Shape-Guided Dual Consistency Semi-Supervised Learning Framework for 3-D Medical Image Segmentation

Tao Lei, *Senior Member, IEEE*, Hulin Liu, Yong Wan, Chenxia Li, Yong Xia, *Senior Member, IEEE*, and Asoke K. Nandi, *Life Fellow, IEEE*

***Abstract*—Popular semi-supervised 3-D medical image segmentation networks commonly suffer from two limitations: First, the geometry shape constraint of targets is frequently disregarded, leading to coarse segmentation results. Second, semi-supervision is only performed on the last layer of the decoder, resulting in the insufficient representation learning of 3-D convolution neural network. To address these issues, we propose a shape-guided dual consistency semi-supervised learning (SDC-SSL) framework for 3-D medical image segmentation. Indeed, the proposed framework has two dominating advantages. Initially, a geometry-aware shape constraint is presented and used to learn the shape representation, which converts the differences between two networks into an unsupervised loss and lets the framework learn the boundary distance information of targets in unlabeled challenging regions. Additionally, a deep-supervised knowledge transfer strategy is developed and employed by the proposed framework, which can upgrade the generalization ability of our framework without increasing any extra parameters and computation costs in the inference phase. Experimental results demonstrate that the proposed framework outperforms state-of-the-art methods on two challenging 3-D medical image segmentation tasks due to effective geometry-aware shape constraint on unlabeled data and the strong ability of knowledge mining on labeled data. The code is available at: https://github.com/SUST-reynole/SDC-SSL.**

***Index Terms*—3-D deep supervision, 3-D medical image segmentation, multitask semi-supervised learning, shape constraint.**

Manuscript received 7 December 2022; revised 10 May 2023; accepted 7 June 2023. Date of publication 16 June 2023; date of current version 4 September 2023. This work was supported in part by the National Natural Science Foundation of China under Program 62271296; in part by the Natural Science Basic Research Program of Shaanxi under Program 2021JC-47; in part by the Key Research and Development Program of Shaanxi under Program 2022GY-436 and Program 2021ZDLGY08-07; in part by the Natural Science Basic Research Program of Shaanxi under Program 2022JQ-634 and Program 2022JQ-018; and in part by the Shaanxi Joint Laboratory of Artificial Intelligence under Grant 2020SS-03. *(Corresponding author: Yong Wan.)*

This work did not involve human subjects or animals in its research.

Tao Lei is with the Department of Geriatric Surgery, First Affiliated Hospital, Xi'an Jiaotong University, Xi'an 710061, China, and also with the Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China (e-mail: leitaoly@163.com).

Hulin Liu is with the Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China (e-mail: liuhulincn@163.com).

Yong Wan is with the Department of Geriatric Surgery, First Affiliated Hospital, Xi'an Jiaotong University, Xi'an 710061, China (e-mail: docwanyong@xjtu.edu.cn).

Chenxia Li is with the First Affiliated Hospital, Xi'an Jiaotong University, Xi'an 710061, China (e-mail: saphirli@sina.com).

Yong Xia is with the School of Computing, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: yxia@nwpu.edu.cn).

Asoke K. Nandi is with the Department of Electronic and Electrical Engineering, Brunel University London, UB8 3PH Uxbridge, U.K., and also with the School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: asoke.nandi@brunel.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TRPMS.2023.3286866.

Digital Object Identifier 10.1109/TRPMS.2023.3286866

I. Introduction

MEDICAL image segmentation task frequently plays a key role in computer-aided diagnosis and intelligent medicine due to the great improvement of the diagnostic efficiency and accuracy [1], [2]. It is essential to segment some crucial organs and lesions in medical images and extract further features from segmentation results to assist clinicians make accurate diagnosis and therapy planning [3].

In recent years, deep convolution neural networks have achieved remarkable performance in medical image segmentation [4], especially the U-shaped encoder–decoder networks, such as U-Net [5], 3-D U-Net [6], V-Net [7], nnU-Net [8], LV-Net [9], Defed-Net [10], QAU-Net [11], and SGU-Net [12]. Thanks to a large amount of labeled training data, the current deep learning techniques have yielded good results on almost all medical image segmentation tasks, but this does not mean that the issue has been faultlessly addressed [13]. In contrast to normal RGB images, medical images usually suffer from low-intensity contrast to adjacent tissues, intensity inhomogeneity and blurred boundaries [14]. Therefore, obtaining fine-grained annotations at the pixel level for medical images is inevitably time consuming and costly, since generating high-quality annotated data requires professional knowledge and clinical experience, particularly for 3-D medical images that need to be labeled manually slice by slice.

Current medical image segmentation methods are still dominated by supervised learning and require massive amounts of high-quality labeled data for model training [15]. To alleviate annotations scarcity, weakly supervised semantic segmentation approaches mainly engage nonpixel weak labels, including but not limited to bounding boxes or image-level category annotations to achieve pixel-level image segmentation, such as A2GNN [16], LBBA [17], and FickleNet [18]. In general, popular weakly supervised conditions can be classified into three types: 1) incomplete supervision; 2) inexact supervision; and 3) inaccurate supervision [19]. Semi-supervised learning is a widely accepted form of incomplete supervision conditioning,

which is halfway between supervised and unsupervised learning and utilizes only a small amount of labeled data and a large amount of unlabeled data for model training [20], [21].

Existing semi-supervised learning methods can be roughly divided into two groups: the self-training and the consistency learning. The self-training is a progressive technique that employs labeled data to train an initial model, which is then used to generate pseudo-labels for unlabeled data, such as deep adversarial networks (DANs) [22], cross pseudo supervision (CPS) [23], MC-Net [24], and MC-Net+ [25]. The consistency learning assumes that even if an image encounters data-level or model-level disturbances, the corresponding prediction from the same sample should not alter, such as uncertainty-aware mean teacher (UA-MT) [26], shape-aware semi-supervised network (SASSNet) [27], uncertainty rectified pyramid consistency (URPC) [28], and dual-task consistency (DTC) [29]. Considering that the collection of unlabeled data is more convenient in practical clinical application scenarios, it is necessary to develop semi-supervised segmentation technique to decrease the dependence on fine-labeled data.

Semi-supervised learning approaches commonly attempt to improve model performance by mining useful features or underlying information from unlabeled data [21]. Although these methods have achieved good results, they still face the following two challenges. First, existing semi-supervised learning approaches suffer from insufficient focus on high-uncertainty regions for contours, and lack explicit modeling on the boundary distance information, hence frequently leading to insufficient targets coverage or poor boundary prediction for 3-D medical images. Second, the training schemes only attach semi-supervision at the last layer of networks, which may spread the wrong guidance information layer by layer and ignore the importance of labeled data compared to unlabeled data. Specifically, under the circumstance that only a small amount of labeled data is available for training a 3-D convolution neural network with a huge number of parameters and a complex network structure [30], [31].

To mitigate the challenges described above, we propose a novel shape-guided dual consistency semi-supervised learning (SDC-SSL) framework for 3-D medical image segmentation. Our framework integrates two independent networks that have an identical encoder and slightly different decoders, and employs the difference of the two decoders to capture the uncertainty. Specifically, the segmentation network is used to generate pixelwise probability maps, and the regression network is applied to generate shape-aware representation. In addition, a geometry-aware autoencoder is adopted to model explicitly the boundary distance information of segmentation targets in a lower-dimensional manifold.

In summary, the major contributions of this article include the following.

1) We introduce a shape-guided constraint strategy based on dual consistency to explore the boundary distance information of targets in 3-D medical images. In addition, we present an explicit shape constraint strategy and try to employ a geometry-aware autoencoder regularization to explore the ability on shape representation of targets in a lower-dimensional manifold.
2) We rethink the deep supervision mechanism and propose a deep-supervised knowledge transfer strategy for 3-D medical image segmentation. Without introducing extra computation costs in the inference phase, auxiliary supervision branches are attached to the decoder, and then the intermediate layer feature representation is transmitted as knowledge within a single network.

Experimental results under two semi-supervised learning settings on widely used public volumetric benchmarks indicate that our framework achieves higher segmentation performance and provides better shape prediction by effectively mining unlabeled data and efficiently utilizing labeled data. The rest of this article is organized as follows. The related work is reviewed in Section II. The detailed description of the proposed framework is provided in Section III. The experimental results and discussion of key issues are reported in Section IV. The summary and conclusion are drawn in Section V.

II. Related Work

A. Semi-Supervised Medical Image Segmentation

In order to reduce the burden of annotations, researchers proposed lots of semi-supervised learning methods for medical image segmentation that utilize only a small amount of labeled data for model training. These approaches can be broadly categorized into the self-training and the consistency learning.

Semi-supervised learning based on the self-training obtains pseudo-labels for unlabeled data from the high-confidence predictions of models, and improves the quality of the generated pseudo-labels by learning both hand-labeled and pseudo-labeled samples [32]. For instance, Bai et al. [33] iteratively updated the pseudo-labels and network parameters, and then used conditional random fields (CRFs) to refine the pseudo-labels. In addition, most self-training methods typically employ adversarial training to improve the quality of pseudo-labels, encouraging the evaluation network to provide discriminative scores between the segmentation results, while encouraging the segmentation network to produce probability maps that are as similar as possible for labeled and unlabeled data, such as DAN [22], deep atlas prior (DAP) [34], Entropy Mini [35], and MTANS [36]. Through the iterative adversarial training process, the model is able to generate progressively accurate segmentation results for unlabeled data. Nevertheless, the pseudo-labels generated by the model perhaps contain a certain amount of noise, which may lead to the accumulation of misinformation during the training process [37]. To avoid this, some approaches incorporate the uncertainty or confidence estimation of pseudo-labels into the self-training, leading to better segmentation results, such as UA-MT [26], URPC [28], UATE [38], and double-uncertainty weighted method (DUWM) [39].

Semi-supervised learning based on consistency regularization learns features from both labeled and unlabeled data, where the regularly supervised loss is computed on labeled data and the unsupervised consistency loss is computed on

both labeled and unlabeled data. The mainstream consistency learning methods can be cursorily classified as mean teacher (MT) [40] and its variants, multitask learning [41], mutual consistency [42], and deep co-training [43]. The most popular methods of these techniques are MT [40] and its variants, which apply different perturbations, such as Gaussian noise, random rotation, and scaling to input images, enforcing consistency between the student and teacher model. For example, Yu et al. [26] explored the uncertainty estimation and encouraged the segmentation results to be consistent for the same input under various perturbations. In addition, Wang et al. [44], [45] developed a tripled-uncertainty guided framework to encourage the auxiliary tasks in the student model to learn more reliable knowledge from the teacher model.

In practical applications, a lot of methods based on MT [40], such as TCSM [46], SCO-SSL [47], UG-MCL [48], local and global structure-aware entropy regularized MT (LG-ER-MT) [49], MTANS [36], and ASE-Net [50], demonstrate the efficiency and effectiveness of medical image segmentation tasks. These approaches proven that the introduction of unlabeled data can produce considerable improvement in segmentation performance, particularly when the existing available training dataset is small.

B. Multitask Semi-Supervised Learning

Most current medical image segmentation networks are based on encoder–decoder architectures, which are frequently used for 2-D images and 3-D volumetric data, respectively. Due to the limitations of training data, Myronenko [51] appended a variational autoencoder branch to the encoder–decoder network to reconstruct the raw medical images. Meanwhile, the variational autoencoder can also be used as an auxiliary task for semi-supervised framework to learn the parameters of a generative model from unlabeled images [52].

Different from the consistency learning utilizing perturbed model or data, multitask learning incorporates inherent predictive perturbations from various tasks. The existing multitask learning techniques could be broadly classified into two types. The first generally contains a common backbone and multiple output branches for different tasks. For instance, SASSNet [27] is a multitask adversarial network that jointly predicts segmentation probability maps and signed distance maps of target surfaces and imposes a geometric shape-aware constraint on segmentation results, which still belongs to data-level regularization. DTC [29], however, is a representative dual-task network that is based on task-level regularization between the global level-set representation of targets and directly predicted pixelwise segmentation probability maps for both labeled and unlabeled images. Furthermore, SMTL [53] is a shape-aware multitask learning network which involves three tasks: main segmentation task, signed distance regression task, and contour detection task.

The second type of multitask semi-supervised learning is based on multiple autonomous segmentation networks with nonshared parameters or features for different tasks. For example, DTML [54] is also a dual-task mutual learning framework with two student networks that simultaneously learn region-based shape constraints and boundary-based surface mismatches. In addition, UG-MCL [48] is an uncertainty-guided mutual consistency learning framework with student model and teacher model (MT) [40], which takes the estimated uncertainty of models as the basis to select comparatively certain predictions. Furthermore, Liu and Zhao [55] designed a dual-view network with shared encoder and twin decoders to decrease the uncertainty of prediction, in which an auxiliary task is applied to learn global geometric information. In a word, these multitask semi-supervised learning methods predict segmentation probability maps and signed distance maps with geometric information perception, prompting the network to focus more on learning geometric-aware shape constraints from unlabeled medical images [56].

Above all, multitask learning [41] is an effective strategy for enhancing semi-supervised learning, allowing different levels of information from various tasks to be supplemented with each other during the training process. In contrast to the aforementioned approaches, our multitask learning framework consists of two independent encoder–decoder networks which have the identical encoder and slightly different decoders, and a geometry-aware autoencoder to perform segmentation tasks from different perspectives.

C. Knowledge Transfer Strategy

The knowledge transfer strategy, also referred as knowledge distillation, uses a large teacher model to regularize the training procedure of a small student model [57], [58]. Different from the one-way transfer between teacher model and student model, mutual learning [42], and deep co-training [43] are multimodal collaborative strategies with the goal of leveraging the large amount of inexpensively unlabeled data to assist in training better models.

Mutual learning [42] improves performance in the process of collaborative training by employing the divergence and complementarity among disparate models. For instance, MC-Net [24] and MC-Net+ [25] enforce mutual consistency of hard regions and encourage the multiple sightly different decoders to generate consistent and low-entropy predictions under a cycled pseudo-label scheme. Moreover, UG-MCL [48] encourages uncertainty-guided mutual consistency to exploit unlabeled data by incorporating intratask consistency learning from up-to-date predictions for self-ensembling multitask consistency learning.

Deep co-training [43] is a multiview learning technique that trains multiple neural networks from different views and exploits adversarial samples to encourage view-differences. For example, the deep multiplanar co-training (DMPCT) [59] cooperatively trains multiple deep networks to mine valuable information from multiple planes for generating reliable pseudo-labels. Moreover, the CPS [23] simultaneously trains two deep segmentation networks perturbed with the same structure but different initializations, which imposes the consistency for same input image. Furthermore, Luo et al. [60] simplified the deep co-training from

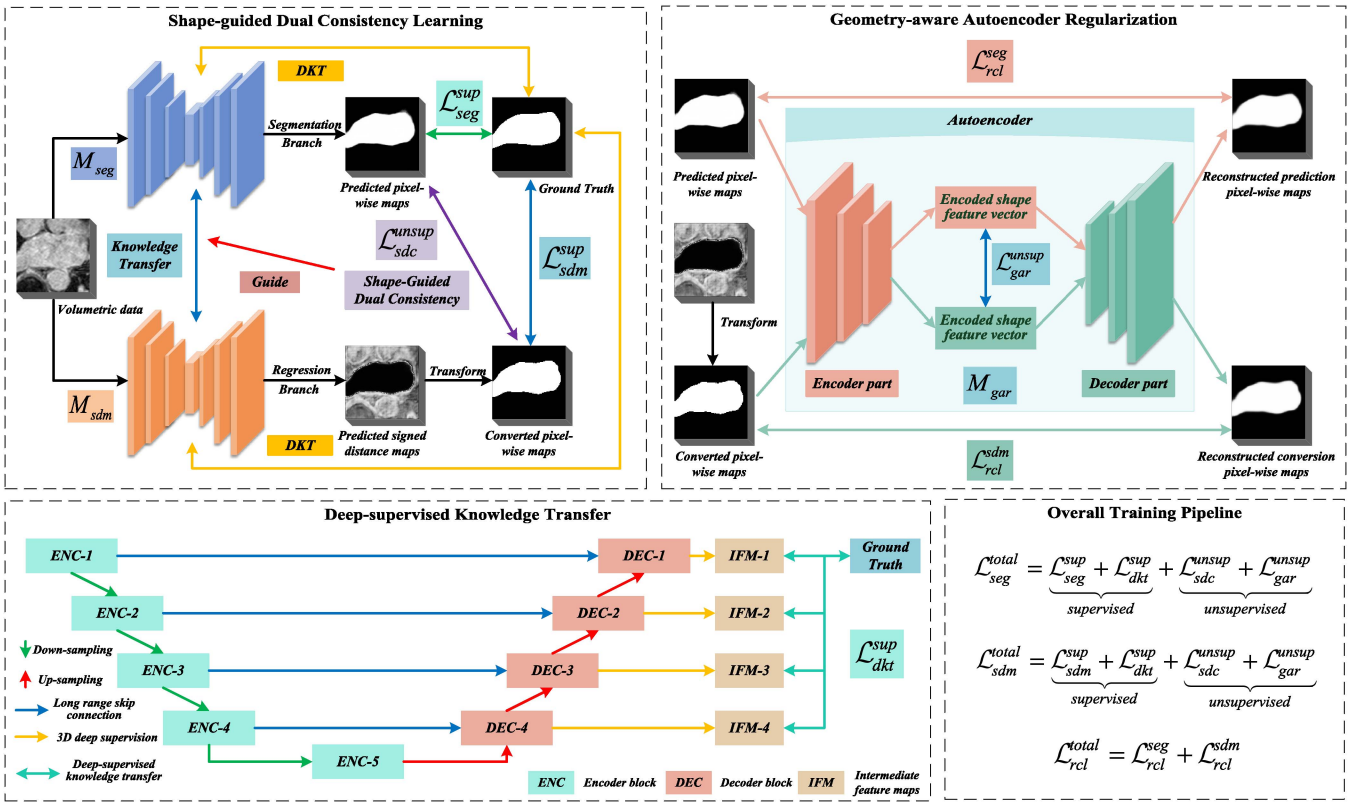

Fig. 1. Overview of the shape-guided dual consistency learning framework for semi-supervised 3-D medical image segmentation. Our framework integrates two independent networks and a geometry-aware autoencoder, the segmentation network M_{seg} is used to generate segmentation probability maps, the regression network M_{sdm} is applied to produce signed distance maps with shape perception, and the geometry-aware autoencoder M_{gar} is engaged to capture the saliency features of the input shape in a lower-dimensional manifold. In addition, auxiliary deep supervision is attached to each decoder stage to accelerate the knowledge transfer procedure within a single network, which is considered as an intermediate supervision to further utilize labeled data. Both networks are optimized by minimizing a combined loss, including the supervised loss $\mathcal{L}_{\text{seg}}^{\text{sup}}$ for M_{seg} or $\mathcal{L}_{\text{sdm}}^{\text{sup}}$ for M_{sdm} of labeled data, the auxiliary supervision loss $\mathcal{L}_{\text{dkt}}^{\text{sup}}$ of deep-supervised knowledge transfer, the shape-guided dual consistency loss $\mathcal{L}_{\text{sdc}}^{\text{unsup}}$, and the geometry-aware autoencoder regularization loss $\mathcal{L}_{\text{gar}}^{\text{unsup}}$ of labeled and unlabeled data. The trainable autocoder M_{gar} learns better shape feature encoding vectors by minimizing the reconstruction loss of both types of segmentation probability maps, which are denoted by $\mathcal{L}_{rcl}^{\text{seg}}$ and $\mathcal{L}_{rcl}^{\text{sdm}}$, respectively.

consistency regularization to cross teaching between CNN and Transformer, where the corresponding predictions from one of the networks was directly used for the pseudo-labels of another.

In summary, the aforementioned methods not only extend the training data by assigning pseudo-labels to unlabeled data, but also encourage the model learning more compact feature encoding in order to improve the performance of model. In addition, further studies [30], [31], [57], [61], [62] suggest that the intermediate feature maps can also be used to facilitate and accelerate the knowledge transfer process, thereby enhancing the model's discrimination capabilities.

III. Methodology

A. Overall Training Pipeline

In this section, we introduce the proposed shape-guided dual consistency framework for semi-supervised 3-D medical image segmentation. As shown in Fig. 1, our framework can be expressed as the integration of two independent segmentation networks and a geometry-aware autoencoder. The two networks share an identical encoder and two slightly different decoders to approximate cognitive uncertainty. The segmentation network denoted by M_{seg} uses the original 3-D transpose convolution to realize upsampling, while the regression network designated by M_{sdm} employs the trilinear interpolation and 3-D convolution block to expand feature maps. Additionally, the geometry-aware autoencoder designated by M_{gar} engages the continuous 3-D convolution to acquire encoded shape feature vectors, then utilizes the consecutive 3-D convolution and trilinear interpolation to obtain the reconstructed segmentation probability maps. They collaborate in the training process and indirectly explore more reliable information from both labeled and unlabeled data, but only need the main segmentation networks M_{seg} in the inference phase, which does not involve any additional memory or computing overhead. The overall training scheme is illustrated in Algorithm 1.

In the 3-D medical image segmentation based on semi-supervised learning, we define the training dataset as $D = D_l \cup D_u$, the labeled dataset as $D_l = \{x_i, y_i\}_{i=1}^{M}$, and the unlabeled dataset as $D_u = \{x_i\}_{i=1}^{N}$, where $M << N$. Furthermore, $x_i \in R^{H\times W\times D}$ and $y_i \in \{0, 1\}^{H\times W\times D}$ refers to the input image and the corresponding ground truth, respectively. For input volumetric data, the segmentation branch of M_{seg} outputs the segmentation probability maps $f_{\text{seg}}(x_i) \in [0, 1]^{H\times W\times D}$, while the regression branch of M_{sdm} generates the signed distance maps $f_{\text{sdm}}(x_i) \in [-1, 1]^{H\times W\times D}$.

Algorithm 1 Training Scheme of SDC-SSL

Input: $x_i \in D_l \cup D_u$, $y_i \in D_l$

Output: The model's weights θ_{seg} for segmentation network M_{seg}, θ_{sdm} for regression network M_{sdm}, and θ_{gar} for geometry-aware autoencoder M_{gar}

$f_{seg}(x_i)$ and $f_{sdm}(x_i)$ indicate the predictions of M_{seg} and M_{sdm} to generate segmentation probability maps and signed distance maps, respectively
while not stopping criterion **do**
  Sample batch: $\{(x_i, y_i) \in D_l, x_i \in D_u\}$
  Produce the segmentation probability maps $f_{seg}(x_i)$ and signed distance maps $f_{sdm}(x_i)$
  Calculate the principal supervised segmentation loss $\mathcal{L}_{seg}^{sup}$ and $\mathcal{L}_{sdm}^{sup}$ as Eq. (6) and Eq. (7)
  Calculate the unsupervised loss of shape-guided dual consistency $\mathcal{L}_{sdc}^{unsup}$ as Eq. (8)
  Calculate the unsupervised loss of geometry-aware autoencoder regularization $\mathcal{L}_{gar}^{unsup}$ as Eq. (9)
  Calculate the supervised loss of deep-supervised knowledge transfer $\mathcal{L}_{dkt}^{sup}$ as Eq. (12)
  Update the weights θ_{seg} with $\mathcal{L}_{seg}^{total} = \mathcal{L}_{seg}^{sup} + \mathcal{L}_{sdc}^{unsup} + \mathcal{L}_{gar}^{unsup} + \mathcal{L}_{dkt}^{sup}$
  Update the weights θ_{sdm} with $\mathcal{L}_{sdm}^{total} = \mathcal{L}_{sdm}^{sup} + \mathcal{L}_{sdc}^{unsup} + \mathcal{L}_{gar}^{unsup} + \mathcal{L}_{dkt}^{sup}$
  Calculate the unsupervised reconstruction loss $\mathcal{L}_{rcl}^{seg}$ and $\mathcal{L}_{rcl}^{sdm}$ as Eq. (10) and Eq. (11)
  Update the weights θ_{gar} with $\mathcal{L}_{rcl}^{total} = \mathcal{L}_{rcl}^{seg} + \mathcal{L}_{rcl}^{sdm}$
end while
return θ_{seg}, θ_{sdm}, and θ_{gar}

The total loss of segmentation network M_{seg} and regression network M_{sdm} are defined as follows:

$$\mathcal{L}_{\text{seg}}^{\text{total}} = \underbrace{\mathcal{L}_{\text{seg}}^{\text{sup}} + \mathcal{L}_{\text{dkt}}^{\text{sup}}}_{\text{supervised}} + \underbrace{\mathcal{L}_{\text{sdc}}^{\text{unsup}} + \mathcal{L}_{\text{gar}}^{\text{unsup}}}_{\text{unsupervised}}, \tag{1}$$

$$\mathcal{L}_{\text{sdm}}^{\text{total}} = \underbrace{\mathcal{L}_{\text{sdm}}^{\text{sup}} + \mathcal{L}_{\text{dkt}}^{\text{sup}}}_{\text{supervised}} + \underbrace{\mathcal{L}_{\text{sdc}}^{\text{unsup}} + \mathcal{L}_{\text{gar}}^{\text{unsup}}}_{\text{unsupervised}} \tag{2}$$

where $\mathcal{L}_{\text{seg}}^{\text{sup}}$ for M_{seg} and $\mathcal{L}_{\text{sdm}}^{\text{sup}}$ for M_{sdm} represent the supervised loss calculated on labeled data to learn a responsible representation from the corresponding ground truth, $\mathcal{L}_{\text{sdc}}^{\text{unsup}}$ indicates the unsupervised loss of shape-guided dual consistency between the generated segmentation probability maps and the converted pixelwise segmentation maps, $\mathcal{L}_{\text{gar}}^{\text{unsup}}$ denotes the unsupervised loss of geometry-aware autoencoder regularization between two types of encoded shape feature vectors in a lower-dimensional manifold, and $\mathcal{L}_{\text{dkt}}^{\text{sup}}$ means the supervised loss of deep-supervised knowledge transfer between the intermediate feature representation and the corresponding ground truth. The total loss of geometry-aware autoencoder M_{gar} is defined as follows:

$$\mathcal{L}_{\text{rcl}}^{\text{total}} = \mathcal{L}_{\text{rcl}}^{\text{seg}} + \mathcal{L}_{\text{rcl}}^{\text{sdm}} \tag{3}$$

where $\mathcal{L}_{\text{rcl}}^{\text{seg}}$ means the unsupervised reconstruction loss between predicted pixelwise maps in the segmentation network M_{seg} and reconstructed prediction pixelwise maps, and $\mathcal{L}_{\text{rcl}}^{\text{sdm}}$ refers to the unsupervised reconstruction loss between converted pixelwise maps transformed from the signed distance maps and reconstructed conversion pixelwise maps.

B. Shape-Guided Dual Consistency Learning

Due to the limitations of medical imaging quality, it is difficult to use a small amount of labeled data to segment targets accurately. Motivated by [27] and [56], a regression branch is attached to the existing encoder–decoder structure to generate signed distance maps and capture the geometric contour information of targets. Then we introduce the transformation from predicted pixelwise maps to signed distance maps [29], [48], [53], [54], [55] as follows:

$$\mathcal{T}(p) = \begin{cases} -\inf\limits_{q \in \partial S} ||p - q||_2, & p \in S_{\text{in}} \\ 0, & p \in \partial S \\ +\inf\limits_{q \in \partial S} ||p - q||_2, & p \in S_{ex} \end{cases} \tag{4}$$

where p and q are two different pixels/voxels in a predicted pixelwise maps, $||p - q||_2$ indicates the Euclidian distance between p and q, ∂S represents the contour of targets, S_{in} and S_{ex} denote the internal region and the external region.

In order to convert signed distance maps into segmentation probability maps, a smoothing approximation function is frequently adopted to achieve the inverse transformation [29], [48], [53], [54], [55], which is defined as:

$$\mathcal{T}^{-1}(z) = \frac{1}{1 + e^{-k \cdot z}} = \sigma(k \cdot z) \tag{5}$$

where z refers to the point corresponding to the pixels/voxels in the signed distance maps and k is a conversion coefficient as large as possible.

For the labeled dataset D_l, each network is principally optimized by the supervised loss to learn a reliable representation of segmentation targets [29], [54] as follows:

$$\mathcal{L}_{\text{seg}}^{\text{sup}} = \sum_{x_i, y_i \in D_l} \mathcal{L}_{\text{dice}}(f_{\text{seg}}(x_i), y_i) + \mathcal{L}_{\text{bce}}(f_{\text{seg}}(x_i), y_i), \tag{6}$$

$$\mathcal{L}_{\text{sdm}}^{\text{sup}} = \sum_{x_i, y_i \in D_l} \mathcal{L}_{\text{dice}}(\sigma(k \cdot f_{\text{sdm}}(x_i)), y_i) \tag{7}$$

where $\mathcal{L}_{\text{seg}}^{\text{sup}}$ represents the assorted loss between the predicted segmentation probability maps and the corresponding ground truth, $\mathcal{L}_{\text{sdm}}^{\text{sup}}$ indicates the compositional loss between the transformed pixelwise probability maps and the corresponding ground truth, $\mathcal{L}_{\text{dice}}$ denotes the commonly used dice loss, and $\mathcal{L}_{\text{bce}}$ means the binary cross-entropy loss.

For the entire training dataset $D_l \cup D_u$, we perform an unsupervised multitask dual consistency loss, which encourages consistent predictions on the same input images to utilize unlabeled data effectively. In the training process, our framework can strengthen shape constraints and collaboratively learn more discriminative features from unlabeled data through joint learning of the segmentation probability maps and the signed distance maps with geometric shape perception representations. Therefore, we focus on enhancing the shape-guided dual consistency between the generated segmentation probability maps and the converted pixelwise segmentation maps,

which can be defined as

$$\mathcal{L}_{\text{sdc}}^{\text{unsup}} = \sum_{x_i \in D_l \cup D_u} \lambda_{\text{sdc}} ||f_{\text{seg}}(x_i) - \sigma(k \cdot f_{\text{sdm}}(x_i))||^2 + \mathcal{BDL}(\partial f_{\text{seg}}(x_i), \partial \sigma(k \cdot f_{\text{sdm}}(x_i))) \quad (8)$$

where λ_{sdc} indicates the rising weighting coefficient. According to [26] and [63], we use a Gaussian ramp-up warming function $\lambda_{\text{sdc}}(t) = e^{-5(1-\frac{t}{t_{\max}})^2}$ to control the balance between the supervised loss and the unsupervised dual consistency loss, where t denotes the present step of iterations and $t_{\max}$ indicates the maximum training step. In addition, $\mathcal{BDL}(\cdot)$ means the boundary distance loss on the geometric space, which evaluates the change between two nearby boundaries [64]. The aforementioned boundary distance loss can alleviate the problem of blurred edges in the segmentation probability maps since it takes the form of a distance metric on the space of contours instead of regions.

C. Geometry-Aware Autoencoder Regularization

For medical image segmentation, the accuracy of contour is crucial, because the segmentation results are frequently used to reconstruct 3-D organs. However, due to the limitations of medical imaging quality, it is difficult to precisely segment targets. In fact, human organs generally have a fixed shape and position, and the incorporation of prior knowledge of target shape is essential to improve the medical image segmentation effect. Due to the fact that shape information is a strong prior knowledge shared among different data samples and is robust to the appearance changes of input data, Liu et al. [65] used shape information to learn feature spaces. The autoencoder is first trained with labeled data, shape features are then captured by mapping the segmentation results to a low-dimensional manifold space, and finally a regression model is learned in the 1-D feature space to predict the quality of segmentation results. Consequently, we attempt to incorporate higher-level shape constraints into the multitask semi-supervised learning framework to make segmentation results more consistent with anatomical prior knowledge [12].

Numerous studies [27], [29], [48], [53], [56] have demonstrated that the signed distance maps apparently capture geometric active contours and involve structural boundary distance information. As everyone knows, it is extremely complicated to measure shape similarity in high-dimensional space. The high-dimensional shape, however, usually lies in a lower-dimensional manifold where each shape would be mapped to an encoded shape feature vector in the subspace. In light of this, Oktay et al. [66] presented an anatomically constrained neural network, which uses an autoencoder to learn the nonlinear compact representation of underlying anatomical structures in low-dimensional spaces.

Motivated by [12], [51], [52], and [66], we present a novel geometry-aware autoencoder regularization by exploring shape representation and guiding the final segmentation results to be closer to labels annotated by physicians. Specifically, we employ a trainable autoencoder to capture the saliency features of the input shape and encode them into a lower-dimensional manifold as shown in Fig. 2. Since the well-designed autoencoder can commendably reconstruct the input shape, and the encoding in a lower-dimensional manifold can be well approximated as a representation of the shape features.

The proposed geometry-aware autoencoder is trained on the whole training dataset $D_l \cup D_u$ by the generated segmentation probability maps and the pixelwise segmentation maps converted from signed distance maps. The aforementioned regularization contains two major loss components, namely, the adversarial loss $\mathcal{L}_{\text{gar}}^{\text{unsup}}$ between two types of encoded shape feature vectors, and the associated reconstruction loss between two kinds of the pixelwise maps, i.e., $\mathcal{L}_{\text{rcl}}^{\text{seg}}$ and $\mathcal{L}_{\text{rcl}}^{\text{sdm}}$.

The first loss term is the adversarial unsupervised loss, which tries to distinguish the predicted shape from the real shape representation by maximizing its distance in a lower-dimensional manifold, which can be defined as

$$\mathcal{L}_{\text{gar}}^{\text{unsup}} = \sum_{x_i \in D_l \cup D_u} ||E(f_{\text{seg}}(x_i)) - E(\sigma(k \cdot f_{\text{sdm}}(x_i)))||^2 \quad (9)$$

where $E(\cdot)$ represents the encoder part of the autoencoder which yields the encoded shape feature vectors of the generated segmentation probability maps or the converted pixelwise segmentation maps in a lower-dimensional manifold.

The second loss term is used to reconstruct the input shape from the encoded shape feature vectors by minimizing the reconstruction loss and then allowing the autoencoder to learn the shape representation, which is defined as

$$\mathcal{L}_{\text{rcl}}^{\text{seg}} = \sum_{x_i \in D_l \cup D_u} \mathcal{L}_{\text{dice}}(f_{\text{seg}}(x_i), AE(f_{\text{seg}}(x_i))) \quad (10)$$

$$\mathcal{L}_{\text{rcl}}^{\text{sdm}} = \sum_{x_i \in D_l \cup D_u} \mathcal{L}_{\text{dice}}(\sigma(k \cdot f_{\text{sdm}}(x_i)) AE(\sigma(k \cdot f_{\text{sdm}}(x_i)))) \quad (11)$$

where $AE(\cdot)$ specifies the geometry-aware autoencoder that initially produces the encoded shape feature vectors in the encoding stage, and then generates the reconstructed pixel-level segmentation probability maps in the decoding stage. $\mathcal{L}_{\text{rcl}}^{\text{seg}}$ means the reconstruction loss between predicted pixelwise maps from M_{seg} and reconstructed prediction pixelwise maps, while $\mathcal{L}_{\text{rcl}}^{\text{sdm}}$ refers to the reconstruction loss between converted pixelwise maps from M_{sdm} and reconstructed conversion pixelwise maps.

In this way, we force the geometry-aware autoencoder to better encode two types of shape and capture their subtle differences, while stimulating the segmentation network to trick autoencoder failing to capture these differences. Therefore, in the training phase, the segmentation network intuitively plays a game-theoretical min-max optimization with the autoencoder to exploit geometry-aware shape information. The segmentation network attempts to predict pixelwise probability maps that are consistent with the anatomical prior knowledge to minimize the distance in a lower-dimensional manifold, while the autoencoder strives to learn better shape feature encoding to maximize these distance. In other words, the trainable autoencoder requires to learn the ability of identifying subtle differences between the predicted contours from M_{seg} and the converted contours from M_{sdm} by maximizing their preceding

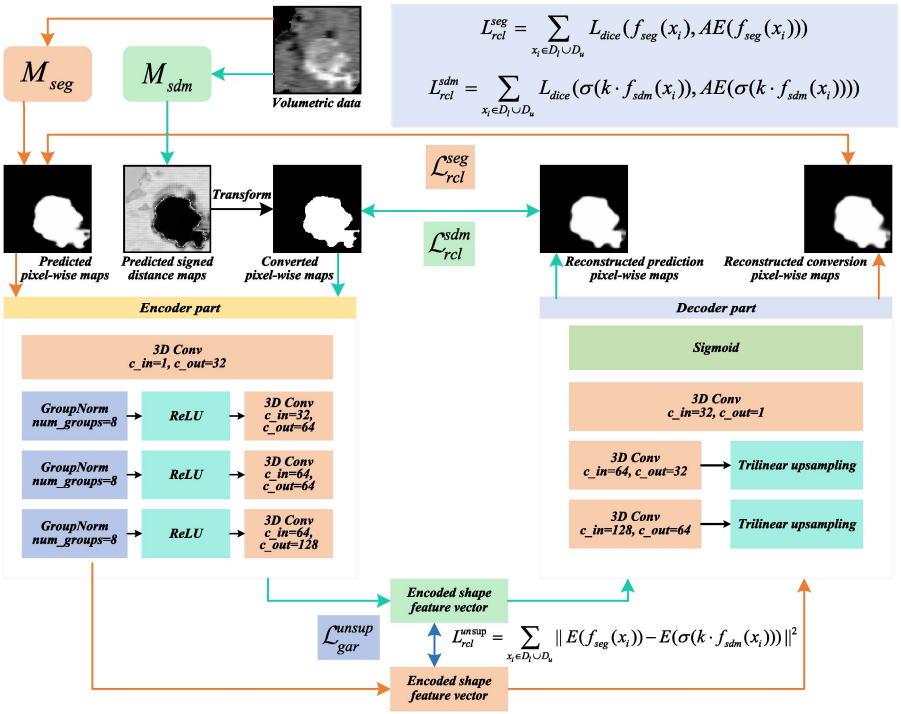

Fig. 2. Overview of the geometry-aware autoencoder regularization for shape-guided dual consistency learning. To begin with, the predicted segmentation probability maps of segmentation network M_{seg} and the pixelwise segmentation maps converted from the signed distance maps of regression network M_{sdm} are fed into the encoder part, and then mean square error loss $\mathcal{L}_{\text{gar}}^{\text{unsup}}$ between the two encoded shape feature vectors is calculated. Therewith, the obtained shape feature vectors are injected into the decoder part separately to acquire the reconstructed pixelwise segmentation probability maps. Subsequently, dice loss $\mathcal{L}_{\text{rcl}}^{\text{seg}}$ and $\mathcal{L}_{\text{rcl}}^{\text{sdm}}$ are assessed between the initial segmentation probability maps and the reconstructed pixelwise maps, which is principally engaged to train the autoencoder.

differences, while the segmentation network tries to trick autoencoder by minimizing these differences.

D. Deep-Supervised Knowledge Transfer

As medical volumetric images commonly suffer from issues, such as blur, noise, low contrast, and the complex anatomical environments, it is more difficult to extract discriminative features from medical images than other normal RGB images [30], [31]. In the existing training schemes for semi-supervised 3-D medical image segmentation, semi-supervision is regularly performed only at the last layer, leading to troubles, such as vanishing gradients and insufficient representation learning of targets [61]. This issue may be exacerbated by the fact that only a small amount of labeled data is available for model training in semi-supervised learning, which inevitably slows down the convergence of the model and reduces its discriminative capacity.

To alleviate the aforementioned shortcomings, we propose a deep-supervised knowledge transfer strategy to emphasize the importance of small amounts of labeled data motivated by [9], [30], and [31]. As shown in Fig. 3, the intermediate feature representation is transferred as knowledge within a single network by attaching an auxiliary supervision branch to the decoder [62]. Specifically, the auxiliary supervision branch is considered as a constraint condition of each upsampling stage in the decoder, which can inject extra gradients from the intermediate feature maps at multiple scales into networks. The bottom of Fig. 3 shows the detailed architecture of the 3-D deep supervision module, where a pointwise convolution is first used for the input feature maps, then a trilinear interpolation is applied for upsampling, and finally a softmax layer is employed to generate the pixelwise probability maps of segmentation results [9].

For the labeled dataset D_l, we estimate the difference between the intermediate feature representation and the corresponding ground truth for each branch denoted by $\mathcal{L}_{\text{dkt}}^{\text{sup}}$, which is defined as

$$\mathcal{L}_{\text{dkt}}^{\text{sup}} = \sum_{x_i, y_i \in D_l} \sum_{k=1}^{K} \mathcal{L}_{\text{dice}}(h^k(x_i), y_i) + \mathcal{L}_{\text{bce}}(h^k(x_i), y_i) \quad (12)$$

where h^k represents the kth 3-D deep supervision branch. Different from existing techniques, we append auxiliary supervision branches and incorporate the obtained intermediate feature maps into the optimization process to fully exploit the small amount of labeled data, thereby upgrading the training efficiency and speeding up the convergence of networks.

IV. Experiments

A. Datasets and Preprocessing

Two popular public datasets are engaged in our experiments, containing the LA dataset [67] and the BraTS dataset [68]. Following the generic semi-supervised learning settings, we adopt the same preprocessing and standard data augmentation techniques, including random cropping and random flipping.

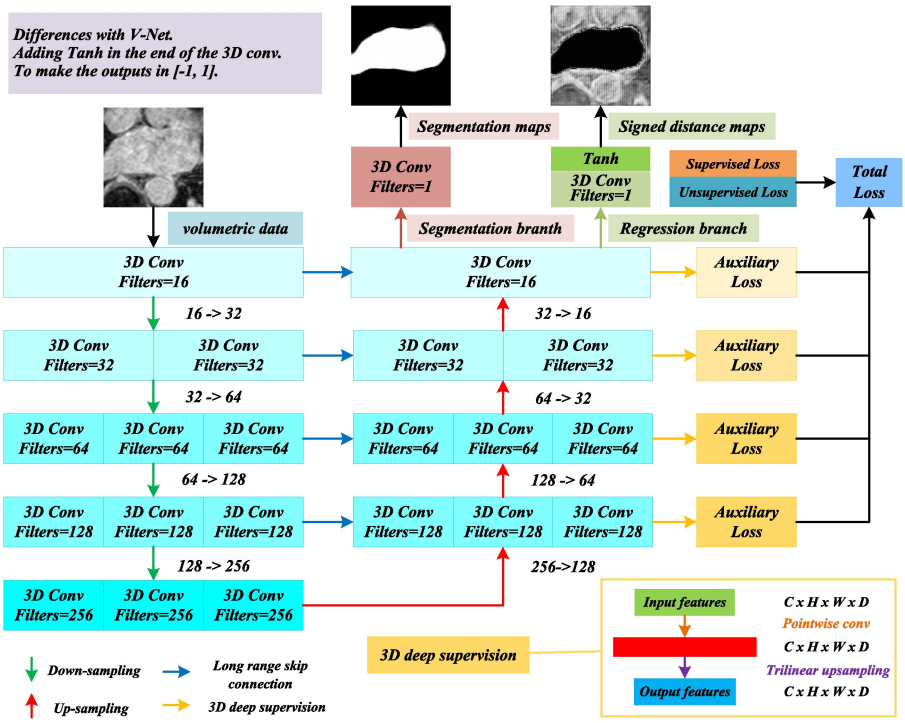

Fig. 3. Architecture of the backbone for segmentation network M_{seg} and regression network M_{sdm} in multitask semi-supervised learning. In contrast to V-Net [7], a regression branch is attached to the existing encoder–decoder structure to generate signed distance maps with the value domain of $[-1, 1]$ and dimension of $H \times W \times D$, which capture the geometric contour information of targets for 3-D medical image segmentation. The segmentation branch outputs pixelwise segmentation probability maps, while the regression branch outputs signed distance maps. In addition, we append an auxiliary supervision branch after each stage of decoder to transfer the intermediate feature maps as knowledge within a single deep neural network and display the detailed structure of the 3-D deep supervision module in the corner.

LA Dataset: The left atrium dataset is derived from the 2018 Atrial Segmentation Challenge[1] and consists of 100 3-D gadolinium-enhanced magnetic resonance imaging scans (GE-MRIs) comprising 3-D binary masks of the left atrial cavity, which are acquired from patients with atrial fibrillation (AF). Ethical approval was granted for all cardiac volume data, and the original isotropic resolution is $0.625 \times 0.625 \times 0.625$ mm^3. Following [26], [27], [29], and [69], we used 80 scans as the training dataset and 20 scans as the validation dataset. We engaged the constant 10%/8 scans or 20%/16 scans as labeled data and the remains of scans for unlabeled data.[2]

BraTS Dataset: Another dataset is the brain tumor segmentation dataset from Multimodal Brain Tumor Segmentation Challenge 2019,[3] which has always been focusing on the evaluation of state-of-the-art methods for the segmentation of brain tumors in multimodal MRI scans. It contains 335 scans, including T1, T1Gd, T2, and T2-FLAIR with the corresponding annotations, all from glioma patients, with the same resolution of $1 \times 1 \times 1$ mm^3. Following [48], we randomly selected 250 scans as the training dataset, 25 scans for validation and 60 scans for testing. We employed the identical 10%/25 scans or 20%/50 scans for labeled data and the rest of scans as unlabeled data.[4]

[1]https://atriaseg2018.cardiacatlas.org/data/
[2]https://github.com/yulequan/UA-MT/tree/master/data/
[3]https://www.med.upenn.edu/cbica/brats-2019/
[4]https://github.com/HiLab-git/SSL4MIS/tree/master/data/BraTS2019/

B. Implementing Details and Evaluation Metrics

Implementing Details: All algorithms were implemented with PyTorch framework and trained on a GeForce RTX 3090Ti GPU with 24-GB VRAM. It is worth mentioning that the random seed of Python, NumPy, PyTorch, CUDA, and DataLoader is fixed at 1337 to ensure the reproducibility of the proposed framework. We employed V-Net [7] as the backbone for all experiments to achieve a fair comparison, and implemented the multitask dual consistency networks by adding a regression branch at the end of the original V-Net [7]. The framework is trained for 6000 iterations by a SGD optimizer, with an initial learning rate of 0.01 decayed by 0.1 every 2500 iterations. The batch size is 4, consisting of two labeled images and two unlabeled images. We randomly cropped $112 \times 112 \times 80$ on the LA dataset [67] or $96 \times 96 \times 96$ on the BraTS dataset [68] subvolume as the input images. According to [26], [27], [29], and [69], we employed the standard on-the-fly data augmentation during the training stage to avoid overfitting and the value of k is set to 1500.

Evaluation Metrics: According to [70], four frequently used metrics are adopted to quantitatively evaluate the segmentation performance, i.e., Dice similarity coefficient (Dice), Jaccard index (Jaccard), average surface distance (ASD), and 95% Hausdorff distance (95HD), which can be defined as

$$Dice(V_{\text{pred}}, V_{gt}) = \frac{2|V_{\text{pred}} \cap V_{gt}|}{|V_{\text{pred}}| + |V_{gt}|} \tag{13}$$

TABLE I
COMPARISONS OF DIFFERENT METHODS ON THE LA DATASET WITH THE SAME 10%/8 OR 20%/16 SCANS AS THE LABELED TRAINING DATASET, THE BEST VALUES ARE HIGHLIGHTED IN BOLD

Method	Scans used		Metrics			
	Labeled	Unlabeled	Dice(%)	Jaccard(%)	ASD(voxel)	95HD(voxel)
V-Net [7]	8	0	79.99	68.12	5.48	21.11
V-Net [7]	16	0	86.03	76.06	3.51	14.26
V-Net [7]	80	0	91.14	83.82	1.52	5.75
DAP [34]	8	72	81.89	71.23	3.80	15.81
UA-MT [26]	8	72	84.25	73.48	3.36	13.48
SASSNet [27]	8	72	87.32	77.72	2.55	9.62
LG-ER-MT [50]	8	72	85.54	75.12	3.77	13.29
DUWM [39]	8	72	85.91	75.75	3.31	12.67
PDC-Net [70]	8	72	86.55	76.57	3.92	13.61
URPC [28]	8	72	85.01	74.43	3.96	15.37
DTC [29]	8	72	87.51	78.17	2.36	8.23
Ours	8	72	**88.31**	**79.25**	**1.94**	**7.56**
DAP [34]	16	64	87.89	78.72	2.74	9.29
UA-MT [26]	16	64	88.88	80.21	2.26	7.32
SASSNet [27]	16	64	89.54	81.24	2.20	8.24
LG-ER-MT [50]	16	64	89.62	81.31	2.06	7.16
DUWM [39]	16	64	89.65	81.35	2.03	7.04
PDC-Net [70]	16	64	89.76	81.57	2.95	10.31
URPC [28]	16	64	88.74	79.93	3.66	12.73
DTC [29]	16	64	89.42	80.98	2.10	7.32
Ours	16	64	**90.44**	**82.73**	**1.75**	**6.02**

TABLE II
COMPARISONS OF VARIOUS APPROACHES ON THE BRATS DATASET WITH THE IDENTICAL 10%/25 OR 20%/50 SCANS AS THE LABELED TRAINING DATASET, THE BEST VALUES ARE HIGHLIGHTED IN BOLD

Method	Scans used		Metrics			
	Labeled	Unlabeled	Dice(%)	Jaccard(%)	ASD(voxel)	95HD(voxel)
V-Net [7]	25	0	79.23	68.40	2.48	15.83
V-Net [7]	50	0	82.40	72.53	2.30	10.19
V-Net [7]	250	0	86.95	78.03	1.75	6.56
MT [40]	25	225	80.31	70.37	2.83	11.69
UA-MT [26]	25	225	80.85	70.32	2.57	13.61
Entropy Mini [35]	25	225	83.96	74.13	2.18	9.61
DAN [22]	25	255	83.43	73.71	2.18	9.30
ICT [72]	25	255	83.07	73.63	2.45	**9.13**
CPS [23]	25	255	82.74	72.86	2.22	9.14
URPC [28]	25	225	81.80	71.63	2.48	11.50
DTC [29]	25	255	83.03	73.41	**1.81**	10.22
Ours	25	255	**84.76**	**75.11**	1.95	11.29
MT [40]	50	200	83.72	74.24	2.11	9.61
UA-MT [26]	50	200	82.94	73.35	2.38	11.24
Entropy Mini [35]	50	200	84.45	74.91	2.15	7.97
DAN [22]	50	200	84.90	75.64	1.97	8.14
ICT [72]	50	200	84.94	75.51	**1.91**	**7.27**
CPS [23]	50	200	84.32	74.75	2.15	8.81
URPC [28]	50	200	82.80	72.72	2.72	12.48
DTC [29]	50	200	84.64	75.45	1.92	8.58
Ours	50	200	**85.45**	**76.23**	1.96	7.61

$$\text{Jaccard}\big(V_{\text{pred}}, V_{gt}\big) = \frac{|V_{\text{pred}} \cap V_{gt}|}{|V_{\text{pred}} \cup V_{gt}|} \tag{14}$$

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

$$HD(A,B) = \max[\sup_{a\in A}\inf_{b\in B} d(a,b), \sup_{b\in B}\inf_{a\in A} d(a,b)] \tag{16}$$

where V_{pred} and V_{gt} specify the set of pixels/voxels in the predicted pixelwise probability maps from the segmentation network M_{seg} and the corresponding ground truth. A and B indicate two sets of contour points, $d(a,b)$ denotes the Euclidean distance between the two points a and b, and $|\cdot|$ means the number of pixels/voxels in the V_{pred} or V_{gt}.

In addition, Dice and Jaccard are volumetric overlap measures, while ASD and 95HD are employed to evaluate the shape or contour accuracy of targets. The former focuses more on the interior of targets based on the region, while the latter concentrates on the geometric similarity of targets based on boundaries. It is worth noting that we calculated the metrics in the 3-D binary segmentation results.

C. Quantitative Comparison

In this section, we conducted extensive comparison experiments on the LA dataset [67] and the BraTS dataset [68] to demonstrate the superiority of the proposed framework under two different semi-supervised learning settings. In addition, we utilized 10%, 20%, and all scans of the training dataset, and presented the segmentation performance obtained by V-Net [7] in the fully supervised settings as shown in the top three rows of Tables I or II, respectively.

Comparison on the LA Dataset: In order to verify the superiority of the proposed framework, we conducted a comprehensive comparison on the LA dataset [67] with eight contemporary semi-supervised methods in two different settings, including DAP [34], UA-MT [26], SASSNet [27], LG-ER-MT [49], DUWM [39], parameter decoupling strategy network (PDC-Net) [69], URPC [28], and DTC [29]. Note that we employed the same V-Net [7] backbone in all these methods for a fair comparison.

Table I shows the quantitative performance comparison of various semi-supervised methods on the LA dataset [67]. Moreover, we specified the performance achieved by V-Net [7] as the upper and lower bound of the left atrium segmentation task under the fully supervised settings. By sufficiently mining a large amount of unlabeled data and efficiently utilizing a small amount of labeled data in the training process, the proposed framework obtains remarkable performance gains (from 86.03% to 90.44% of Dice, 76.06% to 82.73% of Jaccard, 3.51 to 1.75 of ASD, and 14.26 to 6.02 of 95HD). At the same time, our framework achieves the equivalent results compared to the upper bound performance (90.44% versus 91.14% of Dice, 82.73% versus 83.82% of Jaccard, 1.75 versus 1.52 of ASD, and 6.02 versus 5.75 of 95HD), which outperforms the mainstream semi-supervised methods.

In addition, Fig. 4 shows the visualization renderings of different semi-supervised methods on the LA dataset [67]. It is obvious that our framework produces more accurate and complete segmentation results.

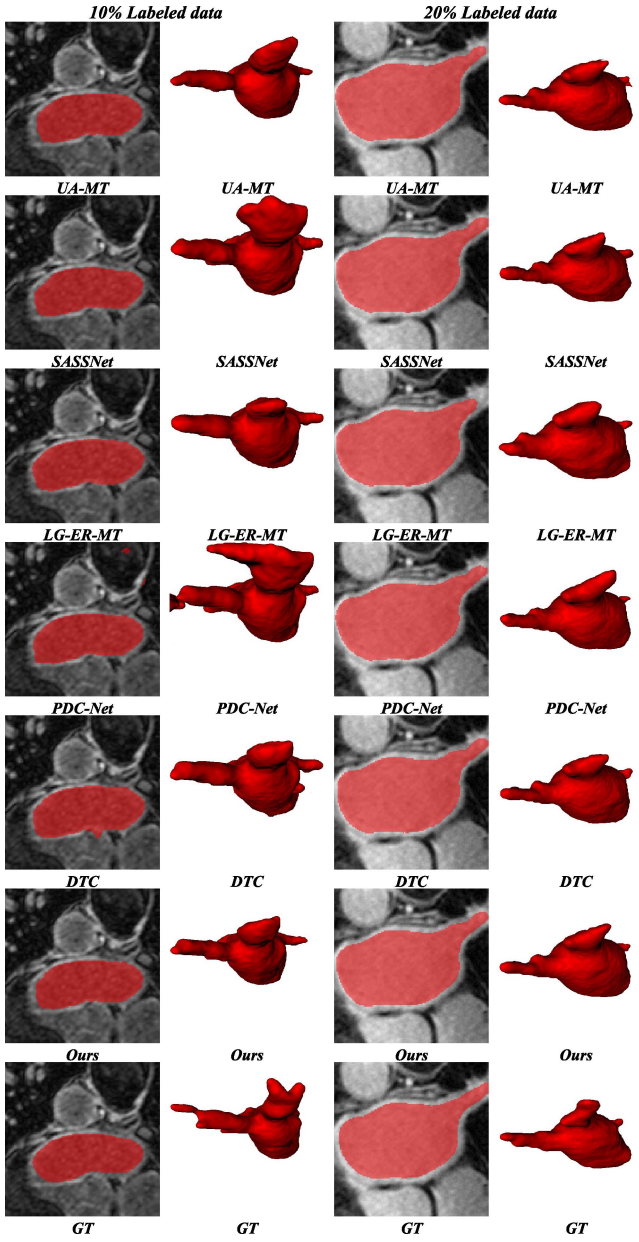

Fig. 4. Comparison of segmentation results using UA-MT [26], SASSNet [27], LG-ER-MT [49], PDC-Net [69], DTC [29], and the proposed framework on the LA dataset [67]. The comparisons using 10% or 20% labeled data are shown on the left or right two columns, respectively.

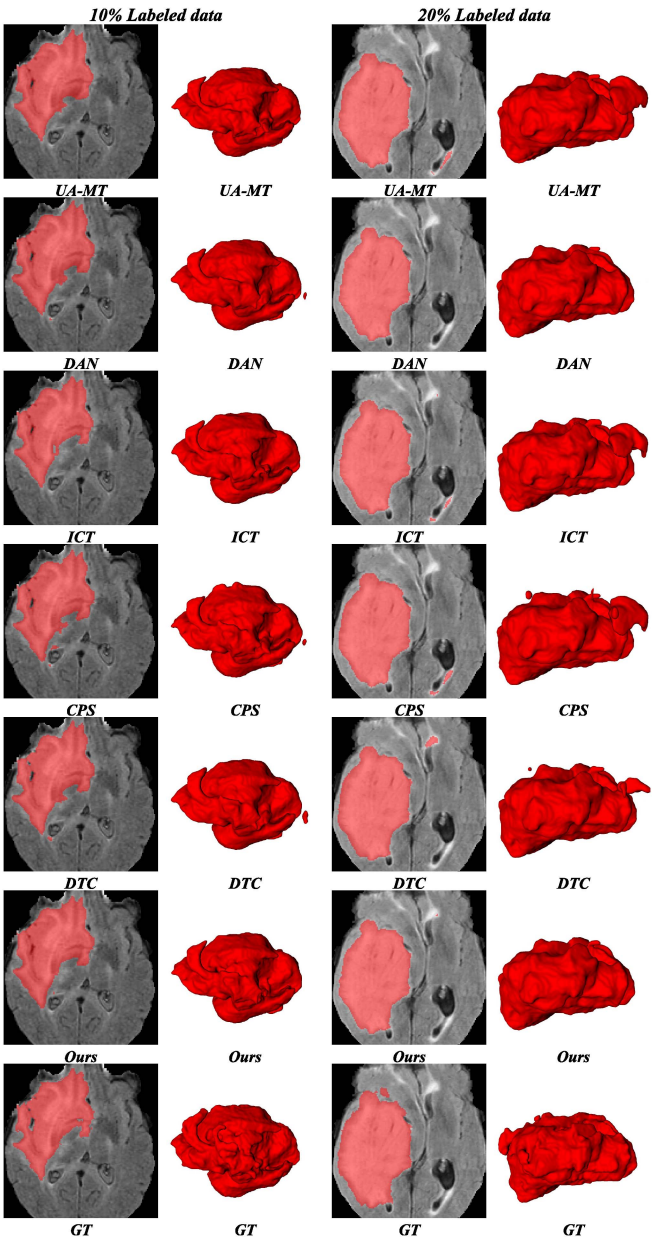

Fig. 5. Comparation of segmentation results using UA-MT [26], DAN [22], ICT [71], CPS [23], DTC [29], and the proposed framework on the BraTS dataset [68]. The comparisons using 10% or 20% labeled data are shown on the left or right two columns, respectively.

Comparison on the BraTS Dataset: To demonstrate the effectiveness of the proposed framework for 3-D medical image segmentation, eight state-of-the-art methods are adopted for comparison on the BraTS dataset [68], containing MT [40], UA-MT [26], adversarial entropy minimization (Entropy Mini) [35], DANs [22], interpolation consistency training (ICT) [71], CPS [23], URPC [28], and DTC [29]. The quantitative analysis of various classic and up-to-date semi-supervised methods on the BraTS dataset [68] is shown in Table II. Similar to the experimental results on the LA dataset [67], our framework achieves considerable performance gains on both region-based and boundary-based metrics. The visual presentations of experimental results on the BraTS dataset [68] is illustrated in Fig. 5. It is clear that our framework produces better segmentation results than state-of-the-art methods in challenging regions and eliminates most isolated areas.

D. Ablation Study

As illustrated in Table III, we conducted ablation experiments using 16 labeled and 64 unlabeled scans on the LA dataset [67] to examine the effectiveness of each component. Compared with the fully supervised V-Net [7] (1) using only 16 labeled scans as the training dataset, the semi-supervised baseline (2) with consistency loss from [29] and [54] has better performance. It is demonstrated that the performance can be improved by mining meaningful underlying information from the 64 unlabeled scans when only a small amount of labeled data is available for training.

TABLE III
ABLATION STUDY OF THE PROPOSED FRAMEWORK USING 16 LABELED AND 64 UNLABELED SCANS ON THE LA DATASET, THE BEST VALUES ARE HIGHLIGHTED IN BOLD

Method	Losses				Metrics			
	$\mathcal{L}_{con}^{unsup}$	$\mathcal{L}_{sdc}^{unsup}$	$\mathcal{L}_{gar}^{unsup}$	$\mathcal{L}_{dkt}^{sup}$	Dice(%)	Jaccard(%)	ASD(voxel)	95HD(voxel)
1)V-Net+					86.03	76.06	3.51	14.26
2)V-Net+	✓				89.79	81.64	1.83	6.95
3)V-Net+	✓		✓		90.09	82.10	1.94	6.63
4)V-Net+	✓			✓	90.31	82.43	1.78	6.61
5)V-Net+	✓		✓	✓	90.37	82.55	1.85	6.16
6)V-Net+				✓	90.18	82.22	**1.73**	6.77
7)V-Net+		✓			90.22	82.31	1.85	6.47
8)V-Net+		✓	✓		90.11	82.16	1.88	6.71
9)V-Net+		✓		✓	90.39	82.57	1.79	6.21
10)V-Net+		✓	✓	✓	**90.44**	**82.73**	1.75	**6.02**

To begin with, we engaged an additional shape constraint to learn the geometric shape representation of targets inspired by the previous semi-supervised consistency learning [29], [54], which facilitates the framework to focus on the boundary distance information during training process and indirectly exploring more robust knowledge from unlabeled data. Experimental results (7)–(10) reveal that the employment of boundary distance information can significantly improve segmentation accuracy. Furthermore, experimental findings (3), (5), (8), (10) illustrate that the deep-supervised knowledge transfer strategy yields significant performance gains by adding auxiliary supervised branches which assist the network in learning more discriminative and significative features from labeled data. In addition, the experimental results (4)–(6), (9) demonstrate the effectiveness of the proposed geometry-aware autoencoder regularization. The aforesaid strategy emphasized the small branches or blurred edges by adequately mining the shape information from signed distance maps at a lower-dimensional manifold, which makes a meaningful and reliable guidance for challenging regions.

E. Discussion

The main contribution of this article focuses on the training of semi-supervised learning frameworks, where not only a large amount of unlabeled data should be mined sufficiently to improve performance, but also a small amount of labeled data should be efficiently exploited to learn more reliable and valuable features. On this basis, the proposed framework contains two knowledge transfer strategies at the same time. On the one hand, our framework consists of two separate segmentation networks and a geometry-aware autoencoder, which can implicitly explore helpful knowledge and also leads to the framework learning a more compact feature encoding. On the other hand, the suggested deep-supervised knowledge transfer strategy is a high-efficiency form of knowledge transfer in a single neural network and facilitates the framework to gradually capture generalized features.

It is extremely complicated to measure shape similarity in a high-dimensional space where shape information involving organs or lesions is commonly located. Nevertheless, high-dimensional shapes usually lie in a certain low-dimensional manifold, where each shape is mapped to a low-dimensional feature vector in a subspace, and then evaluate the difference between the network predictions and the manual labels in a supervised learning manner, forcing the segmentation network to follow the learned statistical shape distribution [66]. The autoencoder is able to represent geometry-aware shape information in a microscopic way and can regularize the predicted target contours by minimizing the difference between the predicted segmentation probability maps and the converted signed distance maps in a lower-dimensional manifold.

In actual clinical diagnosis, different experts usually have different understandings on the same set of CT or MRI scanning images, so clinical diagnosis usually requires at least two experts to make a decision. Inspired by [24], [25], and [54], we use two slightly different decoders to approximate the cognitive uncertainty in actual clinical diagnosis. This strategy utilizes adversarial samples to encourage differences and complementarities between different views, mining low-uncertainty hidden information from multiple perspectives to guide the training process of the student network. Obviously, this strategy of building dual decoders can improve the generalization ability of our segmentation network.

Compared with the previous methods, our framework focuses on exploring the geometric boundary information of targets from unlabeled data and learning discriminative features from labeled data. First, the significance of unlabeled data is emphasized by shape-guided dual consistency learning and geometry-aware autoencoder regularization, which fall into the category of multitask semi-supervised learning. Second, the importance of labeled data is addressed through deep-supervised knowledge transfer, belonging to the category of knowledge transfer strategy. It is worth mentioning that although our framework introduces some parameters and computations during the training process, the regression network, the autoencoder and all auxiliary deep-supervised branches are discarded during the inference phase, and only the segmentation network needs to be deployed, which is in line with the universal semi-supervised learning settings [20], [21].

In addition, the recommended shape-guided dual consistency learning framework can be easily combined with Transformer [72], [73], [74], which has the ability of global semantic information interaction, to compensate for the locality of convolution operations. It may be possible to enforce implicit consistency regularization by cross pseudo supervised learning based on CNN and Transformer, which exploits the difference in learning paradigms to capture both local semantic features and global interaction information.

V. CONCLUSION

This work presents a novel and practical SDC-SSL framework for 3-D medical image segmentation to reduce the human effort of delineating volumetric data. The framework integrates two independent networks and a geometry-aware autoencoder that collaborate with each other to explore more

reliable information from a large amount of unlabeled data, engaging an additional shape constraint to guide the learning of boundary distance information. In addition, we evolve a deep-supervised knowledge transfer strategy to leverage a small amount of labeled data, which upgrades the ability to capture compact features in challenging regions. Extensive experiments on two challenging public benchmarks demonstrate that the proposed framework can provide a general and effective solution to achieve high-quality 3-D medical image segmentation.

ACKNOWLEDGMENT

All authors declare that they have no known conflicts of interest in terms of competing financial interests or personal relationships that could have an influence or are relevant to the work reported in this article.

REFERENCES

[1] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: Achievements and challenges," *J. Digit. Imag.*, vol. 32, no. 4, pp. 582–596, 2019.

[2] J. Ma et al., "AbdomenCT-1K: Is abdominal organ segmentation a solved problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6695–6714, Oct. 2022.

[3] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1597–1605, Jul. 2018.

[4] P. H. Conze, G. A. Miranda, V. K. Singh, V. Jaouen, and D. Visvikis, "Current and emerging trends in medical image segmentation with deep learning," *IEEE Trans. Radiat. Plasma Med. Sci.*, early access, Apr. 10, 2023, doi: 10.1109/TRPMS.2023.3265863.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.

[6] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2016, pp. 424–432.

[7] F. Milletari, N. Navab, and S. A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. IEEE Int. Conf. 3D Vis. (3DV)*, 2016, pp. 565–571.

[8] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier Hein, "NNU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nat. Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[9] T. Lei, W. Zhou, Y. Zhang, R. Wang, H. Meng, and A. K. Nandi, "Lightweight V-Net for liver segmentation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020, pp. 1379–1383.

[10] T. Lei, R. Wang, Y. Zhang, Y. Wan, C. Liu, and A. K. Nandi, "DefED-net: Deformable encoder–decoder network for liver and liver tumor segmentation," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 6, no. 1, pp. 68–78, Jan. 2022.

[11] L. Hong, R. Wang, T. Lei, X. Du, and Y. Wan, "QAU-Net: Quartet attention U-Net for liver and liver-tumor segmentation," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2021, pp. 1–6.

[12] T. Lei, R. Sun, X. Du, H. Fu, C. Zhang, and A. K. Nandi, "SGU-Net: Shape-guided ultralight network for abdominal image segmentation," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 3, pp. 1431–1442, Mar. 2023.

[13] P. Bizopoulos and D. Koutsouris, "Deep learning in cardiology," *IEEE Rev. Biomed. Eng.*, vol. 12, pp. 168–193, 2018.

[14] H. H. Chang and D. J. Valentino, "An electrostatic deformable model for medical image segmentation," *Comput. Med. Imag. Graph.*, vol. 32, no. 1, pp. 22–35, 2008.

[15] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET Image Process.*, vol. 16, no. 5, pp. 1243–1267, 2022.

[16] B. Zhang, J. Xiao, J. Jiao, Y. Wei, and Y. Zhao, "Affinity attention graph neural network for weakly supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8082–8096, Nov. 2022.

[17] B. Dong, Z. Huang, Y. Guo, Q. Wang, Z. Niu, and W. Zuo, "Boosting weakly supervised object detection via learning bounding box adjusters," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 2876–2885.

[18] J. Lee, E. Kim, S. Lee, J. Lee, and S. Yoon, "FickleNet: Weakly and semi-supervised semantic image segmentation using stochastic inference," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 5267–5276.

[19] Z. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, 2018.

[20] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning (Chapelle, O. et al., eds.; 2006)," *IEEE Trans. Neural Netw.*, vol. 20, no. 3, p. 542, Mar. 2009.

[21] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2020.

[22] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2017, pp. 408–416.

[23] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 2613–2622.

[24] Y. Wu, M. Xu, Z. Ge, J. Cai, and L. Zhang, "Semi-supervised left atrium segmentation with mutual consistency training," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2021, pp. 297–306.

[25] Y. Wu et al., "Mutual consistency learning for semi-supervised medical image segmentation," *Med. Image Anal.*, vol. 81, Oct. 2022, Art. no. 102530.

[26] L. Yu, S. Wang, X. Li, C. Fu, and P. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2019, pp. 605–613.

[27] S. Li, C. Zhang, and X. He, "Shape-aware semi-supervised 3D semantic segmentation for medical images," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2020, pp. 552–561.

[28] X. Luo et al., "Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2021, pp. 318–329.

[29] X. Luo, J. Chen, T. Song, and G. Wang, "Semi-supervised medical image segmentation through dual-task consistency," in *Proc. Assoc. Adv. Artif. Intell. (AAAI)*, vol. 35, 2021, pp. 8801–8809.

[30] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P. Heng, "3D deeply supervised network for automatic liver segmentation from CT volumes," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2016, pp. 149–157.

[31] Q. Dou et al., "3D deeply supervised network for automated segmentation of volumetric medical images," *Med. Image Anal.*, vol. 41, pp. 40–54, Oct. 2017.

[32] D. Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 3, 2013, p. 896.

[33] W. Bai et al., "Semi-supervised learning for network-based cardiac MR image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2017, pp. 253–260.

[34] H. Zheng et al., "Semi-supervised segmentation of liver using adversarial learning with deep Atlas prior," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2019, pp. 148–156.

[35] T. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 2517–2526.

[36] G. Chen et al., "MTANS: Multi-scale mean teacher combined adversarial network with shape-aware embedding for semi-supervised brain lesion segmentation," *NeuroImage*, vol. 244, Dec. 2021, Art. no. 118568.

[37] S. Min, X. Chen, Z. Zha, F. Wu, and Y. Zhang, "A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels," in *Proc. Assoc. Adv. Artif. Intell. (AAAI)*, vol. 33, 2019, pp. 4578–4585.

[38] X. Cao, H. Chen, Y. Li, Y. Peng, S. Wang, and L. Cheng, "Uncertainty aware temporal-ensembling model for semi-supervised abus mass segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 431–443, Jan. 2021.

[39] Y. Wang et al., "Double-uncertainty weighted method for semi-supervised learning," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2020, pp. 542–551.
[40] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 1195–1204.
[41] S. Chen, G. Bortsova, A. G.-U. Juárez, G. V. Tulder, and M. D. Bruijne, "Multi-task attention-based semi-supervised learning for medical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2019, pp. 457–465.
[42] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4320–4328.
[43] S. Qiao, W. Shen, Z. Zhang, B. Wang, and A. Yuille, "Deep co-training for semi-supervised image recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 135–152.
[44] K. Wang et al., "Tripled-uncertainty guided mean teacher model for semi-supervised medical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2021, pp. 450–460.
[45] K. Wang, B. Zhan, C. Zu, X. Wu, J. Zhou, and L. Zhou, "Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning," *Med. Image Anal.*, vol. 79, Jul. 2022, Art. no. 102447.
[46] X. Li, L. Yu, H. Chen, C. Fu, L. Xing, and P.-A. Heng, "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 523–534, Feb. 2021.
[47] X. Xu, T. Sanford, B. Turkbey, S. Xu, B. J. Wood, and P. Yan, "Shadow-consistent semi-supervised learning for prostate ultrasound segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 6, pp. 1331–1345, Jun. 2022.
[48] Y. Zhang, R. Jiao, Q. Liao, D. Li, and J. Zhang, "Uncertainty-guided mutual consistency learning for semi-supervised medical image segmentation," *Artif. Intell. Med.*, vol. 138, Apr. 2023, Art. no. 102476.
[49] W. Hang et al., "Local and global structure-aware entropy regularized mean teacher model for 3D left atrium segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2020, pp. 562–571.
[50] T. Lei, D. Zhang, X. Du, X. Wang, Y. Wan, and A. K. Nandi, "Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network," *IEEE Trans. Med. Imag.*, vol. 42, no. 5, pp. 1265–1277, May 2023.
[51] A. Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," in *Proc. Int. MICCAI Brainlesion Workshop*, 2018, pp. 311–320.
[52] S. Sedai, D. Mahapatra, S. Hewavitharanage, S. Maetschke, and R. Garnavi, "Semi-supervised segmentation of optic cup in retinal fundus images using variational autoencoder," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2017, pp. 75–82.
[53] S. Liu, Y. Li, X. Li, and G. Cao, "Shape-aware multi-task learning for semi-supervised 3D medical image segmentation," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, 2021, pp. 1418–1423.
[54] Y. Zhang and J. Zhang, "Dual-task mutual learning for semi-supervised medical image segmentation," in *Proc. Pattern Recognit. Comput. Vis. (PRCV)*, 2021, pp. 548–559.
[55] Z. Liu and C. Zhao, "Semi-supervised medical image segmentation via geometry-aware consistency training," 2022, *arXiv:2202.06104*.
[56] Y. Xue et al., "Shape-aware organ segmentation by predicting signed distance maps," in *Proc. Assoc. Adv. Artif. Intell. (AAAI)*, vol. 34, 2020, pp. 12565–12572.
[57] D. Sun, A. Yao, A. Zhou, and H. Zhao, "Deeply-supervised knowledge synergy," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 6997–7006.
[58] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Comput. Sci.*, vol. 14, no. 7, pp. 38–39, 2015.
[59] Y. Zhou et al., "Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2019, pp. 121–140.
[60] X. Luo, M. Hu, T. Song, G. Wang, and S. Zhang, "Semi-supervised medical image segmentation via cross teaching between CNN and transformer," in *Proc. Int. Conf. Med. Imag. Deep Learn. (MIDL)*, 2022, pp. 820–833.
[61] C. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. Artif. Intell. Stat.*, 2015, pp. 562–570.
[62] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4133–4141.
[63] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," 2016, *arXiv:1610.02242*.
[64] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," in *Proc. Int. Conf. Med. Imag. Deep Learn. (MIDL)*, 2019, pp. 285–296.
[65] F. Liu, Y. Xia, D. Yang, A. L. Yuille, and D. Xu, "An alarm system for segmentation algorithm based on shape model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 10652–10661.
[66] O. Oktay et al., "Anatomically constrained neural networks (ACNNs): Application to cardiac image enhancement and segmentation," *IEEE Trans. Med. Imag.*, vol. 37, no. 2, pp. 384–395, Feb. 2018.
[67] Z. Xiong et al., "A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance Imag," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101832.
[68] S. S. Bakas. "Brats MICCAI brain tumor dataset." 2020. [Online]. Available: https://ieee-dataport.org/competitions/brats-miccai-brain-tumor-dataset
[69] X. Hao, S. Gao, L. Sheng, and J. Zhang, "Parameter decoupling strategy for semi-supervised 3D left atrium segmentation," in *Proc. Int. Conf. Mach. Vis. (ICMV)*, vol. 12084, 2022, pp. 118–124.
[70] L. Li, V. A. Zimmer, J. A. Schnabel, and X. Zhuang, "Medical image analysis on left atrial LGE MRI for atrial fibrillation studies: A review," *Med. Image Anal.*, vol. 77, Apr. 2022, Art. no. 102360.
[71] V. Verma et al., "Interpolation consistency training for semi-supervised learning," *Neural Netw.*, vol. 145, pp. 90–106, Jan. 2022.
[72] F. Shamshad et al., "Transformers in medical imaging: A survey," *Med. Image Anal.*, vol. 88, Aug. 2023, Art. no. 102802.
[73] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.
[74] H. Cao et al., "Swin-UNet: UNet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 205–218.