed-forward passes without corresponding labels.

The TED-Net model underwent training using high-resolution inputs of 512×512 pixels, capitalizing on labels for five distinct classes, namely skeletal muscle (MUS), three types of fat tissue-intermuscular adipose tissue (IMAT), saturated adipose tissue (SAT), and visceral adipose tissue (VAT)-and lastly, a background class. Notably, downsampling or patch extraction was deliberately avoided to maintain the original image resolution, thereby achieving superior segmentation results.

4.1.2 Evaluation metrics

In our analysis, we have computed counting or overlap metrics and distance-based metrics to assess the performance of the proposed method. Counting metrics are calculated directly from a confusion matrix. We include Accuracy (ACC), Specificity (SPE), Precision (PRE), Recall/sensitivity (REC), Dice Similarity Coefficient (DSC), and Intersection over Union (IoU), which are defined as:

$$ACC = \frac{TN + TP}{TN + FP + FN + TP}, \tag{8}$$

$$SPE = \frac{TN}{TN + FP}, \tag{9}$$

$$PRE = \frac{TP}{TP + FP}, \tag{10}$$

$$REC = \frac{TP}{TP + FN}, \tag{11}$$

$$DSC = \frac{2TP}{2TP + FP + FN}, \text{ and} \tag{12}$$

$$IoU = \frac{TP}{TP + FP + FN}. \tag{13}$$

These metrics are computed based on the values of true positive (TP), false negative (FN), true negative (TN), and false positive (FP). It is crucial to note that the output range of these measures is between 0 and 1, where a higher score signifies better performance.

Distance-based metrics are calculated from TP and take into account the distance from each point on predicted boundry to the actual boundary are calculated and include Hausdorff Distance 95% percentile (HD95) and Average Sur-

[1] The release of this dataset to the public is subject to approval from the Institutional Review Board/Ethics Committee of Kyungpook National University Hospital and Kyungpook National University.

face Distance (ASD). HD95 measures the 95% percentile of all shortest distances from one object boundary to the other while ASD calculates average of all distances for each point from one object to another. Given two objects A and B, these metrics are defined as [63]:

$$HD95(A, B) = \max\{d_{95}(A, B), d_{95}(B, A)\}, \tag{14}$$

$$ASD(A, B) = \frac{1}{2}\left(\frac{\sum_{a\in A} d(a, B)}{|A|} + \frac{\sum_{b\in B} d(A, b)}{|B|}\right), \tag{15}$$

where d is a distance measure which is generally Euclidean, $d_{95}(A, B) = \underset{a\in A}{x\ {}_{95}}\ \{\min_{b\in B} d(a, b)\}$ and $d(a, B) = min_{b\in B} d(a, b)$. Equations (14) and (15) denote perfect prediction by 0 while there is no fixed upper bound to indicate otherwise. Binarization of predictions is a prerequisite for the calculation of these metrics. To determine the optimal threshold, we systematically explored values within the interval [0, 1] at increments of 0.01 and identified the threshold that maximized the DSC. Subsequently, final labels were assigned to each predicted pixel according to the class probability with the highest value.

4.2 Implementation details

We primarily investigated the proposed method for UDA's source-to-target and target-to-source modes. This section discusses the training and inference settings for the experiments.

4.2.1 Training procedure

One encoder–decoder network is initialized using the Kaiming weight initialization method, while the other utilizes the Xavier initialization method. TED-Net undergoes end-to-end training for 100 epochs, employing a batch size of 8 to optimize memory usage throughout the training process. At a given instance, a labeled and unlabeled slice is sequentially presented to TED-Net, following which the total loss is computed. This work employs a novel TED-Net and does not rely on pre-existing or widely used segmentation networks. Moreover, the CDCL experiments did not utilize pre-training or fine-tuning methods. The learning rate is fixed at 0.001, and the AdamW optimizer is used for its scale invariance property [64] and better generalization performance [65]. This is essential since the total loss obtained through task-dependent uncertainty was higher than the individual losses. All experiments were carried out on four NVIDIA RTX A4000 GPUs, each with 16 GB of memory, using an opensource machine learning framework PyTorch.

4.2.2 Inference

During the inference phase, we utilized a batch size of 1, averaging over the entire test fold rather than batches. Predictions were exclusively generated using a single encoder–decoder network $f(\theta_1)$.

5 Results

5.1 Quantitative results

The performance comparison between Transfer Learning (TL) and the proposed CDCL is detailed in Table 2, and to analyze the segmentation results, we compare the performance metrics DSC, IoU, HD95, and ASD for each class. Each metric is reported with its mean and standard deviation across five classes, and the last value represents the average performance across the foreground classes.

5.1.1 Source to target UDA

We implemented TED-Net, initially trained it on the source domain, and evaluated its performance on the target domain. For simplicity, this configuration is denoted as $D^s \rightarrow D^t$. Specifically, the datasets denoted as L3 and T1S5 served as the source and target domains, respectively. For the TL approach, first, we utilized the training fold of L3 to train the encoder–decoder network $f(\theta_1)$ in a supervised manner. Subsequently, we used it as a pre-trained model and evaluated its performance on the T1S5 dataset without fine-tuning. The obtained results thus set a lower bound for comparison with the least expected performance from CDCL. In a subsequent experiment, TED-Net was trained to utilize labeled data from the source domain and unlabeled data from the target domain with CDCL enabled.

From the upper half of Table 2, the incorporation of CDCL resulted in significant improvements over TL in foreground segmentation, evidenced by respective gains of 3.31% and 4.40% in terms of average DSC and IoU scores. Particularly noteworthy is the substantial enhancement observed in the DSC score for the minority class IMAT (78.59 ± 1.83), which increased by 11.48%. Additionally, the SAT and VAT classes exhibited improvements of 1.33% or more in their DSC scores, showing consistent segmentation performance by CDCL across the foreground classes. Corresponding improvements in IoU were recorded as 14.11%, 2.49%, and 2.19% for the IMAT, SAT, and VAT classes, respectively. A slight decrease of 0.96% in DSC scores for the MUS class is worth noting. That is because IMAT lies within the MUS class and has overlapped Hounsfield intensity ranges, making the delineation between them challenging even for annotators. Consequently, it can be attributed

Table 2 Comparison between the Transfer Learning (TL) and the proposed CDCL method

Mode	Class	DSC↑		IoU↑		HD95↓		ASD↓	
		TL	CDCL	TL	CDCL	TL	CDCL	TL	CDCL
$D^s \rightarrow D^t$	BKG	$\mathbf{98.90}_{\mathbf{0.32}}$	$98.73_{0.25}$	$\mathbf{97.84}_{\mathbf{0.49}}$	$97.53_{0.48}$	$\mathbf{17.11}_{\mathbf{2.48}}$	$62.46_{5.71}$	$\mathbf{0.54}_{\mathbf{0.03}}$	$0.61_{0.02}$
	MUS	$\mathbf{90.72}_{\mathbf{0.88}}$	$89.76_{0.42}$	$\mathbf{83.64}_{\mathbf{1.19}}$	$82.43_{0.62}$	$\mathbf{14.24}_{\mathbf{2.43}}$	$18.39_{1.31}$	$\mathbf{1.63}_{\mathbf{0.12}}$	$2.73_{0.14}$
	IMAT	$67.11_{2.88}$	$\mathbf{78.59}_{\mathbf{1.83}}$	$51.44_{3.24}$	$\mathbf{65.55}_{\mathbf{2.42}}$	$20.17_{1.39}$	$\mathbf{14.68}_{\mathbf{1.34}}$	$\mathbf{1.27}_{\mathbf{0.10}}$	$2.57_{0.16}$
	SAT	$95.55_{0.63}$	$\mathbf{96.88}_{\mathbf{0.37}}$	$91.91_{1.03}$	$\mathbf{94.40}_{\mathbf{0.63}}$	$17.82_{1.02}$	$\mathbf{16.79}_{\mathbf{0.94}}$	$2.03_{0.22}$	$\mathbf{1.81}_{\mathbf{0.13}}$
	VAT	$87.64_{3.84}$	$\mathbf{89.06}_{\mathbf{3.09}}$	$82.49_{5.05}$	$\mathbf{84.68}_{\mathbf{3.86}}$	$\mathbf{25.47}_{\mathbf{4.42}}$	$28.59_{4.29}$	$2.41_{1.29}$	$\mathbf{1.26}_{\mathbf{1.21}}$
	AVG	$85.26_{1.52}$	$\mathbf{88.57}_{\mathbf{1.11}}$	$77.37_{1.88}$	$\mathbf{81.77}_{\mathbf{1.47}}$	$\mathbf{19.42}_{\mathbf{1.23}}$	$19.61_{1.12}$	$\mathbf{1.83}_{\mathbf{0.32}}$	$2.09_{0.27}$
$D^t \rightarrow D^s$	BKG	$99.57_{0.05}$	$\mathbf{99.60}_{\mathbf{0.03}}$	$99.14_{0.09}$	$\mathbf{99.19}_{\mathbf{0.07}}$	$\mathbf{5.86}_{\mathbf{0.09}}$	$34.51_{0.36}$	$\mathbf{0.32}_{\mathbf{0.54}}$	$0.42_{0.01}$
	MUS	$94.98_{0.20}$	$\mathbf{95.24}_{\mathbf{0.24}}$	$90.48_{0.37}$	$\mathbf{90.95}_{\mathbf{0.43}}$	$\mathbf{4.11}_{\mathbf{0.21}}$	$6.02_{0.52}$	$\mathbf{0.61}_{\mathbf{0.01}}$	$0.71_{0.01}$
	IMAT	$65.53_{0.49}$	$\mathbf{66.89}_{\mathbf{0.65}}$	$49.00_{0.57}$	$\mathbf{50.54}_{\mathbf{0.76}}$	$\mathbf{12.89}_{\mathbf{0.28}}$	$17.31_{0.49}$	$\mathbf{3.03}_{\mathbf{0.08}}$	$4.33_{0.11}$
	SAT	$97.47_{0.27}$	$\mathbf{97.84}_{\mathbf{0.17}}$	$95.15_{0.47}$	$\mathbf{95.83}_{\mathbf{0.30}}$	$13.82_{0.74}$	$\mathbf{13.01}_{\mathbf{0.20}}$	$\mathbf{0.41}_{\mathbf{0.02}}$	$1.56_{0.05}$
	VAT	$95.78_{0.19}$	$\mathbf{96.11}_{\mathbf{0.27}}$	$92.16_{0.35}$	$\mathbf{92.70}_{\mathbf{0.48}}$	$\mathbf{9.01}_{\mathbf{0.65}}$	$10.04_{0.66}$	$\mathbf{0.73}_{\mathbf{0.04}}$	$1.07_{0.03}$
	AVG	$88.44_{0.16}$	$\mathbf{89.02}_{\mathbf{0.24}}$	$81.70_{0.20}$	$\mathbf{82.50}_{\mathbf{0.33}}$	$\mathbf{9.95}_{\mathbf{0.41}}$	$11.60_{0.38}$	$\mathbf{1.19}_{\mathbf{0.03}}$	$1.92_{0.03}$

TED-Net is trained on the labeled and unlabeled training folds from both domains. Only $f(\theta_1)$ is employed for evaluation. The best values are in bold

to better prediction of IMAT pixels intersecting with the MUS class, previously misclassified by the TL approach. Though the TL approach generally performed better for the MUS class (90.72 ± 0.88 for TL and 89.76 ± 0.42 for CDCL), CDCL significantly outperformed TL for IMAT (IoU of 65.55 ± 2.42 vs 51.44 ± 3.24) and SAT classes (IoU of 91.91 ± 1.03 vs 94.40 ± 0.63). We conducted a pair-wise Wilcoxon signed-rank test on all slices in the target domain to find the statistical significance of performance obtained from both methods. The obtained p-values for each class, 3.72×10^{-4}, 1.74×10^{-1}, 9.80×10^{-71}, 2.67×10^{-55}, and 1.37×10^{-8}, were less than 0.05. This shows that the performance gains from CDCL are unlikely due to random chance; hence, it is superior to the TL method.

The average foreground HD95 is comparable between the two approaches (19.42 ± 1.23 for TL and 19.61 ± 1.12 for CDCL), whereas the average foreground ASD is slightly better for TL (1.83 ± 0.32) than CDCL (2.09 ± 0.27). HD95 and ASD calculated for TL performed better for MUS, whereas CDCL performed better for the SAT class. CDCL has higher ASD values for most classes (e.g., 2.73 ± 0.14 for MUS and 2.57 ± 0.16 for IMAT) compared to TL (1.63 ± 0.12 for MUS and 1.27 ± 0.10 for IMAT), suggesting less precise boundary segmentation. Penalizing boundary inaccuracies heavily in the final loss (L_{cdc}), employing multi-scale feature fusion, boundary-specific augmentations, dropout regularization, or morphological operations in post-processing could help in reducing the boundary errors, thereby lowering the distance-based metrics across all classes.

5.1.2 Target to source UDA

We conducted another experiment wherein we swapped the source and target domains, thus altering the experimental configuration to $D^t \rightarrow D^s$, following the methodology outlined in Sect. 5.1.1. From the lower half of Table 2, CDCL demonstrated slightly higher average DSC (89.02 ± 0.24 compared to 88.44 ± 0.16 from TL) and IoU (82.50 ± 0.33 compared to 81.70 ± 0.20 from TL). Specifically, IMAT and SAT are improved with DSC scores of 95.24 ± 0.24 and 66.89 ± 0.65, respectively, compared to 94.98 ± 0.20 and 65.53 ± 0.49 for TL. For HD95, CDCL showed better performance for VAT (13.01 ± 0.20) compared to TL (13.82 ± 0.74), indicating more accurate boundary delineation in this class. The observations of significantly higher HD95 values for the background class and higher ASD values overall are consistent with those made in the previous experiment. P-values for each class in the target domain were calculated through the pair-wise Wilcoxon signed-rank test to analyze the performance of both methods. The results in 6.38×10^{-4}, 2.83×10^{-39}, 5.87×10^{-96}, 9.58×10^{-125}, and 1.02×10^{-13} being less than 0.05 showed highly significant differences and stronger evidence against the null hypothesis.

The observed improvements in average DSC and IoU scores were modest, with an increase of merely 0.58% and 0.80%, respectively. This marginal improvement contrasts with the more substantial gains observed in the case of $D^s \rightarrow D^t$. A plausible explanation could be attributed to the exacerbated class imbalance among adipose classes within the T1S5 datasets as indicated by Fig. 2b. It is noteworthy that the same classes had previously demonstrated notable

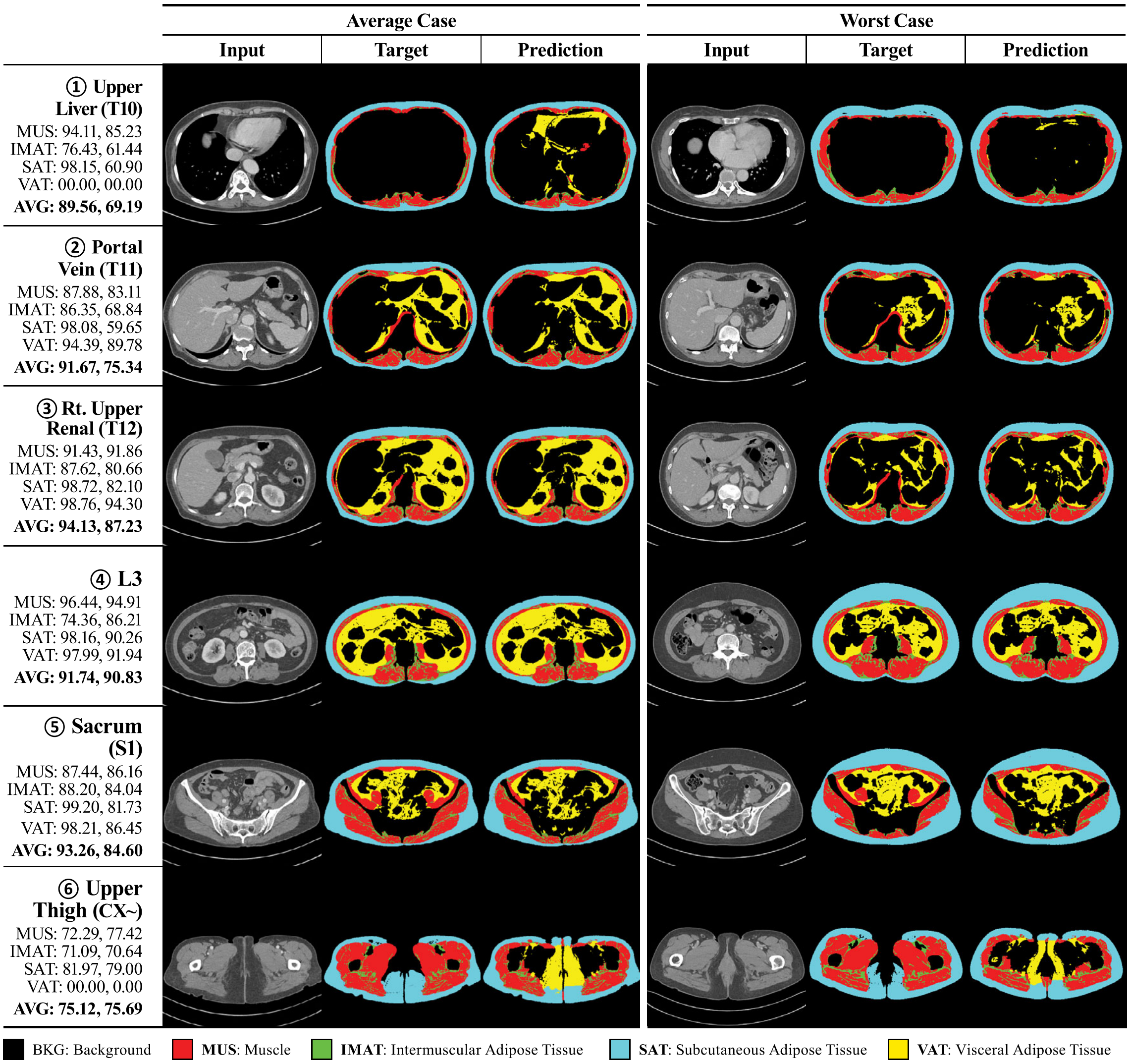


Fig. 5 Qualitative results for six specific vertebral locations of two subjects from the T1S5 dataset indicating an average and poor case with their respective class-wise DSC in the first column

improvements through the CDCL technique. Thus, besides the abundant unlabeled data, our findings underscore the pivotal role of source domain selection in determining the efficacy of the proposed CDCL method. Whenever feasible, the selection of the source domain should prioritize the domain exhibiting a balanced distribution of classes.

5.2 Qualitative results

We calculated the DSC among all patients in the target domain to identify average and worst-case performance scenarios. Figure 5 illustrates six slices from female subjects aged 68 and 53 years and their corresponding class-wise DSC scores obtained for the source to target domain UDA. The slices were selected sequentially along the vertebral column from six specific regions: upper liver, portal vein, right upper renal, L3, sacrum, and upper thigh, shown as the rows in Fig. 5. Including L3 slices from the source domain ensures completeness and facilitates better understanding. For clarity, the input CT slices, target labels, and predicted labels for the first subject are depicted under the Average Case, while those for the second subject are shown under the Worst Case.

The proposed method demonstrates superiority over similar approaches [16, 17] with an average DSC of 91.51%, 93.45%, 93.07%, 92.55%, and 73.65% for MUS, IMAT, SAT, and VAT, respectively. The lower DSC for the upper thigh region is primarily caused by misclassification of the genital area as VAT in most instances and as SAT in a few cases instead of being classified as background. This misclassification is attributed to the low contrast in HU values between these regions. Similarly, the upper liver region also shows a lower DSC for the same reason. Our study also found that the

Table 3 Comparison between the intra-domain and inter-domain cases

Mode	Method	Dataset		Performance Metrics							
		Train	Test	ACC↑	SPE↑	PRE↑	REC↑	DSC↑	IoU↑	HD95↓	ASD↓
$D^s \rightarrow D^s$	SL	L3	L3	$99.51_{0.02}$	$99.73_{0.01}$	$91.91_{0.91}$	$91.11_{1.00}$	$91.27_{1.03}$	$85.42_{1.52}$	$7.99_{0.18}$	$0.69_{0.07}$
$D^s \rightarrow D^t$	TL	L3	T1S5	$99.14_{0.13}$	$99.64_{0.08}$	$84.40_{1.41}$	$88.28_{1.05}$	$85.26_{1.52}$	$77.37_{1.88}$	$\mathbf{19.42}_{1.23}$	$\mathbf{1.83}_{0.32}$
	CDCL	L3/T1S5	T1S5	$\mathbf{99.23}_{0.10}$	$\mathbf{99.79}_{0.03}$	$\mathbf{91.08}_{1.15}$	$\mathbf{88.60}_{0.74}$	$\mathbf{88.57}_{1.11}$	$\mathbf{81.77}_{1.47}$	$19.61_{1.12}$	$2.09_{0.27}$
$D^t \rightarrow D^t$	SL	T1S5	T1S5	$99.76_{0.06}$	$99.86_{0.03}$	$94.11_{1.44}$	$94.18_{1.48}$	$93.73_{1.57}$	$89.57_{2.47}$	$5.01_{1.16}$	$0.43_{0.23}$
$D^t \rightarrow D^s$	TL	T1S5	L3	$99.28_{0.06}$	$99.62_{0.03}$	$92.85_{0.29}$	$86.39_{0.11}$	$88.44_{0.16}$	$81.70_{0.20}$	$\mathbf{9.95}_{0.41}$	$\mathbf{1.19}_{0.03}$
	CDCL	T1S5/L3	L3	$\mathbf{99.33}_{0.04}$	$\mathbf{99.64}_{0.02}$	$\mathbf{93.15}_{0.36}$	$\mathbf{87.07}_{0.19}$	$\mathbf{89.02}_{0.24}$	$\mathbf{82.50}_{0.33}$	$11.60_{0.38}$	$1.92_{0.03}$

SL, TL, and CDCL refer to Supervised Learning, Transfer Learning, and Cross-Domain Consistency Learning methods. The bold values indicate the best scores among the TL and CDCL methods

DSC for the minority foreground class, IMAT, was highest at L3 and decreased as we moved further away from it. This was anticipated, given that the source domain exclusively comprises L3 slices. Consequently, the network is inclined to be more familiar with slices exhibiting a similar class distribution, which contrasts with the characteristics found in the distant regions of the upper liver and upper thigh.

5.3 Ablation studies

5.3.1 Intra-domain and Inter-domain performance comparison

The two intra-domain cases, namely, the source domain to the source domain $D^s \rightarrow D^s$ and the target domain to the target domain $D^t \rightarrow D^t$, represent the upper limits of performance, as shown in the first and fourth rows in Table 3. Conversely, the two inter-domain scenarios source to the target domain $D^s \rightarrow D^t$ and target to the source domain $D^t \rightarrow D^s$, conducted with TL, establish the lower limits of performance.

It is evident that both CDCL and TL did not achieve the performance levels of supervised learning, which aligns with expectations. However, CDCL consistently surpassed TL across all counting metrics. Specifically, when L3 was the source domain and T1S5 was the target domain, CDCL demonstrated notable performance improvements of 6.68%, 3.31%, and 4.40% in precision, DSC, and IoU scores, respectively. Performance gains were also noted when T1S5 and L3 were reversed as the source and target domains; however, these gains were not as significant as those observed in the previous case. In terms of distance-based metrics, the average HD95 for CDCL was 19.61 ± 1.12, which was comparable to that obtained for the TL method, which was 19.42 ± 1.23. Similarly, the average ASD was 19.61 ± 1.12 for the TL method and 19.42 ± 1.23 for CDCL. These findings are consistent with the observations discussed in the preceding sections.

Table 4 Results for source to target domain UDA with and without augmentation

Augmentation	Performance Metrics					
	ACC↑	SPE↑	PRE↑	REC↑	DSC↑	IoU↑
✓	99.17	99.75	89.3	87.82	87.27	79.97
	99.22	**99.79**	**90.95**	**89.9**	**88.69**	**81.93**

5.3.2 Image augmentations for performance enhancement

In this study, we leveraged the image augmentations suggested in [66] to enhance performance. Our experimentation encompassed various transformation types and policies, involving the random application of 1 to n transformations, where $n = 8$ represents the total number of transformations. Furthermore, we varied the probability of applying these transformations and explored their application on source, target, or both domains. These transformations include sharpness from 0.1 to 0.5, Gaussian blur ($k = [3, 11]$), Gaussian noise ($\mu = 0, \sigma = 0.02$), brightness, contrast, and gamma adjustments in the range [0.85, 1.15], horizontal flips, resized crop with scales ranging from 0.75 to 1.25 and aspect ratios from 0.84 to 1.15.

Despite thorough exploration, we observed negligible performance enhancements with CDCL (Table 4). Surprisingly, the network failed to derive substantial benefits from the abovementioned strategies. Although we meticulously adjusted the hyperparameters to ensure that the transformed output images closely resembled the inputs, the network struggled to extract meaningful representations, hindering performance improvements. This discrepancy likely stems from the network's high regularization, which penalizes deviations from normal class distributions.

5.3.3 Effect of source domain size on performance

We assessed our method's performance using varying amounts of labeled training data from the source domain, ranging from 0 to 100% (804 slices) in 5% increments (40 slices). Figure 6

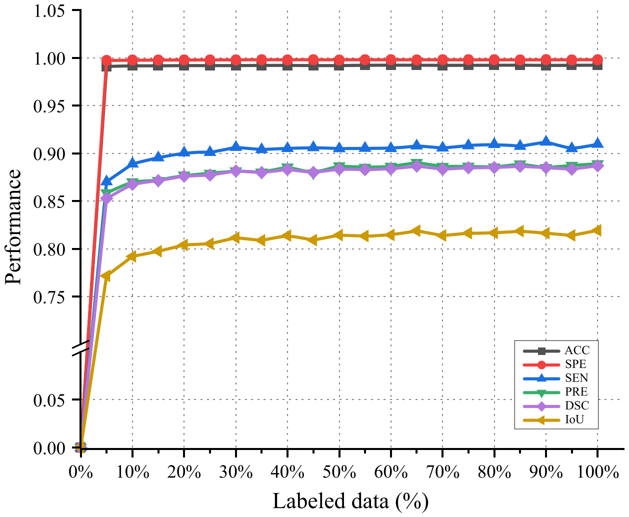


Fig. 6 Comparison of network performance with the percentage of labeled data from the training fold of the source domain when utilized to train TED-Net

illustrates the relationship between the percentage of labeled data utilized for network training and the resulting performance. The network was trained for 100 epochs with fixed hyperparameters. We observed that the network converged much earlier when trained on higher amounts of labeled data than on lower amounts. The highest DSC achieved was 88.69% when all labeled data was utilized. Using this as a reference point, we noticed that further improvements in DSC were marginal for labeled data beyond 20%, increasing by approximately 1%, and beyond 30%, rising by around 0.5%. However, when the labeled data is less than 30% of the unlabeled data, the network could be trained for more epochs to achieve reasonable performance.

6 Conclusions

We demonstrated that our proposed CDCL method provided reliable and detailed predictions for unlabeled thoracic, lumbar, and sacral vertebrae slices using only labeled L3 vertebra slices. It is particularly effective when dealing with unlabeled or noisy data, as it utilizes generated pseudo-labels to supplement missing labels, thereby expanding the labeled training set. The performance achieved for foreground classes is particularly remarkable, with the minority class of IMAT benefiting the most. These findings imply that integrating our proposed network into existing architectures could improve their adaptability to unlabeled target domains.

Although our proposed method exhibited notable improvement in utilizing unlabeled CT scans from another domain, it is crucial to recognize its limitations. The lack of an accessible public dataset for body composition precluded validating the proposed method using external data. Another limitation could be the network training using a cohort inclusive of diverse age groups. Future research could address these constraints and enhance the applicability of the proposed method.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00360226) and also supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2024-00437718) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

Author Contributions A.S.—Conceptualized, developed the deep learning model, designed experiments, reviewed literature, and wrote the manuscript. L. Y.—Curated and analyzed data and interpreted the results. P. S.—Curated and analyzed data and interpreted the results. T. W. - Supervised the annotation process and provided critical feedback. J. S.—Supervised the research, secured funding, contributed to experimental design, and revised the manuscript.

Data availibility No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Englesson, E., Azizpour, H.: Generalized Jensen–Shannon divergence loss for learning with noisy labels. Adv. Neural Inf. Process. Syst. **34**, 30284–30297 (2021)
2. Ali, S., Lee, Y.R., Park, S.Y., Tak, W.Y., Jung, S.K.: Volumetric body composition through cross-domain consistency training for unsupervised domain adaptation. In: International Symposium on Visual Computing, pp. 289–299. Springer (2023)
3. Tagliafico, A.S., Bignotti, B., Torri, L., Rossi, F.: Sarcopenia: how to measure, when and why. Radiol. Med. (Torino) **127**(3), 228–237 (2022)
4. Roh, E., Choi, K.M.: Health consequences of sarcopenic obesity: a narrative review. Front. Endocrinol. **11**, 530178 (2020)
5. Marzetti, E., Calvani, R., Tosato, M., Cesari, M., Di Bari, M., Cherubini, A., Collamati, A., D'Angelo, E., Pahor, M., Bernabei, R.: Sarcopenia: an overview. Aging Clin. Exp. Res. **29**, 11–17 (2017)

6. Elhakim, T., Trinh, K., Mansur, A., Bridge, C., Daye, D.: Role of machine learning-based CT body composition in risk prediction and prognostication: current state and future directions. Diagnostics **13**(5), 968 (2023)
7. Tosato, M., Marzetti, E., Cesari, M., Savera, G., Miller, R.R., Bernabei, R., Landi, F., Calvani, R.: Measurement of muscle mass in sarcopenia: from imaging to biochemical markers. Aging Clin. Exp. Res. **29**, 19–27 (2017)
8. Kullberg, J., Brandberg, J., Angelhed, J.-E., Frimmel, H., Bergelin, E., Strid, L., Ahlström, H., Johansson, L., Lönn, L.: Whole-body adipose tissue analysis: comparison of MRI, CT and dual energy x-ray absorptiometry. Br. J. Radiol. **82**(974), 123–130 (2009)
9. Lee, K., Shin, Y., Huh, J., Sung, Y.S., Lee, I.-S., Yoon, K.-H., Kim, K.W.: Recent issues on body composition imaging for sarcopenia evaluation. Korean J. Radiol. **20**(2), 205–217 (2019)
10. Ponti, F., Santoro, A., Mercatelli, D., Gasperini, C., Conte, M., Martucci, M., Sangiorgi, L., Franceschi, C., Bazzocchi, A.: Aging and imaging assessment of body composition: from fat to facts. Front. Endocrinol. **10**, 488049 (2020)
11. Zaffina, C., Wyttenbach, R., Pagnamenta, A., Grasso, R.F., Biroli, M., Del Grande, F., Rizzo, S.: Body composition assessment: comparison of quantitative values between magnetic resonance imaging and computed tomography. Quant. Imaging Med. Surg. **12**(2), 1450 (2022)
12. Faron, A., Sprinkart, A.M., Kuetting, D.L., Feisst, A., Isaak, A., Endler, C., Chang, J., Nowak, S., Block, W., Thomas, D.: Body composition analysis using CT and MRI: intra-individual intermodal comparison of muscle mass and myosteatosis. Sci. Rep. **10**(1), 11765 (2020)
13. Shen, W., Punyanitya, M., Wang, Z., Gallagher, D., St.-Onge, M.-P., Albu, J., Heymsfield, S.B., Heshka, S.: Total body skeletal muscle and adipose tissue volumes: estimation from a single abdominal cross-sectional image. J. Appl. Physiol. **97**(6), 2333–2338 (2004)
14. Shen, W., Punyanitya, M., Wang, Z., Gallagher, D., St-Onge, M.-P., Albu, J., Heymsfield, S.B., Heshka, S.: Visceral adipose tissue: relations between single-slice areas and total volume. Am. J. Clin. Nutr. **80**(2), 271–278 (2004)
15. Zopfs, D., Theurich, S., Grosse Hokamp, N., Knuever, J., Gerecht, L., Borggrefe, J., Schlaak, M., Santos, D.: Single-slice CT measurements allow for accurate assessment of sarcopenia and body composition. Eur. Radiol. **30**, 1701–1708 (2020)
16. Paris, M.T., Tandon, P., Heyland, D.K., Furberg, H., Premji, T., Low, G., Mourtzakis, M.: Automated body composition analysis of clinically acquired computed tomography scans using neural networks. Clin. Nutr. **39**, 3049–3055 (2020)
17. Alavi, D.H., Sakinis, T., Henriksen, H.B., Beichmann, B., Fløtten, A.-M., Blomhoff, R., Lauritzen, P.M.: Body composition assessment by artificial intelligence from routine computed tomography scans in colorectal cancer: introducing bodysegai. JCSM Clin. Rep. **7**(3), 55–64 (2022)
18. Blankemeier, L., Desai, A., Chaves, J.M.Z., Wentland, A., Yao, S., Reis, E., Jensen, M., Bahl, B., Arora, K., Patel, B.N., et al.: Comp2comp: open-source body composition assessment on computed tomography. arXiv preprint arXiv:2302.06568 (2023)
19. Liu, X., Yoo, C., Xing, F., Oh, H., El Fakhri, G., Kang, J.-W., Woo, J., et al.: Deep unsupervised domain adaptation: a review of recent advances and perspectives. APSIPA Trans. Sign. Inf. Process. **11**(1) (2022)
20. Zhou, K., Loy, C.C., Liu, Z.: Semi-supervised domain generalization with stochastic stylematch. arXiv preprint arXiv:2106.00592 (2021)
21. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Trans. Biomed. Eng. **69**(3), 1173–1185 (2021)
22. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014)
23. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: International Conference on Machine Learning, pp. 97–105. PMLR (2015)
24. Sun, B., Feng, J., Saenko, K.: Correlation alignment for unsupervised domain adaptation. Domain Adapt. Comput. Vis. Appl. (2017). https://doi.org/10.1007/978-3-319-58347-1_8
25. Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., : Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25–30, 2017, Proceedings 25, pp. 597–609, Springer (2017)
26. Ouyang, C., Kamnitsas, K., Biffi, C., Duan, J., Rueckert, D.: Data efficient unsupervised domain adaptation for cross-modality image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22, pp. 669–677, Springer (2019). https://doi.org/10.1007/978-3-030-32245-8_74
27. Chen, H., Jiang, Y., Loew, M., Ko, H.: Unsupervised domain adaptation based Covid-19 CT infection segmentation network. Appl. Intell. **52**(6), 6340–6353 (2022)
28. Ren, C.-X., Liu, Y.-H., Zhang, X.-W., Huang, K.-K.: Multi-source unsupervised domain adaptation via pseudo target domain. IEEE Trans. Image Process. **31**, 2122–2135 (2022)
29. Zheng, S., Yang, X., Wang, Y., Ding, M., Hou, W.: Unsupervised cross-modality domain adaptation network for x-ray to CT registration. IEEE J. Biomed. Health Inform. **26**(6), 2637–2647 (2021)
30. Li, R., Jiao, Q., Cao, W., Wong, H.-S., Wu, S.: Model adaptation: Unsupervised domain adaptation without source data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9641–9650 (2020)
31. Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: International Conference on Machine Learning, pp. 1989–1998, Pmlr (2018)
32. Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A.: Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. IEEE Trans. Med. Imaging **39**(7), 2494–2505 (2020)
33. Ge, Y., Chen, Z.-M., Zhang, G., Heidari, A.A., Chen, H., Teng, S.: Unsupervised domain adaptation via style adaptation and boundary enhancement for medical semantic segmentation. Neurocomputing **550**, 126469 (2023)
34. Wang, Y., Wang, H., Shen, Y., Fei, J., Li, W., Jin, G., Wu, L., Zhao, R., Le, X.: Semi-supervised semantic segmentation using unreliable pseudo-labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4248–4257 (2022)
35. Wu, H., Li, X., Lin, Y., Cheng, K.-T.: Compete to win: enhancing pseudo labels for barely-supervised medical image segmentation. IEEE Trans. Med. Imaging **42**(11), 3244–3255 (2023)
36. Shen, Z., Cao, P., Yang, H., Liu, X., Yang, J., Zaiane, O.R.: Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, pp. 4199–4207 (2023)
37. Tian, Q., Sun, C.: Rethinking confidence scores for source-free unsupervised domain adaptation. Neural Comput. Appl. (2024). https://doi.org/10.1007/s00521-024-09867-9
38. Fang, Y., Yap, P.-T., Lin, W., Zhu, H., Liu, M.: Source-free unsupervised domain adaptation: a survey. Neural Netw. (2024). https://doi.org/10.1016/j.neunet.2024.106230
39. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In:

International Conference on Machine Learning, pp. 1597–1607, PMLR (2020)
40. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.-L.: Fixmatch: simplifying semi-supervised learning with consistency and confidence. Adv. Neural. Inf. Process. Syst. **33**, 596–608 (2020)
41. Melas-Kyriazi, L., Manrai, A.K.: Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12435–12445 (2021)
42. Fan, Y., Kukleva, A., Dai, D., Schiele, B.: Revisiting consistency regularization for semi-supervised learning. Int. J. Comput. Vis. **131**(3), 626–643 (2023)
43. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
44. Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12674–12684 (2020)
45. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
46. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6023–6032 (2019)
47. Miyato, T., Maeda, S.-I., Koyama, M., Ishii, S.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Trans. Patt. Anal. Mach. Intell. **41**(8), 1979–1993 (2018)
48. Zhang, Y., Jiao, R., Liao, Q., Li, D., Zhang, J.: Uncertainty-guided mutual consistency learning for semi-supervised medical image segmentation. Artif. Intell. Med. **138**, 102476 (2023)
49. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, pp. 646–661, Springer (2016)
50. Wu, Y., Xu, M., Ge, Z., Cai, J., Zhang, L.: Semi-supervised left atrium segmentation with mutual consistency training. In: Medical Image Computing and Computer Assisted intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24, pp. 297–306, Springer (2021)
51. Ke, Z., Qiu, D., Li, K., Yan, Q., Lau, R.W.: Guided collaborative training for pixel-wise semi-supervised learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16, pp. 429–445, Springer (2020)
52. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2613–2622 (2021)
53. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Adv. Neural Inf. Process. Syst. **30** (2017)
54. Huang, H., Chen, Z., Chen, C., Lu, M., Zou, Y.: Complementary consistency semi-supervised learning for 3d left atrial image segmentation. Comput. Biol. Med. **165**, 107368 (2023)
55. Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y.: Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7236–7246 (2023)
56. Han, K., Sheng, V.S., Song, Y., Liu, Y., Qiu, C., Ma, S., Liu, Z.: Deep semi-supervised learning for medical image segmentation: a review. Exp. Syst. Appl. (2024). https://doi.org/10.1016/j.eswa.2023.123052
57. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, pp. 7482–7491 (2018)
58. Liebel, L., Körner, M.: Auxiliary tasks in multi-task learning. arXiv preprint arXiv:1805.06334 (2018)
59. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241, Springer (2015)
60. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
61. Ali, S., Mahmood, A., Jung, S.K.: Lightweight encoder–decoder architecture for foot ulcer segmentation. In: Communications in Computer and Information Science 1578 CCIS, pp. 242–253 (2022). https://doi.org/10.1007/978-3-031-06381-7_17/TABLES/2
62. Pu, L., Gezer, N.S., Ashraf, S.F., Ocak, I., Dresser, D.E., Dhupar, R.: Automated segmentation of five different body tissues on computed tomography using deep learning. Med. Phys. **50**(1), 178–191 (2023)
63. Reinke, A., Tizabi, M.D., Sudre, C.H., Eisenmann, M., Rädsch, T., Baumgartner, M., Acion, L., Antonelli, M., Arbel, T., Bakas, S., et al.: Common limitations of image processing metrics: a picture story. arXiv preprint arXiv:2104.05642 (2021)
64. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
65. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
66. Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B.J., Roth, H., Myronenko, A., Xu, D.: Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE Trans. Med. Imaging **39**(7), 2531–2540 (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Shahzad Ali is currently pursuing a Ph.D. degree in computer science and engineering at the School of Computer Science and Engineering, Kyungpook National University (KNU), Daegu, South Korea. He received a B.S. degree in computer and information sciences from the Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad, Pakistan, in 2012 and an M.S. degree in Computer Science from the National University of Computer and Emerging Sciences (NUCES), Lahore, Pakistan, in 2016. His research interests include computer vision, machine learning, biomedical image analysis, and remote sensing.

Yu Rim Lee received a Ph.D. in internal medicine from Kyungpook National University (KNU) in 2018. Since 2020, she has been an Assistant Professor at the Department of Internal Medicine at Kyungpook National University Hospital (KNUH) in Chilgok, South Korea. She is the author of more than 50 publications in hepatology. Her research interests are viral hepatitis, liver cirrhosis, and hepatocellular carcinoma. Through meaningful collaborations with clinicians, engineers, and basic medical scientists, she has presented promising research results. She serves as an academic committee member of the Korean Association for the Study of the Liver and the Korean Liver Cancer Association.

Soo Young Park received a Ph.D. in internal medicine from Kyungpook National University in 2006. From 2007 to 2008, he completed a Fellowship at the Department of Gastroenterology and Hepatology, Kyungpook National University Hospital (KNUH). Since 2011, he has been with the Department of Gastroenterology and Hepatology, KNUH, where he is currently a Professor. From 2018 to 2019, he was a Visiting Scientist at the NAFLD Research Center, University of California, San Diego. He is the author of over 200 publications on hepatology and gastroenterology. He holds several awards for his research, including Best Oral Presentation awards at national and international conferences. His research interests focus on liver diseases, including hepatitis B and C, non-alcoholic fatty liver disease, hepatocellular carcinoma, and the clinical applications of artificial intelligence in healthcare. He serves on the Scientific Board of the Korean Association of Clinical Ultrasound and is a member of the Editorial Board of the Korean Association for the Study of the Liver.

Won Young Tak graduated from Kyungpook National University College of Medicine in 1998, earning his MD. He completed his residency in internal medicine at Kyungpook National University Hospital (KNUH), becoming a board-certified internal medicine specialist in 1994. Following his fellowship in gastroenterology, he has served as a professor in the Department of Gastroenterology and Hepatology at KNUH since 1998. He is a full member of the National Academy of Medicine of Korea and currently presides over the Convergence Liver Cancer Study Association. His research primarily focuses on liver diseases, including viral hepatitis, cirrhosis, and hepatocellular carcinoma. He has significantly contributed to global clinical trials in these areas, publishing over 240 research papers and over 150 in high-impact SCI journals. His contributions have been recognized with the Academic Paper Award from the Korean Association for the Study of the Liver. Since 2019, he has also delved into medical artificial intelligence, collaborating on international projects that leverage medical imaging and electronic health records to foster innovation.

Soon Ki Jung received a Ph.D. in computer science from KAIST in 1997. From 1997 to 1998, he was a Research Associate with the University of Maryland Institute for Advanced Computer Studies (UMIACS). Since 1998, he has been with the School of Computer Science and Engineering, Kyungpook National University (KNU), Daegu, South Korea, where he is currently a Professor. From 2001 to 2002, he was a Research Associate, and from 2008 to 2009, he was a Visiting Faculty with the IRIS Computer Vision Laboratory, University of Southern California. He is the author of over 200 articles on computer vision and graphics. He holds more than 20 patents deriving from his research. His research interests include improving the understanding and performance of intelligent vision systems and VR/AR systems, mainly through the application of 3D computer vision, computer graphics, visualization, and HCI. He serves as the Vice President of the Korean Computer Graphics Society, the Korean HCI Society, and the Korean Multimedia Society.