Transformation-Consistent Self-Ensembling Model for Semisupervised Medical Image Segmentation

Xiaomeng Li, Lequan Yu, *Member, IEEE*, Hao Chen, *Member, IEEE*, Chi-Wing Fu, *Member, IEEE*, Lei Xing, and Pheng-Ann Heng, *Senior Member, IEEE*

Abstract—A common shortfall of supervised deep learning for medical imaging is the lack of labeled data, which is often expensive and time consuming to collect. This article presents a new semisupervised method for medical image segmentation, where the network is optimized by a weighted combination of a common supervised loss only for the labeled inputs and a regularization loss for both the labeled and unlabeled data. To utilize the unlabeled data, our method encourages consistent predictions of the network-in-training for the same input under different perturbations. With the semisupervised segmentation tasks, we introduce a transformation-consistent strategy in the self-ensembling model to enhance the regularization effect for pixel-level predictions. To further improve the regularization effects, we extend the transformation in a more generalized form including scaling and optimize the consistency loss with a teacher model, which is an averaging of the student model weights. We extensively validated the proposed semisupervised method on three typical yet challenging medical image segmentation tasks: 1) skin lesion segmentation from dermoscopy images in the International Skin Imaging Collaboration (ISIC) 2017 data set; 2) optic disk (OD) segmentation from fundus images in the Retinal Fundus Glaucoma Challenge (REFUGE) data set; and 3) liver segmentation from volumetric CT scans in the Liver Tumor Segmentation Challenge (LiTS) data set. Compared with state-of-the-art, our method shows superior performance on the challenging 2-D/3-D medical images, demonstrating the effectiveness of our semisupervised method for medical image segmentation.

Xiaomeng Li and Lequan Yu are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, and also with the Department of Radiation Oncology, Stanford University, Stanford, CA 94305 USA (e-mail: xmli@cse.cuhk.edu.hk; lqyu@cse.cuhk.edu.hk).

Hao Chen and Chi-Wing Fu are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: hchen@cse.cuhk.edu.hk; cwfu@cse.cuhk.edu.hk).

Lei Xing is with the Department of Radiation Oncology, Stanford University, Stanford, CA 94305 USA (e-mail: lei@stanford.edu).

Pheng-Ann Heng is with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, and also with Guangdong-Hong Kong-Macao Joint Laboratory of Human–Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: pheng@cse.cuhk.edu.hk).

I. INTRODUCTION

SEGMENTING anatomical structural or abnormal regions from medical images, such as dermoscopy images, fundus images, and 3-D computed tomography (CT) scans, is of great significance for clinical practice, especially for disease diagnosis and treatment planning. Recently, deep learning techniques have made impressive progress on semantic image segmentation tasks and become a popular choice in both computer vision and medical imaging community [1], [2]. The success of deep neural networks usually relies on the massive labeled data set. However, it is hard and expensive to obtain labeled data, notably in the medical imaging domain where only experts can provide reliable annotations [3]. For example, there are thousands of dermoscopy image records in the clinical center, but melanoma delineation by experienced dermatologists is very scarce, see Fig. 1. Such cases can also be observed in the optic disk (OD) segmentation from the retinal fundus images, and especially in liver segmentation from CT scans, where delineating organs from volumetric images in a slice-by-slice manner is very time consuming and expensive.

The lack of the labeled data motivates the study of methods that can be trained with limited supervision, such as semisupervised learning [4]–[6], weakly supervised learning [7]–[9], and unsupervised domain adaptation [10]–[12]. In this article, we focus on the semisupervised segmentation approaches, considering that it is relatively easy to acquire a large amount of unlabeled medical image data.

Semisupervised learning aims to learn from a limited amount of labeled data and an arbitrary amount of unlabeled data, which is a fundamental, challenging problem, and has a high impact on real-world clinical applications. The semisupervised problem has been widely studied in medical image research community [13]–[17]. Recent progress in semisupervised learning for medical image segmentation has featured deep learning [5], [18]–[21]. Bai *et al.* [18] present a semisupervised deep learning model for cardiac MR image segmentation, where the segmented label maps from unlabeled data are incrementally added into the training set to refine the segmentation network. Other semisupervised
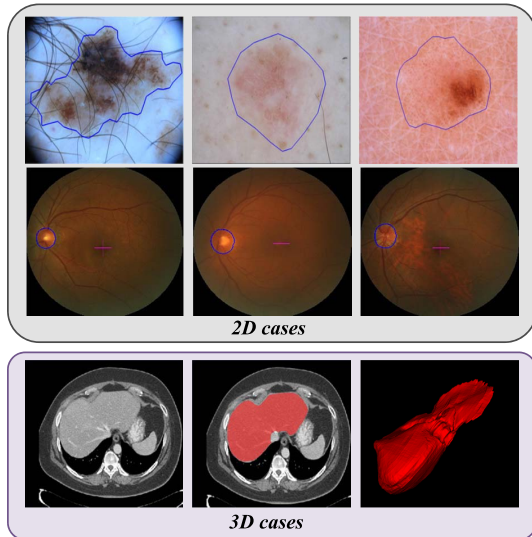
Fig. 1. Three common medical image segmentation tasks. The first, second, and third rows show the skin lesion in the dermoscopy image, the OD in retinal fundus images, and liver segmentations from CT scans, respectively. Blue color denotes the structure boundary and red color represents the liver.

learning methods are based on the recent techniques, such as variational autoencoder (VAE) [5] and generative adversarial network (GAN) [19]. We tackle the semisupervised segmentation problem from a different point of view. With the success of the self-ensembling model in the semisupervised classification problem [22], we further advance the method to medical image segmentation tasks, including 2-D cases and 3-D cases.

In this article, we present a new semisupervised learning method based on the self-ensembling strategy for medical image segmentation. The whole framework is trained with a weighted combination of supervised and unsupervised losses. The supervised loss is designed to utilize the labeled data for accurate predictions. To leverage the unlabeled data, our self-ensembling method encourages a consistent prediction of the network for the same input under different regularizations, e.g., randomized Gaussian noise, network dropout, and randomized data transformation. In particular, our method accounts for the challenging segmentation task, in which a pixel-level classification is required to be predicted. We observe that in the segmentation problem if one transforms (e.g., rotates) the input image, the expected prediction should be transformed in the same manner. When the inputs of convolutional neural networks (CNNs) are rotated, the corresponding network predictions will not rotate in the same way [23]. In this regard, we take advantage of this property by introducing a transformation (i.e., rotation and flipping) consistent scheme at the input and output space of our network. Specifically, we design the unsupervised loss by minimizing the differences between the network predictions under different transformations of the same input. To further improve the regularization, we extend the transformation consistency regularization with the scaling operation and optimize the network under a consistent scaling scheme. In addition, we adopt a teacher model to evaluate images under perturbations to construct better targets. We extensively evaluate our methods for semisupervised medical image segmentation

on three representative segmentation tasks, i.e., skin lesion segmentation from dermoscopy images, OD segmentation from retinal images, and liver segmentation from CT scans. In summary, our semisupervised method achieves significant improvements compared with the supervised baseline and also outperforms other semisupervised segmentation methods.

The main contributions of this article are as follows.

1) We present a novel and effective semisupervised method, namely, transformation-consistent self-ensembling model (TCSM_v2) for medical image segmentation. Our method is flexible and can be easily applied to both 2-D and 3-D CNNs.
2) We regularize unlabeled data with the transformation-consistent strategy and demonstrate effective semisupervised medical image segmentation.
3) Extensive experiments on three representative yet challenging medical image segmentation tasks, including 2-D and 3-D data sets, demonstrate the effectiveness of our semisupervised method over other methods.
4) Our method excels with the other state-of-the-art methods and establishes a new record in the International Skin Imaging Collaboration (ISIC) 2017 skin lesion segmentation data set with the semisupervised method.

This article extends our previous work TCSM [24] in three aspects. First, multiscale inference is an effective technique utilized in many image recognition tasks [25]–[27]. To enhance the regularization, we extend TCSM with more generalized transformation, such as random scaling. Through this, we utilize the unlabeled data to improve the regularization of the network. Second, our preliminary TCSM evaluates the inputs with perturbations on the same network. To avoid the misrecognition, we incorporate a teacher model to construct better targets, where the teacher model is an exponential moving average (EMA) of the student model. Third, we evaluate our method on three data sets, including the skin lesion data set, retinal fundus data set, and liver CT data set. Experiments on all three data sets show the effectiveness of our method over existing methods for semisupervised medical image segmentation.

II. RELATED WORK

A. Semisupervised Segmentation for Medical Images

Early semisupervised works segment medical images mainly using hand-crafted features [13], [14], [16], [17]. You *et al.* [13] combined radial projection and self-training learning to improve the segmentation of retinal vessel from fundus image. Portela *et al.* [14] presented a clustering-based Gaussian mixture model to automatically segment brain MR images. Later on, Gu *et al.* [16] constructed forest oriented superpixels for vessel segmentation. For skin lesion segmentation, Jaisakthi *et al.* [17] explored the *K*-means clustering and flood fill algorithm. These semisupervised methods are, however, based on hand-crafted features, which suffer from limited representation capacity.

Recent works for semisupervised segmentation are mainly based on deep learning. An iterative method is proposed by Bai *et al.* [18] for cardiac segmentation from MR images,

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

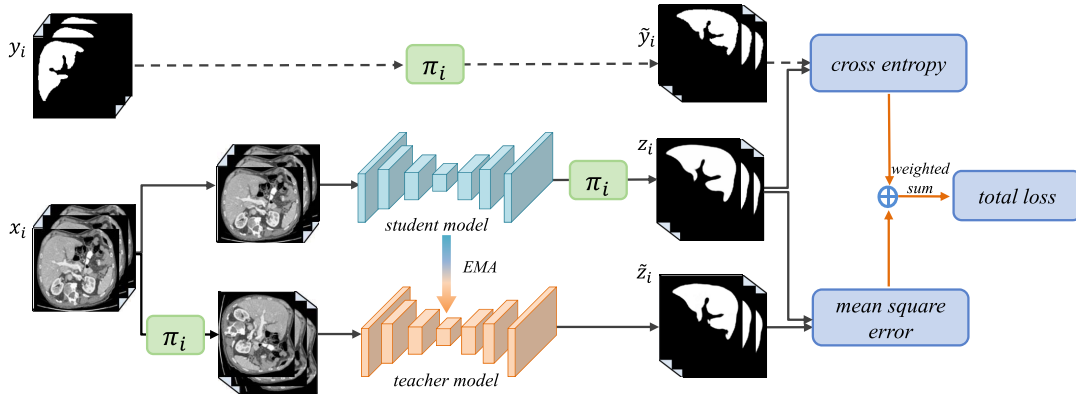LI *et al.*: TCSM FOR SEMISUPERVISED MEDICAL IMAGE SEGMENTATION

3

Fig. 2. TCSM_v2 for semisupervised medical image segmentation (we use liver CT scans as examples). The teacher and student models share the same architecture, and the weight of the teacher model is the EMA of the student model. The student model is trained by the total loss, which is a weighted combination of the cross-entropy loss on labeled data, and mean square error loss on both labeled and unlabeled data. The model encourages the teacher and student models to be transformed consistently by utilizing the unlabeled data. π_i refers to the transformation-consistent regularization, including rotation, flipping, and scaling operations.

where network parameters and segmentation masks for unlabeled data are alternatively updated. Generative model based semisupervised approaches are also popular in the medical image analysis community [5], [19], [28]. Sedai *et al.* [5] introduced a VAE for optic cup (OC) segmentation from retinal fundus images. They learned the feature embedding from unlabeled images using VAE and then combined the feature embedding with the segmentation autoencoder trained on the labeled images for pixelwise segmentation of the cup region. To involve unlabeled data in training, Nie *et al.* [19] presented an attention-based GAN approach to select trustworthy regions of the unlabeled data to train the segmentation network. Another GAN-based work [28] employed the cycle-consistency principle and performed experiments on cardiac MR image segmentation. More recently, Ganaye *et al.* [21] proposed a semisupervised method for brain structures segmentation by taking advantage of the invariant nature and semantic constraint of anatomical structures. Multiview co-training-based methods [4], [29] have also been explored on 3-D medical data. Differently, our method takes advantage of transformation consistency and self-ensembling model, which is simple yet effective for medical image segmentation tasks.

B. Transformation Equivariant Representation

Next, we review equivariance representations, to which the transformation equivariance is encoded in the network to explore the network equivariance property [23], [30]–[32]. Cohen and Welling [30] proposed a group equivariant neural network to improve the network generalization, where equivariance to 90°-rotations and dihedral flips are encoded by copying the transformed filters at different rotation-flip combinations. Concurrently, Dieleman *et al.* [31] designed four different equivariance to preserve feature map transformations by rotating the feature maps instead of the filters. Recently, Worrall *et al.* [23] restricted the filters to circular harmonics to achieve continuous 360°-rotations equivariance. However, these works aim to encode equivariance into the network

to improve its generalization capability, while our method aims to better utilize the unlabeled data in semisupervised learning.

C. Medical Image Segmentation

Early methods for medical image segmentation mainly focused on using thresholding [33], statistical shape models [34] and machine learning [35]–[37], while recent ones are mainly deep learning-based [38]–[40]. Deep learning methods showed promising results on skin lesion segmentation, OD segmentation, and liver segmentation [41]–[45]. Yu *et al.* [41] explored the network depth property and developed a deep residual network for automatic skin lesion segmentation by stacking residual blocks to increase the network's representative capability. Yuan *et al.* [46] trained a 19-layer deep CNN in an end-to-end manner for skin lesion segmentation. As for OD segmentation, Fu *et al.* [42] presented an M-Net for joint OC and OD segmentation. In addition, a disk-aware network [42] was designed for glaucoma screening by an ensemble of different feature streams of the network. For liver segmentation, Chlebus *al.* [37] presented a cascaded FCN combined with hand-crafted features. Li *et al.* [47] presented a 2-D–3-D hybrid architecture for liver and tumor segmentation from CT images. Although these methods achieved good results, they are based on fully supervised learning, requiring massive pixelwise annotations from experienced dermatologists or radiologists.

III. METHOD

Fig. 2 overviews our TCSM_v2 for semisupervised medical image segmentation. First, we randomly sample x_i raw data, including both the labeled and unlabeled cases from the training data set, followed by performing random transformations on these images. Teacher and student models are formulated in our framework, where the student model is trained by the loss function, and the teacher model is an average of consecutive student models. To train the student model, the transformed inputs are fed into the student model, and the softmax output

is compared with a one-hot label using classification cost (cross entropy in Fig. 2) and with the teacher output using consistency cost (mean square error in Fig. 2). After the weights of the student model have been updated with gradient descent, the teacher model weights are updated as an EMA of the student weights. Hence, the label information is passed to the unlabeled data by constraining the model outputs to be consistent with the unlabeled data.

A. Mean Teacher-Based Semisupervised Framework

To ease the description of our method, we first formulate the semisupervised segmentation task. In the semisupervised segmentation problem, the training set consists of N inputs in total, including M labeled inputs and $N - M$ unlabeled inputs. We denote the labeled set as $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^{M}$ and the unlabeled set as $\mathcal{U} = \{x_i\}_{i=M+1}^{N}$. For the 2-D images, $x_i \in \mathbb{R}^{H \times W \times 3}$ denotes the input image and $y_i \in \{0, 1\}^{H \times W}$ is the ground-truth segmentation mask. For the 3-D volumes, $x_i \in \mathbb{R}^{H \times W \times D}$ denotes the input volume and $y_i \in \{0, 1\}^{H \times W \times D}$ is the ground-truth segmentation volume. The general semisupervised segmentation learning tasks can be formulated to learn the network parameters θ by optimizing

$$\min_{\theta} \sum_{i=1}^{M} l(f(x_i; \theta), y_i) + \lambda R(\theta, \mathcal{L}, \mathcal{U}) \tag{1}$$

where l is the supervised loss function, R is the regularization (unsupervised) loss, and $f(\cdot)$ is the segmentation neural network and θ denotes the model weights. λ is a weighting factor that controls how strong the regularization is. The first term in the loss function is trained by the cross-entropy loss, aiming at evaluating the correctness of network output on labeled inputs only. The second term is optimized with the regularization loss, which utilizes both the labeled and unlabeled inputs.

The key point of this semisupervised learning is based on the *smoothness* assumption, i.e., data points close to each other in the image space are likely to be close in the label space [22], [48]. Specifically, these methods focus on improving the target quality using self-ensembling and exploring different perturbations, which include the input noise and the network dropout. The network with the regularization loss encourages the predictions to be consistent and is expected to give better predictions. The regularization loss R can be described as

$$R(\theta, \mathcal{L}, \mathcal{U}) = \sum_{i=1}^{N} \mathbb{E}_{\xi', \xi} \| f(x_i; \theta, \xi') - f(x_i; \theta, \xi) \|^2 \tag{2}$$

where ξ and ξ' denote to different regularization and perturbations of input data, respectively. In this article, we share the same spirit as these methods by designing different perturbations for the input data. Specifically, we design the regularization term as a consistency loss to encourage smooth predictions for the same data under different regularization and perturbations (e.g., Gaussian noise, network dropout, and randomized data transformation).

In the earlier, we evaluate the model twice to get two predictions under different perturbations. In this case, the
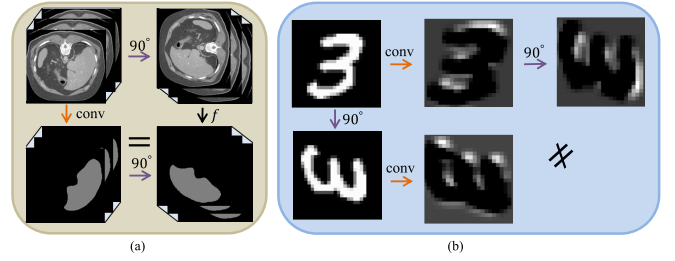

Fig. 3. (a) Segmentation is desired to be rotation equivariant. If the input image is rotated, the ground-truth mask should be rotated in the same manner. (b) Convolutions are not rotation equivariant in general. If the input image is rotated, the generated output is not the same with the original output that rotated in the same manner.

model assumes a dual role as a teacher and as a student. As a student, it learns as before, while, as a teacher, it generates targets, which are then used by itself as a student for learning. The model generates the targets by itself, and thus, it may be incorrect, especially when excessive weight is given to the generated targets. To construct better targets, we employ the mean teacher-based framework [49], where the teacher model $f_{\theta'}$ uses the EMA weights of the student model f_{θ}, i.e., $\theta'_t = \alpha \theta'_{t-1} + (1 - \alpha)\theta_t$. Specifically, the weight of teacher model θ'_t is updated through $\theta'_t = \alpha \theta'_{t-1} + (1 - \alpha)\theta_t$, where θ_t is the student model parameters and θ'_t is the teacher model parameters. α is a smoothing coefficient hyperparameter that affects how the teacher model relies on the current student model parameter. If α is large, the teacher model relies more on the previous teacher model in the last step; otherwise, the teacher model relies more on the current student model parameters. According to the empirical evidence in [49], setting $\alpha = 0$ makes the model as a variation of the π model, and the performance is the best when setting $\alpha = 0.999$. We follow this empirical experience and set α to 0.999 in our experiments. Then, the transformation-consistent regularization is performed on the input to the teacher model $f_{\theta'}$, and the consistency loss is added on the two predictions of the teacher and student model, respectively.

B. Transformation Consistent Self-Ensembling Model

Next, we introduce how we design the randomized data transformation regularization for segmentation, i.e., the TCSM_v2.

1) Motivation: In general self-ensembling semisupervised learning, most regularization and perturbations can be easily designed for the classification problem. However, in the medical image domain, accurate segmentation of important structures or lesions is a very challenging problem, and the perturbations for segmentation tasks are more worthy of exploring. One prominent difference between these two common tasks is that the classification problem is transformation *invariant* while the segmentation task is expected to be transformation *equivariant*. Specifically, for image classification, the CNN only recognizes the presence or absence of an object in the whole image. In other words, the classification result should remain the same, no matter what the data transformation (i.e., translation, rotation, and flipping) are applied to

the input image. While in the image segmentation task, if the input image is rotated, the segmentation mask is expected to have the same rotation with the original mask, although the corresponding pixelwise predictions are the same; see examples in Fig. 3(a). However, in general, convolutions are not transformation (i.e., flipping and rotation) equivariant,[1] meaning that if one rotates or flips the CNN input, then the feature maps do not necessarily rotate in a meaningful manner [23], as shown in Fig. 3(b). Therefore, the convolutional network consisting of a series of convolutions is also not transformation equivariant. Formally, every transformation $\pi \in \Pi$ of input $\mathbf{x}$ associates with a transformation $\psi \in \Psi$ of the outputs; that is $\psi[f(\mathbf{x})] = f(\pi[\mathbf{x}])$, but in general $\pi \neq \psi$.

2) Mechanism of TCSM: This phenomenon limits the unsupervised regularization effect of randomized data transformation for segmentation [22]. To enhance the regularization and more effectively utilize unlabeled data in our segmentation task, we introduce a transformation-consistent scheme in the unsupervised regularization term. Specifically, this transformation-consistent scheme is embedded in the framework by approximating ψ to π at the input and output space. Fig. 2 shows the detailed illustration of the framework, and Algorithm 1 shows the pseudocode. Under the transformation-consistent scheme and other different perturbations (e.g., Gaussian noise and network dropout), each input x_i is fed into the network for twice evaluation to acquire two outputs z_i and $\tilde{z}_i$. More specifically, the transformation-consistent scheme consists of triple π_i operations; see Fig. 2. For one training input x_i, in the first evaluation, the operation π_i is applied to the input image, while in the second evaluation, the operation π_i is applied to the prediction map. Random perturbations (e.g., Gaussian noise and network dropout) are applied in the network during the twice evaluations. By minimizing the difference between z_i and $\tilde{z}_i$ with a mean square error loss, the network is regularized to be transformation-consistent and thus increase the network generalization capacity. Notably, the regularization loss is evaluated on both the labeled and unlabeled inputs. To utilize the labeled data $x_i \in \mathcal{L}$, the same operation π_i is also performed on y_i and optimized by the standard cross-entropy loss. Finally, the network is trained by minimizing a weighted combination of unsupervised regularization loss and supervised cross-entropy loss.

3) Loss Function: It has cross-entropy loss on the labeled inputs and the regularization term on both the labeled and unlabeled inputs. The overall loss function is then defined as

$$\text{loss} = \mathcal{L} + \lambda(T)\mathcal{R} \tag{3}$$

where $\mathcal{L}$ and $\mathcal{R}$ are the supervised term and regularization term, respectively. The time-dependent warming up function $\lambda(T)$ is a weighting factor for supervised loss and regularization loss. This weighting function is a Gaussian ramp-up curve, i.e., $\lambda(T) = k*e^{(-5(1-T)^2)}$, where T denotes the training epoch and k scales the maximum value of the weighting function. In our experiments, we empirically set k as 1.0.

[1]Transformation in this article refers to flipping, scaling, and rotation.

Algorithm 1 TCSM_v2 Algorithm pseudocode.

Input: $x_i \in \mathcal{L} + \mathcal{U}$, $y_i \in \mathcal{L}$
1: $\lambda(T)$ = unsupervised weight function
2: $f_\theta(x)$ = student model with parameters θ
3: $f_{\theta'}(x)$ = teacher model with parameters θ'
4: $\pi_i(x)$ = transformation operations
5: α = smoothing coefficient hyperparameter.
6: **for** T in $[1, numepochs]$ **do**
7:    **for** each minibatch B **do**
8:       randomly update $\pi_i(x)$
9:       $z_{i \in B} \leftarrow \pi_i(f_\theta(x_{i \in B}))$
10:      $\tilde{z}_{i \in B} \leftarrow f_{\theta'}(\pi_i(x_{i \in B}))$
11:      $loss \leftarrow -\frac{1}{|B'|}\sum_{i \in (B')}\log z_i[\pi_i(y_i)] +$
12:      $\lambda(T)\frac{1}{|B|}\sum_{i \in B}\|z_i - \tilde{z}_i\|^2$
13:      update θ using optimizer
14:      update $\theta'_T \leftarrow \alpha\theta'_{T-1} + (1-\alpha)\theta_T$
15:    **end for**
16: **end for**
17: **return** θ';

We randomly sample $x_{i \in B}$ images from the training data and the supervised term $\mathcal{L}$ within one minibatch is defined as

$$\mathcal{L} = -\frac{1}{|B'|}\sum_{i \in B'}\log z_i[\pi_i(y_i)] \tag{4}$$

where $B' \in B$ denotes the labeled images within a minibatch, z_i and y_i are the network prediction and ground-truth segmentation label, respectively. At the beginning of the network training, $\lambda(T)$ is small, and training is mainly dominated by the supervised loss on the labeled data. In this way, the network is able to learn accurate information from the labeled data. As the training progresses, the network gets a reliable model and can generate output for the unlabeled data. The regularization term optimizes the prediction differences by calculating the differences on the predictions

$$\mathcal{R} = \frac{1}{|B|}\sum_{i \in B}\|z_i - \tilde{z}_i\|^2 \tag{5}$$

where z_i and $\tilde{z}_i$ denote the network predictions of the student model and teacher model, respectively, i.e., $z_i = \pi_i(f_\theta(x_{i \in B}))$ and $\tilde{z}_i = f_{\theta'}(\pi_i(x_{i \in B}))$. f_θ and $f_{\theta'}$ denote the student model and teacher model, respectively. The student model f_θ is updated by the gradient descent while the teacher model is updated by $\theta'_T = \alpha\theta'_{T-1} + (1-\alpha)\theta_T$, where $\alpha = 0.999$.

4) Implementation of π Operation: Multiscale ensembling is shown to be effective for image recognition [25]–[27]. To enlarge the regularization effect for semisupervised learning, we extend the TCSM to a more generalized form, including random scaling. The main goal is to keep the consistency of the teacher and student model after multiscale inference. Specifically, π_i operation includes not only rotation but also random scaling operation. For the student model, we give the input and generate the prediction z_i. For the teacher model, we randomly scale the input image and generate the prediction result, which is then rescaled to the original size of the input
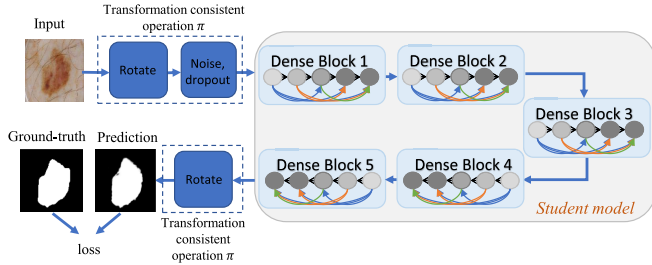
This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

6   IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

Fig. 4. Visualization of transformation-consistent operation π in the DenseUnet architecture. We omit the U-Net connection and the decoder part for simplification.

image, and we finally get $\tilde{z}_i$. Then, two predictions z_i and $\tilde{z}_i$ are minimized through a mean square error loss.

The transformation-consistent scheme includes random scaling, the horizontal flipping operation and four kinds of rotation operations to the input with angles of $\gamma \cdot 90°$, where $\gamma \in \{0, 1, 2, 3\}$. During each training pass, one rotation operation and one scaling operation within the scaling ratio of 0.8–1.2 is randomly chosen and applied to the input image, i.e., $\pi_i(x_i)$. To keep two terms in the loss function, we evenly and randomly select the labeled and the unlabeled samples in each minibatch. Note that we employed the same data augmentation in the training procedure of all the experiments for a fair comparison. However, our method is different from traditional data augmentation. Specifically, our method utilized the unlabeled data by minimizing the network output difference under the transformed inputs, while complying with the *smoothness* assumption.

C. Technical Details of TCSM_v2

For dermoscopy images and retinal fundus images, we employ the 2-D DenseUNet architecture [47] as both our teacher and student models. Compared with the standard DenseNet [50], we add the decoder part for the segmentation tasks. The decoder has four blocks, and each block consists of "upsampling, convolutional, batch normalization, and ReLU activation" layers. The UNet-like skip connection is added between the final convolution layer of each dense block in the encoder part and the convolution layer in the decoder part. The final prediction layer is a convolutional layer with a channel number of two. Before the final convolution layer, we add a dropout layer with a drop rate of 0.3. The network was trained with Adam algorithm [51] with a learning rate of 0.0001. All the experiments are trained for a total of 8000 iterations. We also visualize the network structure diagram to show the implementation of TCSM in Fig. 4.

To generalize our method to 3-D medical images, e.g., liver CT scans, we train TCSM_v2 with 3-D U-Net [38]. For training with 3-D U-Net, we follow the original setting with the following modifications. We modify the base filter parameters to 32 to accommodate this input size. The optimizer is stochastic gradient descent (SGD) with a learning rate of 0.01. The batch normalization layer is employed to facilitate the training process, and the loss function is modified to the standard weighted cross-entropy loss. All the experiments are trained for a total of 9000 iterations.

We implemented the model using PyTorch [52]. The experiments differ slightly from that in [24] due to the different

implementation platforms. We used the standard data augmentation techniques on-the-fly to avoid overfitting, including randomly flipping, rotating, and scaling with a random scale factor from 0.9 to 1.1. Note that all the experiments employed data augmentation for a fair comparison. In the inference phase, we remove the transformation operations in the network and do one single test with the original input for a fair comparison. After getting the probability map from the network, we first apply thresholding with 0.5 to generate the binary segmentation result, and then use morphology operation, i.e., filling holes, to obtain the final segmentation result.

IV. EXPERIMENTS

A. Data Sets

To evaluate our method, we conduct experiments on various modalities of medical images, including dermoscopy images, retinal fundus images, and liver CT scans.

1) Dermoscopy Image Data Set: The dermoscopy image data set in our experiments is the 2017 ISIC skin lesion segmentation challenge data set [53]. It includes a training set with 2000 annotated dermoscopic images, a validation set with 150 images, and a testing set with 600 images. The image size ranges from 540×722 to 4499×6748. To balance the segmentation performance and computational cost, we first resize all the images to 248×248 using bicubic interpolation.

2) Retinal Fundus Image Data Set: The fundus image data set is from the MICCAI 2018 Retinal Fundus Glaucoma Challenge (REFUGE)[2]. Manual pixelwise annotations of the OD were obtained by seven independent ophthalmologists from the Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China. Experiments were conducted on the released training data set, which contains 400 retinal images. The training data set is randomly split to training and test sets, and we resize all the images to 248×248 using bicubic interpolation.

3) Liver Segmentation Data Set: The liver segmentation data set are from the 2017 Liver Tumor Segmentation Challenge (LiTS)[3] [54], [55]. The LiTS data set contains 131 and 70 contrast-enhanced 3-D abdominal CT scans for training and testing, respectively. The data set was acquired by different scanners and protocols at six different clinical sites, with a largely varying in-plane resolution from 0.55 to 1.0 mm and slice spacing from 0.45 to 6.0 mm.

B. Evaluation Metrics

For dermoscopy image data set, we use Jaccard index (JA), Dice coefficient (DI), pixelwise accuracy (AC), sensitivity (SE), and specificity (SP) to measure the segmentation performance

$$AC = \frac{TP + TN}{TP+FP+TN+FN}$$

$$SE = \frac{TP}{TP+FN}, \quad SP = \frac{TN}{TN+FP}$$

$$JA = \frac{TP}{TP+FN+FP}, \quad DI = \frac{2 \cdot TP}{2 \cdot TP + FN + FP} \quad (6)$$

[2]https://refuge.grand-challenge.org/REFUGE2018/
[3]https://competitions.codalab.org/competitions/17094#participate-get_data

TABLE I

COMPARISON OF SUPERVISED LEARNING AND SEMISUPERVISED LEARNING (50 LABELED/1950 UNLABELED) ON THE VALIDATION SET IN THE DERMOSCOPY IMAGE DATA SET. "SUPERVISED+REG" DENOTES SUPERVISED WITH REGULARIZATION

Metric	Supervised	Supervised+regu	Ours
JA	71.17	72.28	**75.24**
DI	79.91	81.10	**83.44**
AC	91.95	93.52	**94.46**
SE	75.90	81.17	**83.07**
SP	97.04	97.02	**97.07**

TABLE II

ABLATION OF SEMISUPERVISED METHOD (50 LABELED/1950 UNLABELED) ON THE VALIDATION SET IN DERMOSCOPY IMAGE DATA SET. "R" DENOTES ROTATION TRANSFORMATION. "ND" DENOTES GAUSSIAN NOISE AND DROPOUT. "NDR" DENOTES GAUSSIAN NOISE, DROPOUT, AND ROTATION. "SHIFT (r)" DENOTES SHIFTING TRANSFORMATION WITH RATIO r. "NDRSCALE" DENOTES THE TRANSFORMATIONS INCLUDING NOISE, DROPOUT, ROTATION, AND RANDOM SCALING. (UNIT: %)

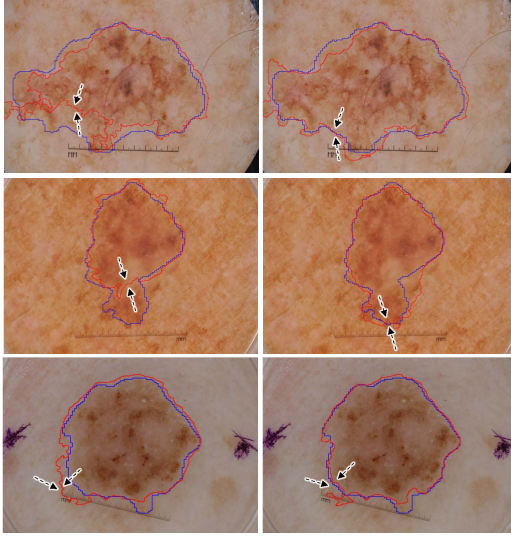| Setting	JA	DI	AC	SE	SP
Supervised	71.17	79.91	91.95	75.90	97.04
TCSM-R	73.76	82.05	94.87	84.14	**97.65**
TCSM-ND	74.03	82.33	94.30	83.84	96.77
TCSM-NDR	74.46	82.22	**95.06**	82.51	97.42
TCSM_v2-NDR	74.91	82.90	94.63	82.80	97.61
TCSM_v2-Shift(0.3)	71.26	75.11	79.5	78.32	85.40
TCSM_v2-Shift(0.2)	71.69	79.24	92.34	80.12	95.42
TCSM_v2-Shift(0.1)	72.74	80.93	93.35	81.32	96.62
TCSM_v2-Scale(0.3)	71.62	79.96	92.69	79.86	97.76
TCSM_v2-Scale(0.2)	71.87	79.78	92.75	79.98	96.04
TCSM_v2-Scale(0.1)	73.57	81.51	94.24	80.57	96.90
TCSM_v2-NDRScale(ours)	**75.24**	**83.44**	94.46	**83.07**	97.07

Fig. 5. Examples of the segmentation results of supervised learning (left) and our method (right) on the validation set in the dermoscopy image data set. The blue and red contours denote the ground truth and our segmentation result, respectively.

where TP, TN, FP, and FN refer to the number of true positives, true negatives, false positives, and false negatives, respectively. For the retinal fundus image data set, we use JA to measure the OD segmentation accuracy. For the liver CT data set, Dice per case score is employed to measure the accuracy of the liver segmentation result, according to the evaluation of the 2017 LiTS challenge [54].

C. Experiments on Dermoscopy Image Data Set

1) Quantitative and Visual Results With 50 Labeled Data: We report the performance of our method trained with only 50 labeled images and 1950 unlabeled images. Note that the labeled image is randomly selected from the whole data set. Table I shows the experiments with the supervised method, supervised with regularization, and our semisupervised method on the validation data set. We use the same network architecture (DenseUNet) in all these experiments for a fair comparison. The supervised experiment is optimized by the standard cross-entropy loss on the 50 labeled images. The supervised with regularization experiment is also trained with 50 labeled images, but differently, the total loss function is a weighted combination of the cross-entropy loss and the regularization loss, which is the same with our loss function. Our method is trained with 50 labeled and 1950 unlabeled

images in a semisupervised manner. From Table I, it is observed that our semisupervised method achieves higher performance than a supervised counterpart on all the evaluation metrics, with prominent improvements of 4.07% on JA and 3.47% on DI, respectively. It is worth mentioning that supervised with regularization experiment improves the supervised training due to the regularization loss on the labeled images; see "Supervised+regu" in Table I. The consistent improvements of "Supervised+regu" on all evaluation metrics demonstrate the regularization loss is also effective for the labeled images. Fig. 5 presents some segmentation results (red contour) of supervised method (left) and our method (right). Comparing with the segmentation contour achieved by the supervised method (left column), the semisupervised method fits more consistently with the ground-truth boundary. The observation shows the effectiveness of our semisupervised learning method, compared with the supervised method.

2) Effectiveness of TCSM and TCSM_v2: To show the effectiveness of our proposed transformation-consistent method, we conducted an ablation analysis of our method on the dermoscopy image data set, as the results are shown in Table II. The experiments were performed with randomly selected 50 labeled data and 1950 unlabeled data, and tested on the validation set. In the "Supervised" setting, we trained the network with only 50 labeled data. "TCSM-ND" refers to semisupervised learning with Gaussian noise and dropout regularization. "TCSM-R" refers to semisupervised learning with transformation-consistent regularization (only rotation), and "TCSM" refers to the experiment with all of these regularizations. As shown in Table II, both kinds of regularizations independently contribute to the performance gains of semisupervised learning. The resulting improvement with transformation-consistent regularization is very competitive, compared with the performance increment with Gaussian noise and dropout regularizations. These two regularizations are complementary, and therefore, when they are employed together, the performance can be further enhanced.

TABLE III

RESULTS OF OUR METHOD ON THE VALIDATION SET UNDER DIFFERENT
NUMBER OF LABELED/UNLABELED IMAGES (UNIT: %)

Label/Unlabel	Metric	Supervised	TCSM	TCSM_v2
50/1950	JA	71.17	74.46	**75.24**
	DI	79.91	82.22	**83.44**
	AC	91.95	**95.06**	94.46
	SE	75.90	82.51	**83.07**
	SP	97.04	**97.42**	97.07
100/1900	JA	73.92	74.75	**75.52**
	DI	82.37	83.02	**83.35**
	AC	93.87	**94.03**	93.28
	SE	83.95	**86.02**	83.78
	SP	**97.08**	96.89	94.65
300/1700	JA	76.66	77.19	**77.52**
	DI	84.32	**85.38**	85.20
	AC	94.25	95.11	**95.90**
	SE	85.82	86.02	**86.31**
	SP	95.77	95.78	**95.85**
2000/0	JA	78.80	79.15	**79.27**
	DI	86.67	87.01	**88.13**
	AC	95.12	95.03	**96.01**
	SE	88.75	89.20	**89.35**
	SP	97.02	96.91	**97.14**

Fig. 6. Results on the validation set of the dermoscopy image data set with different number of labeled/unlabeled data.

Moreover, we analyze the effects of shifting and scaling operations. Shift (r) denotes randomly shifting the image by $r'W$ or $r'H$, where $r' \in [1-r, 1+r]$, and W and H denote the image width and height. Scale (r) denotes randomly scaling the image to $(r'W, r'H)$, where $r' \in [1-r, 1+r]$. From the experiments in Table II, we can see that random scaling with ratio 0.1 could improve the semisupervised learning results, while other transformation settings have limited improvements. "TCSM_v2-NDR" denotes the mean teacher-based semisupervised learning with transformation-consistent strategy (only rotation). "TCSM_v2-NDRScale" refers to the mean teacher based semisupervised learning with our transformation-consistent strategy, including both rotation and scaling. From these two comparisons, we can see that the generalized form of transformation-consistent strategy improves the semisupervised learning. "TCSM" and "TCSM_v2-NDR" utilizes the same regularization. From these two comparisons, we can find that the weight-averaged consistency targets improve the semisupervised deep learning results. Our final model achieves 75.24% JA and 83.44% DI, surpassing the supervised baseline by 5.7% JA and 4.4% DI.

3) Results Under Different Number of Labeled Data: Table III shows the lesion segmentation results of our TCSM and TCSM_v2 (trained with labeled data and unlabeled data) and supervised method (trained only with labeled data) under a different number of labeled/unlabeled images. We draw the JA score of the results in Fig. 6. It is observed that the semisupervised methods consistently performs better than the supervised method in different labeled/unlabeled data settings, demonstrating that our method effectively utilizes the unlabeled data and brings performance gains. Note that in all semisupervised learning experiments, we train the network with 2000 images in total, including labeled images and unlabeled images. As 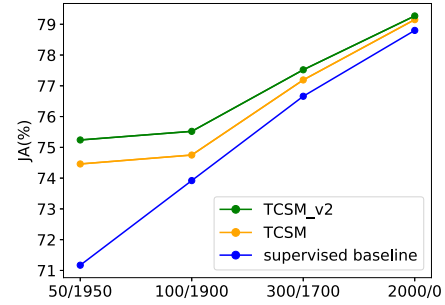expected, the performance of supervised training increases when more labeled training images are available; see the blue line in Fig. 6. At the same time, the segmentation performance of semisupervised learning can also increase with more labeled training images; see the orange line in Fig. 6. The performance gap between supervised training and semisupervised learning narrows as more labeled samples are available, which conforms to our expectations. When the amount of labeled data set is small, our method can gain a large improvement, since the regularization loss can effectively leverage more information from the unlabeled data. Comparatively, as the number of labeled data increases, the improvement becomes limited. This is because the labeled and unlabeled data are randomly selected from the same data set, and a large amount of labeled data may reach the upper bound performance of the data set.

From the comparison between TCSM and TCSM_v2, we can see that TCSM_v2 consistently improve TCSM under different label and unlabeled settings. From the comparison between the semisupervised method and supervised method trained with 2000 labeled images in Fig. 6, it can be observed that our method increases the JA performance when all labels are used. The improvement indicates that the unsupervised loss can also provide regularization to the labeled data. In other words, the consistency requirement in the regularization term can encourage the network to learn more robust features to improve the segmentation performance.

4) Comparison With Other Semisupervised Segmentation Methods: We compare our method with two latest semisupervised segmentation methods [18], [56] in the medical imaging community and an adversarial learning-based semisupervised method [57]. In addition, we extend the semisupervised classification model [49] to segmentation for comparison. Note that the method [19] for medical image segmentation adopts a similar idea with the adversarial learning-based method [57].

For a fair comparison, we re-implemented their methods with the same network backbone on this data set. All the experiments utilized the same data augmentation and training strategies. We conducted experiments with the setting of 50 labeled images and 1950 unlabeled images. Table IV shows the JA performance of different methods on the validation set. As shown in Table IV, our method achieves 4.07% JA improvement by utilizing unlabeled data. Compared with other methods, we achieve the greatest improvement over the supervised baseline. The comparison shows the effectiveness of our semisupervised segmentation method, compared with other semisupervised methods.

TABLE IV

JA PERFORMANCE OF DIFFERENT SEMISUPERVISED METHODS ON THE
VALIDATION DATA SET OF THE DERMOSCOPY IMAGE DATA SET.
"SUPERVISED" DENOTES TRAINING WITH 50 LABELED
DATA (UNIT: %)

Method	Backbone	Result	Improvement
Supervised		71.17	-
GAN [57]		73.31	2.14
DAN [56]	DenseUNet	73.82	2.65
Mean Teacher [49]		74.23	3.06
Self-training [18]		74.40	3.23
Ours		**75.24**	**4.07**

TABLE V

RESULTS ON THE TEST DATA SET IN THE ISIC 2017 DERMOSCOPY
LESION SEGMENTATION CHALLENGE (UNIT: %)

Team	Labels	JA	DI	AC	SE	SP
Our semi-supervised	300	**78.1**	**86.0**	**94.1**	**86.2**	96.8
Our baseline		76.8	84.8	93.6	83.6	96.7
Yuan and Lo [58]		76.5	84.9	93.4	82.5	97.5
Venkatesh et al. [59]		76.4	85.6	93.6	83.0	97.6
Berseth [60]		76.2	84.7	93.2	82.0	97.8
Bi et al. [61]		76.0	84.4	93.4	80.2	**98.5**
RECOD	2000	75.4	83.9	93.1	81.7	97.0
Jer		75.2	83.7	93.0	81.3	97.6
NedMos		74.9	83.9	93.0	81.0	98.1
INESC		73.5	82.4	92.2	81.3	96.8
Shenzhen U (Lee)		71.8	81.0	92.2	78.9	97.5

5) Comparison With Methods on the Challenge Leaderboard: We also compare our method with state-of-the-art methods submitted to the ISIC 2017 skin lesion segmentation challenge [53]. There are a total of 21 submissions, and the top results are listed in Table V. Note that the final rank is determined according to JA on the testing set. We trained two models: TCSM_v2 and baseline. TCSM_v2 was trained with 300 labeled images, and the left are utilized as the unlabeled images. The baseline model is trained with only 300 labeled data. Other methods use all labeled data as the training data. The supervised model is denoted as our baseline model. As shown in Table V, our semisupervised method achieved the best performance on the benchmark, outperforming the state-of-the-art method [58] with 1.6% improvement on JA (from 76.5% to 78.1%). The performance gains on DI and SE are consistent with that on JA, with 1.1% and 3.7% improvement, respectively. Our baseline model with 300 labeled data also outperforms other methods due to state-of-the-art network architecture. Based on this strong baseline, our semisupervised learning method further makes significant improvements, which demonstrates the effectiveness of the overall semisupervised learning method.

D. Experiments on Retinal Fundus Image Data Set

We report the performance of our method for OD segmentation from retinal fundus images. The 400 training images from the REFUGE challenge [62] were randomly separated into training and test data set with the ratio of 9:1. For a semisupervised training model, only a portion of labels (i.e., 10% and 20%) in the training set were used. We preprocessed all the input images by subtracting the mean RGB values of all the training data set. When training the supervised

TABLE VI

JA PERFORMANCE OF DIFFERENT METHODS ON THE FUNDUS IMAGE
DATA SET. "10%" AND "20%" DENOTE TRAINING WITH "10%" AND
"20%" LABELED DATA IN THE TRAINING SET, RESPECTIVELY.
"IMP" REFERS TO THE IMPROVEMENT OVER THE
SUPERVISED BASELINE

Method	Backbone	10%	Imp	20%	Imp
Supervised		93.61	-	94.61	-
Self-training [18]		94.20	0.59	95.01	0.40
GAN [57]	DenseUNet	94.32	0.71	94.93	0.32
DAN [56]		94.47	0.86	94.87	0.26
Mean Teacher [49]		94.76	1.15	95.02	0.41
Ours		**95.43**	**1.82**	**95.35**	**0.74**

TABLE VII

DICE PERFORMANCE OF DIFFERENT SEMISUPERVISED METHODS ON THE
LiTS DATA SET. "10%" AND "20%" DENOTE TRAINING WITH "10%"
AND "20%"LABELED DATA, RESPECTIVELY. "IMP" REFERS TO THE
IMPROVEMENT OVER THE SUPERVISED BASELINE

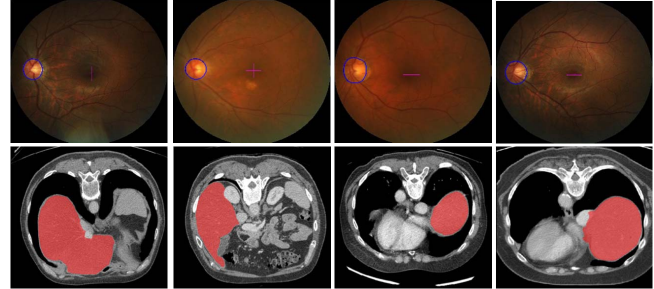| Method	Backbone	10%	Imp	20%	Imp
Supervised		87.97	-	88.55	-
Self-training [18]		89.64	1.67	89.65	1.10
DAN [56]	3D U-Net	92.90	4.93	92.29	3.74
GAN [57]		92.93	4.96	93.75	5.20
Mean Teacher [49]		93.28	5.31	93.37	4.82
Ours		**93.30**	**5.33**	**94.27**	**5.72**

Fig. 7. Examples of our semisupervised (20%) segmentation results for the fundus image and liver CT scans. Blue color denotes the segmented boundary of OD and red color represents the segmented liver.

model, the loss function was the traditional cross-entropy loss, and we used the SGD algorithm with learning rate 0.01 and momentum 0.9. To train the semisupervised model, we added the extra unsupervised regularization loss, and the learning rate was changed to 0.001.

We report JA of the supervised and semisupervised results under the setting of 10% labeled training images, and 20% labeled training images, respectively. As shown in Table VI, we also compare with the other semisupervised methods. It can be observed that our method achieves 1.82% improvement under the 10% labeled training set, which ranked top among all these methods. In addition, the improvement achieved by our method under the 20% training setting is also the highest. Fig. 7 shows some visual segmentation results of our semisupervised method. We can see that our method can better capture the boundary of the OD structure.

E. Experiments on LiTS Data Set

For this data set, we evaluate the performance of liver segmentation from CT volumes. Under our semisupervised setting, we randomly separated the original 131 training data

from the challenge into 121 training volumes and 10 testing volumes. For image preprocessing, we truncated the image intensity values of all scans to the range of $[-200, 250]$ HU to remove the irrelevant details. We run experiments with 3-D U-Net [38] to verify the effectiveness of our method. For the 3-D U-Net, the input size is randomly cropped to $112 \times 112 \times 32$ to leverage the information from the third dimension.

According to the evaluation of the 2017 LiTS challenge, we employed Dice per case score to evaluate the liver segmentation result, which refers to an average Dice score per volume. We report the performance of our method and the other three semisupervised methods under the settings of 10% labeled training images and 20% labeled training images, respectively, in Table VII. We can see that our approach achieves the highest performance improvement in both the 10% and 20% labeled training settings, with 5.33% and 5.72% improvements, respectively. In semisupervised learning, it is obvious that our method gains higher performance consistently than other methods [18], [49], [56], [57] in both 10% and 20% settings, respectively. We also visualize some liver segmentation results from CT scans in the second row of Fig. 7.

V. DISCUSSION

Supervised deep learning has been proven extremely effective for many problems in the medical image community. However, promising performance heavily relies on the number of annotations. Developing new methods with limited annotation will largely advance real-world clinical applications. In this article, we focus on developing semisupervised learning methods for medical image segmentation, which have great potential to reduce the annotation effort by taking advantage of numerous unlabeled data. The key insight of our semisupervised learning method is the transformation-consistent self-ensembling strategy. Extensive experiments on three representative and challenging data sets demonstrated the effectiveness of our method.

Medical image data have different formats, e.g., 2-D in-plane scans (e.g., dermoscopy images and fundus images) and 3-D volumetric data (e.g., MRI and CT). In this article, we use both 2-D and 3-D networks for segmentation. Our method is flexible and can be easily applied to both 2-D and 3-D networks. It is worth noting that recent works [4], [29] are specifically designed for 3-D volume data by considering three-view co-training, i.e., the coronal, sagittal, and axial views of the volume data. However, we aim for a more general approach that is applicable to 2-D and 3-D medical images. For 3-D semisupervised learning, it could be a promising direction to design specific methods by considering the 3-D natural property of the volumetric data.

Recent works on network equivariance [23], [30], [31] improve the generalization capacity of trained networks by exploring the equivariance property. E.g., Cohen and Welling [30] presented a group equivariant neural network. Our method also leverages the transformation consistency principle, but differently, we aim for semisupervised segmentation. Moreover, if we trained these works, i.e., harmonic network [23], in a semisupervised way to leverage unlabeled data, the transformation regularization will have no effect ideally, since the network outputs are the same when applying the transformation to the input images. Hence, the limited regularization would restrict the performance improvement from unlabeled data.

One limitation of our method is that we assume both labeled and unlabeled data come from the same distribution. However, in clinical applications, labeled and unlabeled data may have different distributions with a domain shift. Oliver *et al.* [63] demonstrated that the performance of semisupervised learning methods could degrade substantially when the unlabeled data set contains out-of-distribution samples. However, most of the current semisupervised approaches for medical image segmentation do not consider this issue. Therefore, in the future, we would explore domain adaptation [10] and investigate how to adapt it with a self-ensembling strategy.

In this article, the selection of transformation is based on the property of neural networks. The convolution layer is not rotation equivariant. To tackle the segmentation task, we need to train the network to be rotation equivariant. Moreover, the neural network is not scale equivariant due to padding, upsampling, and so on. The rotation and scaling transformations are the general transformations used in medical images. Thus, learning to minimize the output differences caused by these transformations will regularize the network to be transformation-consistent. Our method is flexible to extend to more general transformation cases, such as affine transformations. The transformation-consistent module consists of a transformation on the input image that will be fed to the teacher model, and the same transformation on the output space generated by the student model. It is flexible without additional training costs to be applied to the neural network. Moreover, recent automatic augmentation search works [64]–[66] explored the best transformations for a specific data set. It is an interesting future work to explore more useful transformations for our semisupervised segmentation framework through automatic data augmentation. In addition, the experiments reported are the averaged result over three trials, which also indicate the robustness of our method.

VI. CONCLUSION

This article presents a novel and effective transformation-consistent self-ensembling semisupervised method for medical image segmentation. The whole framework is trained with a teacher-student scheme and optimized by a weighted combination of supervised and unsupervised losses. To achieve this, we introduce a TCSM for the segmentation task, enhancing the regularization and can be easily applied on 2-D and 3-D networks. Comprehensive experimental analysis on three medical imaging data sets, i.e., skin lesion, retinal image, and liver CT data sets, demonstrated the effectiveness of our method. Our method is general and can be widely used in other semisupervised medical imaging problems.

REFERENCES

[1] D. Nie, L. Wang, Y. Gao, J. Lian, and D. Shen, "STRAINet: Spatially varying sTochastic residual AdversarIal networks for MRI pelvic organ segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1552–1564, May 2019.

[2] M. Mahmud, M. S. Kaiser, A. Hussain, and S. Vassanelli, "Applications of deep learning and reinforcement learning to biological data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2063–2079, Jun. 2018.

[3] M. D. Kohli, R. M. Summers, and J. R. Geis, "Medical image data and datasets in the era of machine learning—Whitepaper from the 2016 C-MIMI meeting dataset session," *J. Digit. Imag.*, vol. 30, no. 4, pp. 392–399, 2017.

[4] 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)*, Jan. 2019, pp. 121–140.

[5] 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. MICCAI*. Springer, 2017, pp. 75–82.

[6] V. Cheplygina, M. de Bruijne, and J. P. W. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Med. Image Anal.*, vol. 54, pp. 280–296, May 2019.

[7] Y. Hu *et al.*, "Weakly-supervised convolutional neural networks for multimodal image registration," *Med. Image Anal.*, vol. 49, pp. 1–13, Oct. 2018.

[8] W. M. Gondal, J. M. Kohler, R. Grzeszick, G. A. Fink, and M. Hirsch, "Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 2069–2073.

[9] X. Feng, J. Yang, A. F. Laine, and E. D. Angelini, "Discriminative localization in CNNs for weakly-supervised segmentation of pulmonary nodules," in *Proc. MICCAI*. Springer, 2017, pp. 568–576.

[10] K. Kamnitsas *et al.*, "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *Proc. Int. Conf. Inf. Process. Med. Imag.* Springer, 2017, pp. 597–609.

[11] N. Dong, M. Kampffmeyer, X. Liang, Z. Wang, W. Dai, and E. Xing, "Unsupervised domain adaptation for automatic estimation of cardiothoracic ratio," in *Proc. MICCAI*. Springer, 2018, pp. 544–552.

[12] F. Mahmood, R. Chen, and N. J. Durr, "Unsupervised reverse domain adaptation for synthetic medical images via adversarial training," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2572–2581, Dec. 2018.

[13] X. You, Q. Peng, Y. Yuan, Y.-M. Cheung, and J. Lei, "Segmentation of retinal blood vessels using the radial projection and semi-supervised approach," *Pattern Recognit.*, vol. 44, nos. 10–11, pp. 2314–2324, Oct. 2011.

[14] N. M. Portela, G. D. C. Cavalcanti, and T. I. Ren, "Semi-supervised clustering for MR brain image segmentation," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1492–1497, Mar. 2014.

[15] A. Masood, A. Al-Jumaily, and K. Anam, "Self-supervised learning model for skin cancer diagnosis," in *Proc. 7th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, Apr. 2015, pp. 1012–1015.

[16] L. Gu, Y. Zheng, R. Bise, I. Sato, N. Imanishi, and S. Aiso, "Semi-supervised learning for biomedical image segmentation via forest oriented super pixels (Voxels)," in *Proc. MICCAI*. Springer, 2017, pp. 702–710.

[17] S. M. Jaisakthi, A. Chandrabose, and P. Mirunalini, "Automatic skin lesion segmentation using semi-supervised learning technique," 2017, *arXiv:1703.04301*. [Online]. Available: http://arxiv.org/abs/1703.04301

[18] W. Bai, O. Oktay, and M. E. A. Sinclair, "Semi-supervised learning for network-based cardiac MR image segmentation," in *Proc. MICCAI*. Springer, 2017, pp. 253–260.

[19] D. Nie, Y. Gao, L. Wang, and D. Shen, "ASDNet: Attention based semi-supervised deep networks for medical image segmentation," in *Proc. MICCAI*. Springer, 2018, pp. 370–378.

[20] C. S. Perone and J. Cohen-Adad, "Deep semi-supervised segmentation with weight-averaged consistency targets," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 12–19.

[21] P.-A. Ganaye, M. Sdika, and H. Benoit-Cattin, "Semi-supervised learning for segmentation under semantic constraint," in *Proc. MICCAI*. Springer, 2018, pp. 595–602.

[22] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Proc. ICLR*, 2017, pp. 1–13.

[23] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow, "Harmonic networks: Deep translation and rotation equivariance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jul. 2017, pp. 5028–5037.

[24] X. Li, L. Yu, H. Chen, C.-W. Fu, and P.-A. Heng, "Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model," in *Proc. BMVC*, 2018, pp. 1–12.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[26] H. Jégou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 117–128, Jan. 2011.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[28] A. Chartsias *et al.*, "Factorised spatial representation learning: Application in semi-supervised myocardial segmentation," in *Proc. MICCAI*. Springer, 2018, pp. 490–498.

[29] Y. Xia *et al.*, "3D semi-supervised learning with uncertainty-aware multi-view co-training," 2018, *arXiv:1811.12506*. [Online]. Available: http://arxiv.org/abs/1811.12506

[30] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *Proc. ICML*, 2016, pp. 2990–2999.

[31] S. Dieleman, J. D. Fauw, and K. Kavukcuoglu, "Exploiting cyclic symmetry in convolutional neural networks," in *Proc. ICML*, 2016, pp. 1889–1898.

[32] X. Li, L. Yu, C.-W. Fu, and P.-A. Heng, "Deeply supervised rotation equivariant network for lesion segmentation in dermoscopy images," in *Proc. Int. Workshop Comput.-Assist. Robot. Endoscopy*, 2018, pp. 235–243.

[33] M. Emre Celebi, Q. Wen, S. Hwang, H. Iyatomi, and G. Schaefer, "Lesion border detection in dermoscopy images using ensembles of thresholding methods," *Skin Res. Technol.*, vol. 19, no. 1, pp. e252–e258, Feb. 2013.

[34] T. Heimann and H.-P. Meinzer, "Statistical shape models for 3D medical image segmentation: A review," *Med. Image Anal.*, vol. 13, no. 4, pp. 543–563, Aug. 2009.

[35] J. Cheng *et al.*, "Superpixel classification based optic disc and optic cup segmentation for glaucoma screening," *IEEE Trans. Med. Imag.*, vol. 32, no. 6, pp. 1019–1032, Jun. 2013.

[36] M. D. Abràmoff *et al.*, "Automated segmentation of the optic disc from stereo color photographs using physiologically plausible features," *Invest. Ophthalmol. Vis. Sci.*, vol. 48, pp. 1665–1673, Apr. 2007.

[37] G. Chlebus, A. Schenk, J. H. Moltz, B. van Ginneken, H. K. Hahn, and H. Meine, "Automatic liver tumor segmentation in CT with fully convolutional neural networks and object-based postprocessing," *Sci. Rep.*, vol. 8, no. 1, p. 15497, Dec. 2018.

[38] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. MICCAI*. Springer, 2016, pp. 424–432.

[39] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*. Springer, 2015, pp. 234–241.

[40] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis.*, Oct. 2016, pp. 565–571.

[41] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Trans. Med. Imag.*, vol. 36, no. 4, pp. 994–1004, Apr. 2017.

[42] H. Fu *et al.*, "Disc-aware ensemble network for glaucoma screening from fundus image," *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2493–2501, Nov. 2018.

[43] J. H. Tan, U. R. Acharya, S. V. Bhandary, K. C. Chua, and S. Sivaprasad, "Segmentation of optic disc, fovea and retinal vasculature using a single convolutional neural network," *J. Comput. Sci.*, vol. 20, pp. 70–79, May 2017.

[44] F. Lu, F. Wu, P. Hu, Z. Peng, and D. Kong, "Automatic 3D liver location and segmentation via convolutional neural network and graph cut," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 12, no. 2, pp. 171–182, Feb. 2017.

[45] D. Yang *et al.*, "Automatic liver segmentation using an adversarial image-to-image network," in *Proc. MICCAI*. Springer, 2017, pp. 507–515.

[46] Y. Yuan, M. Chao, and Y.-C. Lo, "Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance," *IEEE Trans. Med. Imag.*, vol. 36, no. 9, pp. 1876–1886, Sep. 2017.

[47] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[48] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *Proc. NIPS*, 2016, pp. 1163–1171.

[49] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. NIPS*, 2017, pp. 1195–1204.

[50] G. Huang, Z. Liu, L. van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, Jul. 2017, vol. 1, no. 2, p. 3.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[52] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *Proc. NIPS Autodiff Workshop*, 2017.

[53] N. C. Codella *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *Proc. IEEE 15th Int. Symp. Biomed. Imag.*, Apr. 2018, pp. 168–172.

[54] P. Bilic *et al.*, "The liver tumor segmentation benchmark (LiTS)," 2019, *arXiv:1901.04056*. [Online]. Available: http://arxiv.org/abs/1901.04056

[55] H. Seo, C. Huang, M. Bassenne, R. Xiao, and L. Xing, "Modified U-Net (mU-Net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in CT images," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1316–1325, May 2020.

[56] 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. MICCAI*. Springer, 2017, pp. 408–416.

[57] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," in *Proc. BMVC*, 2018, pp. 1–12.

[58] Y. Yuan and Y.-C. Lo, "Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 519–526, Mar. 2019.

[59] G. Venkatesh, Y. Naresh, S. Little, and N. E. O'Connor, "A deep residual architecture for skin lesion segmentation," in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Springer, 2018, pp. 277–284.

[60] M. Berseth, "ISIC 2017-skin lesion analysis towards melanoma detection," 2017, *arXiv:1703.00523*. [Online]. Available: https://arxiv.org/abs/1703.00523

[61] L. Bi, J. Kim, E. Ahn, and D. Feng, "Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks," 2017, *arXiv:1703.04197*. [Online]. Available: http://arxiv.org/abs/1703.04197

[62] J. I. Orlando *et al.*, "REFUGE challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Med. Image Anal.*, vol. 59, Jan. 2020, Art. no. 101570.

[63] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. J. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," in *Proc. NIPS*, 2018, pp. 3235–3246.

[64] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 113–123.

[65] S. Lim, I. Kim, T. Kim, C. Kim, and S. Kim, "Fast autoaugment," in *Proc. NIPS*, 2019, pp. 6662–6672.

[66] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, Feb. 2020.

Xiaomeng Li received the Ph.D. degree from the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, in 2019.

She is currently a Post-Doctoral Research Fellow with the Department of Radiation Oncology, Stanford University, Stanford, CA, USA. Her research interests include medical image analysis, computer vision, and deep learning.

Lequan Yu (Member, IEEE) received the B.Eng. degree from the Department of Computer Science and Technology, Zhejiang University, Hangzhou, China, in 2015, and the Ph.D. degree from the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, in 2019.

He is currently a Post-Doctoral Research Fellow with the Department of Radiation Oncology, Stanford University, Stanford, CA, USA. His research lies at the intersection of medical image analysis and artificial intelligence.

Hao Chen (Member, IEEE) received the B.Eng. degree from the School of Information Engineering, Beihang University, Beijing, China, in 2013, and the Ph.D. degree from the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, in 2017.

He focuses on translating advanced deep learning techniques into real-world clinical products.
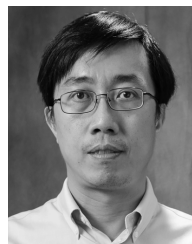
Chi-Wing Fu (Member, IEEE) is currently an Associate Professor with The Chinese University of Hong Kong, Hong Kong. His recent research interests include computation fabrication, 3-D computer vision, user interaction, and data visualization.

Dr. Fu served as a Panel Member for SIGGRAPH 2019 Doctoral Consortium and a program committee member in various research conferences, including the SIGGRAPH Asia Technical Brief, the SIGGRAPH Asia Emerging Technology, the IEEE Visualization, computer vision and pattern recognition (CVPR), the IEEE Conference on Virtual Reality (VR) and 3D User Interfaces, Virtual Reality Software and Technology (VRST), Pacific Graphics, and good manufacturing practices (GMP). He served as the Co-Chair for SIGGRAPH ASIA Technical Brief and Poster Program in 2016. He also served as an Associate Editor for the *Computer Graphics Forum*

Lei Xing received the Ph.D. degree in physics from Johns Hopkins University, Baltimore, MD, USA, in 1992.

He is currently the Jacob Haimson Professor of medical physics and the Director of the Department of Radiation Oncology, Medical Physics Division, Stanford University, Stanford, CA, USA, where he has been a member of the Radiation Oncology Faculty, since 1997. His research has been focused on inverse treatment planning, artificial intelligence in medicine, tomographic image reconstruction, CT, optical, and positron emission tomography (PET) imaging instrumentations, image-guided interventions, nanomedicine, and applications of molecular imaging in radiation oncology.

Pheng-Ann Heng (Senior Member, IEEE) received the B.Sc. degree in computer science from the National University of Singapore, Singapore, in 1985, and the M.Sc. degree in computer science, the M.Art. degree in applied math, and the Ph.D. degree in computer science from Indiana University, Bloomington, IN, USA, in 1987, 1988, and 1992, respectively.

He was the Head of the Graduate Division, The Chinese University of Hong Kong, Hong Kong, from 2005 to 2008 and from 2011 to 2016, and the Chairman with the Department of Computer Science and Engineering from 2014 to 2017, where he is currently a Professor. His research interests include artificial intelligence and the IEEE Conference on Virtual Reality (VR) and 3D User Interfaces for medical applications, surgical simulation, visualization, graphics, and human–computer interaction.