Computers in Biology and Medicine 156 (2023) 106493

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Semi-supervised segmentation of coronary DSA using mixed networks and multi-strategies

Yao Pu [a,b,1], Qinghua Zhang [c,1], Cheng Qian [a], Quan Zeng [a], Na Li [d], Lijuan Zhang [a,**], Shoujun Zhou [a,*], Gang Zhao [e]

[a] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, China
[b] University of Chinese Academy of Sciences, Beijing, 100049, China
[c] Department of Neurosurgery, The 6th Affiliated Hospital of Shenzhen University, Huazhong University of Science and Technology Union Shenzhen Hospital, 518052, Shenzhen, Guangdong, China
[d] Department of Biomedical Engineering, Guangdong Medical University, Dongguan, Guangdong, 523808, China
[e] Neurosurgery Department, General Hospital of Southern Theater Command, PLA, Guangzhou, China

ARTICLE INFO

Keywords:
Semi-supervised
Inception-SwinUnet
Pyramid-consistency learning
Confidence learning

ABSTRACT

The coronary arteries supply blood to the myocardium, which originate from the root of the aorta and mainly branch into the left and right. X-ray digital subtraction angiography (DSA) is a technique for evaluating coronary artery plaques and narrowing, that is widely used because of its time efficiency and cost-effectiveness. However, automated coronary vessel classification and segmentation remains challenging using a little data. Therefore, the purpose of this study is twofold: one is to propose a more robust method for vessel segmentation, the other is to provide a solution that is feasible with a small amount of labeled data. Currently, there are three main types of vessel segmentation methods, *i.e.*, graphical- and statistical-based; clustering theory based, and deep learning-based methods for pixel-by-pixel probabilistic prediction, among which the last method is the mainstream with high accuracy and automation. Under this trend, an Inception-SwinUnet (ISUnet) network combining the convolutional neural network and Transformer basic module was proposed in this paper. Considering that data-driven fully supervised learning (FSL) segmentation methods require a large set of paired data with high-quality pixel-level annotation, which is expertise-demanding and time-consuming, we proposed a Semi-supervised Learning (SSL) method to achieve better performance with a small amount of labeled and unlabeled data. Different from the classical SSL method, *i.e.*, Mean-Teacher, our method used two different networks for cross-teaching as the backbone. Meanwhile, inspired by deep supervision and confidence learning (CL), two effective strategies for SSL were adopted, which were denominated Pyramid-consistency Learning (PL) and Confidence Learning (CL), respectively. Both were designed to filter the noise and improve the credibility of pseudo labels generated by unlabeled data. Compared with existing methods, ours achieved superior segmentation performance over other FSL and SSL ones by using data with a small equal number of labels. Code is available in https://github.com/Allenem/SSL4DSA.

1. Introduction

1.1. Background

According to the World Health Organization [1] and some studies [2,3], coronary artery disease (CAD) is the leading cause of death in both developed and developing countries. Early detection can be beneficial in the treatment of CAD [4]. As an imaging modality with better visualization, DSA can be used to identify the coronary plaques and narrowing, allowing for the early detection of CAD. During DSA procedure, a mask image is captured before the introduction of contrast medium, followed by a series intensifier imaging of the targeted area at a set rate (1–7.5 frames per second) after the intensifier is administered. A set of DSA images are acquired by subtracting the mask from the subsequent

* Corresponding author.
** Corresponding author.
E-mail addresses: lj.zhang@siat.ac.cn (L. Zhang), sj.zhou@siat.ac.cn (S. Zhou).
[1] These authors contributed equally.

https://doi.org/10.1016/j.compbiomed.2022.106493
Received 19 October 2022; Received in revised form 11 December 2022; Accepted 27 December 2022
Available online 30 December 2022

images. In general, the background appears light gray, and structures with contrast passing appear dark gray.

Many segmentation methods were proposed to distinguish the background and the vessels. In the past few years, convolutional neural networks (CNNs) have achieved milestones in the medical image segmentation field, especially for the purpose of image-guided intervention and radiation therapy [5,6]. Among them, U-shaped architecture and skip-connections (e.g., ResNet) have shown good advantages and promise for natural image segmentation and feature extraction. Unet [7] is the typical U-shaped network composed of the encoder, decoder, and skip connections. Many networks for angiographic image segmentation have been improved based on the basic Unet structure. However, it is difficult to learn global and long-range semantic information due to the limitation of convolution operation locality. In contrast, Transformer has achieved a better performance in natural language processing (NLP) [8] and computer vision (CV) [9], with strong ability of multi-head self-attention (MSA) mechanism to build long dependencies in long sequences. Combined the convolution and transformer, researchers proposed the Inception Transformer (iFormer) [10], which contains a mixer for splitting and mixing high- and low-frequency using two different operations. This architecture exhibits outstanding performance on image classification.

1.2. Related work

Segmentation methods for coronary angiographic image can be divided into three categories according to Ref. [11], *i.e.*, region growth methods, partitioning methods and pixel-wise probability map (PBM) prediction. As the most commonly used traditional method, ***region growth*** relies on iteratively adding neighboring pixel points that are similar enough until each region contains a class. Different criteria including homogeneity, edge attractiveness, curvature and other image features can be used for a decision process during adding new pixel points. With a large number of potential criteria used for region growth, the flexibility and manual tunability of the parameters remain tricky for specific applications. ***Partitioning methods*** include clustering algorithms such as k-means clustering, and graph-based methods which represent the image pixels/voxels in a graph and then partition the graph to some classes. Partitioning methods are not as commonly used as the other categories, possibly due to the additional constraints and difficulties of acquiring prior knowledge. ***Pixel-wise PBMs*** operate on the individual pixel by using intensity, tubular characteristics, or other features extracted via statistics and deep learning methods. In this category, the individual pixel is classified as yes or no to build artery tree. Generally, morphological opening and erosion operations of traditional methods are often used to remove artifacts, then, threshold filters are often used to filter out non-tubular objects, based on different contrasts in between the vessel and the surrounding tissue in image. Besides, a new set of fractional-order Legendre moments and deep neural networks [12] are used to extract the features from the images encrypted by fractional discrete Meixner moments [13]. Recently, with the improvement of computer power and the development of deep learning algorithms, a series of neural networks based on convolutional and Transformer operations have outperformed the other traditional methodologies in both accuracy and time-consuming [11].

For vascular segmentation tasks, fully-supervised learning (FSL) has achieved state-of-the-art results in the past few years [14]. However, FSL method required a large number of annotated data to feed the network to train the model for better performance. Manually annotating of coronary artery in DSA is reliable, but also time-consuming and laborious. Fortunately, many works about semi-supervised learning (SSL) method were proposed, which attempt to train a powerful model using only a small amount of labeled data and a lot of unlabeled data. So, the key step is how to use the unlabeled data effectively. There are two main categories of SSL methods: Pseudo-labeled-based iterative learning strategy [15] and consistency-based joint training strategy [16]. The former category of methods is usually implemented using adversarial learning [17]. In this category, unlabeled data is fed into the network trained by labeled data and then pseudo labels are generated. After filtering the noise of pseudo labels, they are used to calculate the loss and correct the parameters of previous network. In some methods, uncertainty or confidence estimation of pseudo labeling is introduced into the training process [18,19], thus the negative impact of noise from pseudo labeling is reduced. In the other aspect, although the noise is reduced by the previous methods, the outputs may differ because of the inherent dropout architecture and the different fed iterations. In order to achieve a better consistency in the prediction, loss functions that minimize the prediction variance were proposed in some representative networks, such as TCSM_v2 [20], SCO-SSL [21], DDT [22], SASSNet [23], etc. These belong to the second category.

1.3. Our work

With the FSL, we proposed a new segmentation network Inception-SwinUnet (ISUnet) combining convolutional and shift window attention. Its performance is comparable to classical Unet and better than the SwinUnet [24]. With the SSL combined the pseudo label and consistency theories, we proposed three strategies to optimize the SSL results. The first is two different networks cross teaching, rather than using two homogeneous networks as the student and teacher model for training. Pseudo label cross teaching between the CNN and Transformer reduces the multi-network discrepancy. On one hand, this structure takes full advantage of feature capture capabilities of two networks, and on the other hand, network consistency constrains the networks to progress together. The second is the pyramid-consistency learning, which can reduce the multi-scale discrepancy. Inspired by deep supervision theory [25,26], we proposed pyramid consistency based on the assumption that the predictions of the same object in different scales should be close to each other. The third is the confidence learning [27], which aims to refine the pseudo label of the better network output. Assuming the better network outputs pseudo-labels with noisy data, on which confidence learning and correction operation is performed, then the other network is instructed and optimized by the filtered pseudo label. The last two strategies are mainly created to get the key information and reduce the redundancy of pseudo label, which are inspired by Refs. [28,29]. In summary, our contributions are as follows.

1) A new segmentation network with different operations (convolution and sliding window attention) in different channels is proposed and outperforms SwinUnet.
2) A semi-supervised learning framework with differentiated dual network is proposed for cross-teaching.
3) Pyramid consistency is firstly used for pseudo label learning of unlabeled coronary angiographic image.
4) Confidence learning is skillfully used for pseudo label denoising and optimization.

2. Methodology

2.1. Framework overview

The proposed SSL learning method via pseudo label cross teaching, pyramid consistency, and confidence learning is described in Fig. 1. The inputs of the whole architecture are small amount of labeled gray images with their binary label images, and large amount of unlabeled gray images. The outputs are the probability maps and predicted binary images from two networks, corresponding to ①②③④ in Fig. 1, respectively. And they all have a resolution of 512*512. On one hand, the data with ground truth are fed into the Unet and ISUnet, to calculate the standard supervised loss for training the networks. On the other hand, the unlabeled data are fed into the dual networks as well, only for mutual guidance through pseudo-labeling data generated by the dual

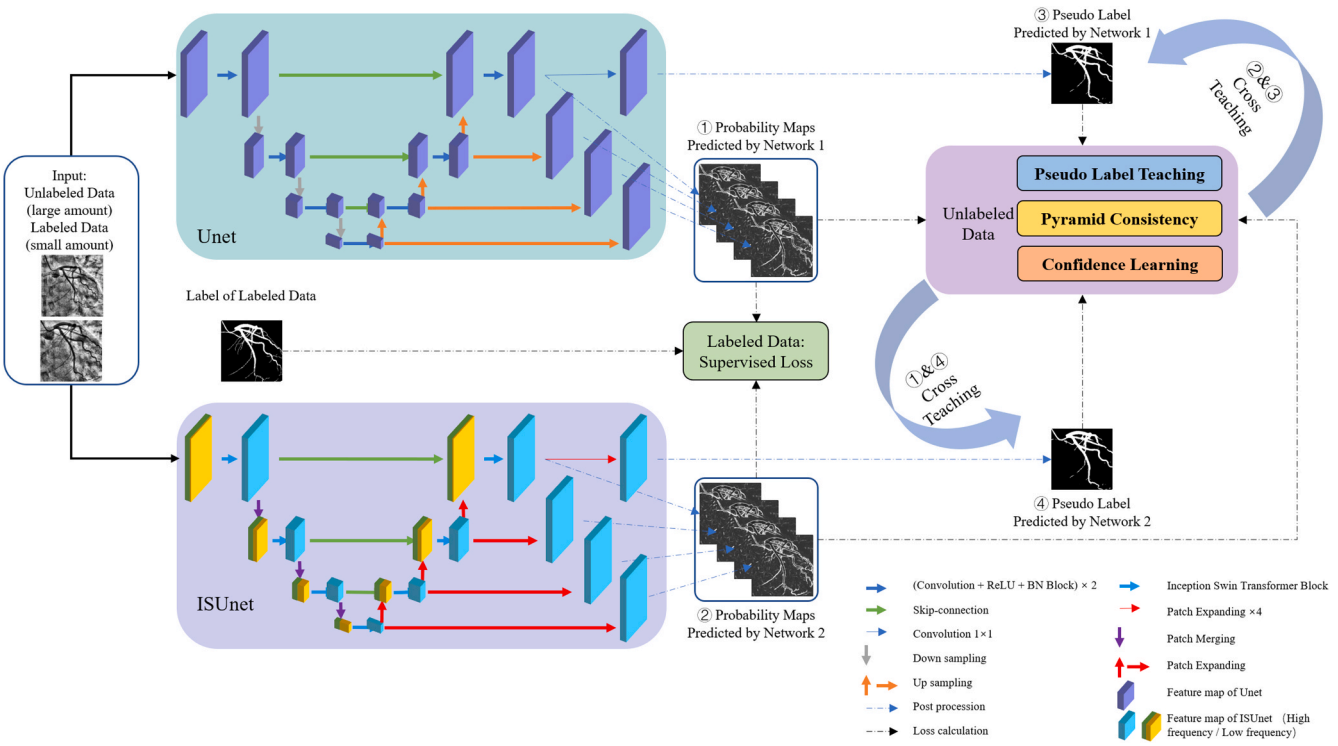


Fig. 1. Illustration of the proposed SSL framework combining supervised loss, pseudo label cross teaching, pyramid consistency learning, and confidence learning. The supervised loss of each network is calculated for updating the respective network parameters separately. The last three losses of unlabeled data are calculated for cross-teaching between the dual networks.

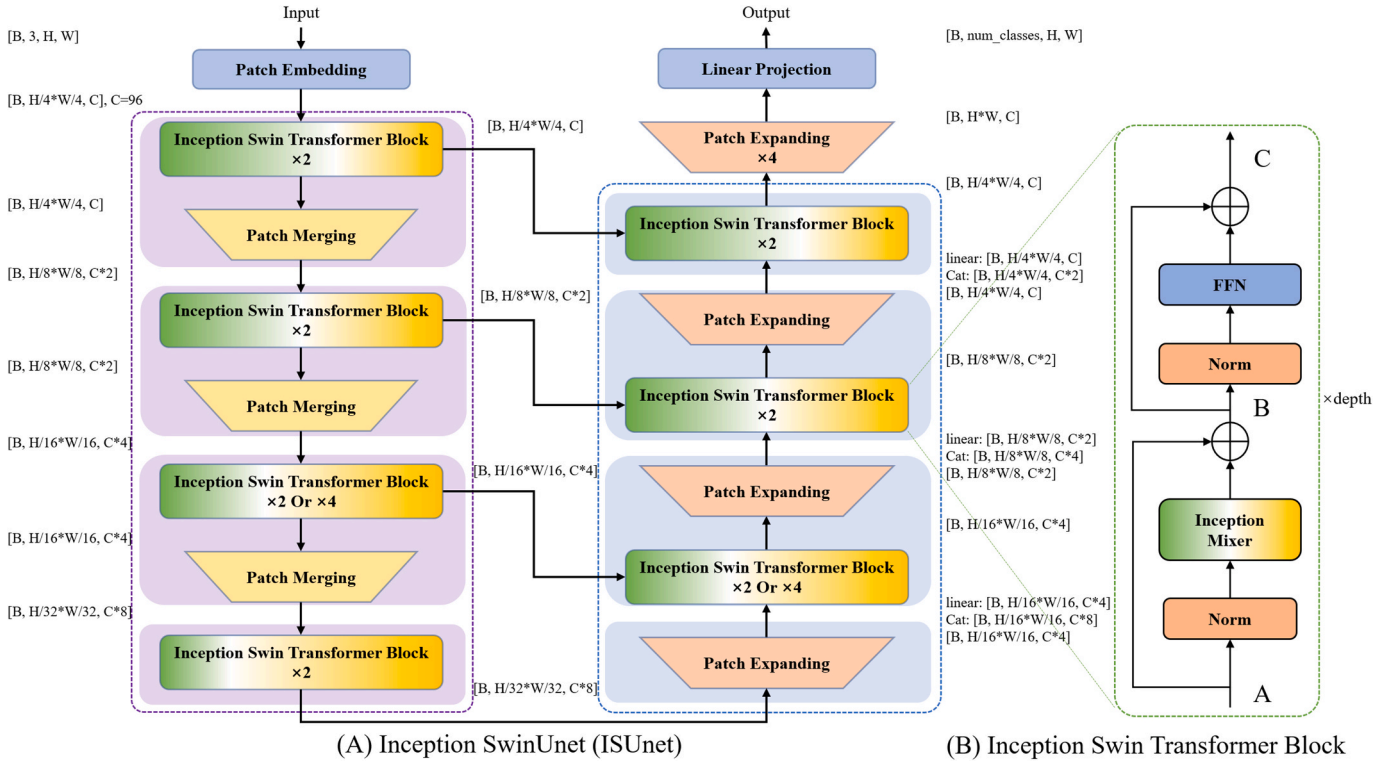


Fig. 2. (A): an architecture of Inception SwinUnet (ISUnet) consisting of encoder, decoder, and skip connections, where the encoder and decoder are all constructed by the Inception transformer blocks. (B): the Inception Transformer Block consists of mixed color in view, and the orange and green indicate low- and high-frequency information, respectively. Another, the information about the batch size, width and height, channels, and some detailed operations of each layer image are also shown in this view.

networks. For the unlabeled data, the pyramid consistency mitigates the differences from deep and shallow scales in the dual networks, while the confidence learning corrects and denoises the weak labels of the outputs. In summary, predicted PBMs for labeled data through the dual networks are used to compute supervised loss and guide the pseudo-labeled data, and the pseudo labels generated by unlabeled data are used to calculated the semi-supervised loss and correct the other network (or say cross-teach) by pyramid consistency and confidence learning.

Unet is one of the backbone networks of our framework and we follow the original setup. It consists of encoder, decoder, and skip-connections. In each layer of the encoder and decoder, 2D convolution, batch normalization, and ReLU activation functions (whose derivative can be calculated easily and quickly), are repeated twice; the encoder and decoder inter-layer operations consist of maximum pooling with length reduced by half and expanded by double, respectively; the skip-connection is implemented by merging encoder same-layer features and upsampled features in the channel dimension. ISUnet is another backbone network. The choice of its pooling & activation function will be explained in Section 3.3 on ablation experiments.

In the next Section 2.2, we describe the details of the proposed ISUnet. In Section 2.3, the supervised learning and the SSL strategies for the labeled and unlabeled data are described, separately. Finally, we summarize the total loss function in Section 2.3.5.

2.2. Implement of Inception SwinUnet

2.2.1. ISUnet overview

Inspired by Inception Transformer [10], Unet [7], and SwinUnet [24], ISUnet consists of the encoder, decoder, and skip connections. As shown the general ISUnet architecture in Fig. 2-left, the basic unit of it is the Inception Swin Transformer (IST) block. For the encoder, the medical image is split into non-overlapping patches with each patch size of 1/(4 × 4), only to transform the input into a sequential embedding. By this partition, the feature dimension of each token is changed from 3 channels to 4 × 4 × 3 = 48. In addition, a linear embedding layer is applied to project the feature dimension to an arbitrary dimension (denoted as C, default is 96). In the encoder, the transformed patch tokens pass through several IST blocks and Patch Merging layers to generate hierarchical feature representations. Specifically, the IST blocks are responsible for feature representation learning, while the Patch Merging layer is responsible for down sampling to reduce resolution and increase dimensionality. In Patch Merging operation, the feature map is divided into 4 parts in height- and width-dimensions, and then it is merged according to channel with a resolution of 1/2*1/2 and a dimension of 4 × , and finally the dimension is reduced to 2 × by a linear layer. The symmetric Inception Transformer decoder consists of IST blocks and Patch Expanding layers. The extracted contextual features are concatenated with the multiscale encoder features through skip-connections to complement the spatial information loss due to downsampling. In contrast to the Patch Merging layer, the Patch Expanding layer comes to perform upsampling. The Patch Expanding layer expands the channel to 2 times, and then enlarges the neighboring feature maps with a resolution of 2 × 2 × and reduces the dimension to 1/2. Finally, 4 × upsampling is performed using the last facet extension layer, to restore the resolution of the feature map to the input resolution (W × H). Then, a linear projection layer is applied to these up-sampled features to output pixel-level segmentation predictions. The details of each module will be elaborated in the following sections.

2.2.2. Inception Swin Transformer Block

The Inception Swin Transformer Block (IST) is mainly responsible for the extraction of image features, where Inception Token Mixer (ITM) is used. Unlike convolution-only operation in Unet or multi-headed self-attention (MSA) only in SwinUnet, the ITM employs multi-headed attention at low frequencies and convolution at high frequencies, and its' superiority is confirmed in Ref. [10]. Also, the structure of Inception Mixer and feedforward network (FFN) plus Layer Norm (LN) are adopted with residuals connecting, as shown in Fig. 2-right. Generally, the deeper the network, the less variation between layers, and the more difficult to learn its signal changing. In contrast, the residual learning can approach the variation between layers in easy-to-learn. The number of IST blocks repetitions in different layers is set by the depth parameter. The final Inception Swin Transformer Block is formally defined as follows.

$$B = A + ITM(LN(A)) \tag{1}$$

$$C = B + FFN(LN(B)) \tag{2}$$

where $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ represent input, intermediate variables and output of IST block. ITM, FFN and LN represent the operation of Inception Token Mixer, Feedforward Network and Layer Norm.

2.2.3. Inception Mixer

Inception Mixer is the key module of the IST block, in which the feature maps are split into two parts, low and high frequency (**l** and **h**) respectively along the channel for different operations, instead of directly feeding image patch tokens into the convolution or MSA. Here, input feature maps and the outputs are denoted as **X**, **Y**. As shown in Fig. 3, the orange background represents the low-frequency channel, and the green represents the high part. For U-shaped architecture, the higher layers have greater resolution and smaller dimensionality, which plays a greater role in capturing high-frequency information, and vice versa. Therefore, a frequency ramp transformation can be used when splitting low and high frequencies, as shown in the following equation.

$$\frac{C_l}{C} = k\, s + b \tag{3}$$

$$\frac{C_l}{C} + \frac{C_h}{C} = 1 \tag{4}$$

where $\boldsymbol{C_l}$, $\boldsymbol{C_h}$, and $\boldsymbol{C}$ represent the number of low-frequency, high-frequency, and all channels, respectively; $\boldsymbol{C_l} + \boldsymbol{C_h} = \boldsymbol{C}$; $\boldsymbol{k}$ and $\boldsymbol{b}$ represent the corresponding slope and intercept of the ramp function; $\boldsymbol{s}$ represents the layer index of the block. Note, the inputs $\boldsymbol{X} \in R^{N\times C}$ are split into $\boldsymbol{X}_l \in R^{N\times C_l}$ and $\boldsymbol{X}_h \in R^{N\times C_h}$.

Low-frequency Mixer. Unlike the MSA only approach in the original Inception Transformer, the operation in the low-frequency part is constructed with shifted window attention. The low-frequency image patch tokens are fed into the average pooling, attention, and upsample module step by step. The outputs $\boldsymbol{Y}_l$ of this part can be calculated as:

$$\boldsymbol{Y}_l = Upsample(Attention(Avgpool(\boldsymbol{X}_l))) \tag{5}$$

The average pooling can reduce the spatial scale of $\boldsymbol{X}_l$, while upsample following attention is used to recover the same resolution of $\boldsymbol{X}_l$. The above design can largely reduce the computational overhead. Fig. 3-right presents attention unit details, where two consecutive base blocks are connected. Each block is composed of LayerNorm (LN) layer, multi-head self-attention (MSA) module, residual connection, MLP with 2-layers linear and GELU nonlinear mapping. The window of multi-head self-attention (W-MSA) and the shifted window multi-head self-attention (SW-MSA) are applied in the first and second MSA module, where the W-MSA and SW-MSA take full advantage of the Transformer to obtain the association information of the larger receptive field. Consecutive base blocks are computed according to:

$$\hat{z}^l = WMSA\left(LN\left(z^{l-1}\right)\right) + z^{l-1},$$

$$z^l = MLP\left(LN\left(\hat{z}^l\right)\right) + \hat{z}^l,$$

$$\hat{z}^{l+1} = SWMSA\left(LN\left(z^l\right)\right) + z^l,$$

$$z^{l+1} = MLP\left(LN\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1} \tag{6}$$

where $\hat{z}^l$ and z^l represent the outputs of the (S)W-MSA module and the MLP module of the l_{th} block, respectively.

High-frequency Mixer. A maximum pooling and convolutional parallel structure are adopted to take advantage of the sensitivity of maximum pooling and the detail-awareness of convolution. As a result, the high-frequencies are divided into two parts with an equal number of channels, *i.e.*, $X_{h1} \in R^{N\times\frac{C_h}{2}}$, $X_{h2} \in R^{N\times\frac{C_h}{2}}$. As shown the green background in Fig. 3-left, the maximum pooling and linear mapping operation are performed for X_{h1}, and the linear mapping and convolution operation are performed for X_{h2}, then we get following two outputs Y_{h1} and Y_{h2}:

$$Y_{h1} = Linear(MaxPool(X_{h1})) \tag{7}$$

$$Y_{h2} = Conv(Linear(X_{h2})) \tag{8}$$

The above three parts are eventually concatenated into one feature map, which is then fed into a convolution and a linear layer with residual connection. And this structure is designed to overcome the weakness of over-smoothing of neighboring pixels caused by upsample at low frequency, while the involved calculation process is as follows.

$$Y_c = Concat(Y_l, Y_{h1}, Y_{h2}) \tag{9}$$

$$Y = Linear(Conv(Y_c) + Y_c) \tag{10}$$

2.3. Supervised and SSL strategies

During the training process, we divide the data into two parts, *i.e.*, the labeled data and unlabeled data. For a small amount of labeled data, we use a FSL approach, while for a large amount of unlabeled data, we use pseudo label cross-teaching, pyramid consistency learning, and confidence learning. The experimental DSA sets involve the left and right coronary arteries with equal number. In this task, the entire training set contains N samples, in which M are manually labeled with $M < N$. We denote the labeled dataset as $S_L = \{X_L(i), Y_L(i)\}$ and unlabeled data as $S_U = \{X_U(i)\}$, where $X(i) \in R^{\Omega(i)}$ and $Y_L(i) \in \{0,1\}^{\Omega(i)}$ represent the input images and the labels of binary segmentation, respectively. Meanwhile, we denote the feature map as $X = \{X_l, X_h\}$ with X_l and X_h represent the low- and high-frequency data, respectively.

2.3.1. Details about supervised learning

To make full use of a small number of label information, we perform a detailed loss calculation for the labeled part. Specifically, output of per scale and per network is linearly interpolated to the original image size as a prediction. And the error between the ground truth and the predicted image is then calculated. Finally, we calculate the weighted average of the losses for each scale as the supervised loss. The weight value is decreasing along with layer depth increasing, the reason is that the deeper the layer, the lower the resolution, and the lower the credibility. Note, both cross-entropy and Dice are used as loss evaluation of supervised learning. The formula for quantification calculation is as follows.

$$soft\ dice(P, Y) = \frac{2|P \cap Y| + \varepsilon}{|P|^2 + |Y|^2 + \varepsilon} \tag{11}$$

$$L_{dice}(P, Y) = 1 - soft\ dice(P, Y) \tag{12}$$

$$L_{ce}(P, Y) = -\frac{1}{N}\sum_i \sum_{j=1}^{C} y_{ij} \log\left(p_{ij}\right) \tag{13}$$

$$L_{supervise,n} = 0.5*\left(\sum_s \alpha_s \left(L^s_{dice} + L^s_{ce}\right)\right) \tag{14}$$

where Y is the ground truth matrix; P is the predicted PBM of each layer output; ε is the smoothing parameter; N denotes the total number of pixel points; C denotes the total number of categories (2 in our task); y_{ij} denotes the value (0 or 1) of the i_{th} pixel of matrix Y belonging to the j_{th} category; p_{ij} denotes the probability of the i_{th} pixel of the prediction matrix P belonging to the j_{th} category; s denotes the scale index of the layer; α_s is the weight of the corresponding scale; L^s_{dice} and L^s_{ce} denote the dice loss and cross entropy loss of the s_{th} scale.

2.3.2. Pseudo learning cross teaching strategy

Pseudo-label cross teaching between the CNN and Transformer is adopted to reduce the multi-network discrepancy. As present before, the labeled data is trained firstly. After some iterations, unlabeled data is fed into the different pre-trained networks to generated two groups of PBMs and segmentation results (pseudo label). We let the results network 0 guide the PBM of network 1, and vice versa. In this procession, we denote the predicted PBM of network n after activation as $OUT_{soft,n}$, and pseudo label after argmax operation as $OUT_{pseudo,n}$, where n ∈ {0,1}, and $(1-n)$ means the other network. The specific calculation formula is as follows.

$$L_{pseudo,n} = L_{dice}\left(OUT_{soft,n}, OUT_{pseudo,(1-n)}\right)$$

2.3.3. Pyramid-consistency learning strategy

It is necessary to reduce the multi-scale discrepancy. In each network, because of the different resolutions in the pyramid structure, we use re-sampling to let the multi-scale outputs have the same resolution. Considering different spatial frequencies of these outputs, *i.e.*, the deeper the layer, the more low-frequency components are captured, which can cause different detail information loss or model collapse on images. We use pyramid uncertainty estimation and uncertainty correction for model training to overcome this problem. The specific steps are as follows: first, for each network, we calculate the weighted average ($p_{avg,n}$) of the PBMs in different scales. According to the general SSL approach, the weighted average can be used as a pseudo-label for the loss calculation of another network after the argmax operation. However, we add an uncertainty weighting factor as a correction term (Var) for calculating the differences between $p_{avg,(1-n)}$ and p_s. The parameters can be calculated as follows.

$$p_{avg,n} = \sum_S \alpha_s \cdot OUT^s_{soft,n} \tag{15}$$

$$Var^i_{s,n} = KL\left(p^i_s \middle\| p^i_{avg,(1-n)}\right) = \sum_{j=1}^{C} p^{i,j}_s \cdot \ln \frac{p^{i,j}_s}{p^{i,j}_{avg,(1-n)}} \tag{16}$$

where $p_{avg,n}$ denotes the weighted average of PBMs in the n_{th} network (multi-scale); $Var^i_{s,n}$ means the uncertainty on the s_{th} scale and the i_{th} pixel in the n_{th} network; and uncertainty of the n_{th} network is calculated according to the weighted average PBM of the other network (1-n); C denotes the classes in the images (here is 2, and j is one of it); KL means that we adopt the Kullback-Leibler divergence as the uncertainty measurement.

By calculating the uncertainty, the correction weight parameters can be further calculated in pixel-wise operation and conforms to the principle of high uncertainty with low weight. Then, the corrected pyramidal consistency loss is calculated by combining the mean square error (MSE) and the correction weight parameter, which is the first term of the final pyramid loss. Considering that decrease of the prediction entropy can improve the model robustness [30], we add the average uncertainty directly to the previous loss, such as follows:

$$L_{pl,n} = \frac{1}{S}\sum_{s=1}^{S}\left[\beta_{pl} \cdot \frac{\sum_{i,j}\left(\left\|p^i_s - p^{i,j}_{avg,(1-n)}\right\|^2 \cdot w^i_s\right)}{\sum_i w^i_s} + \left(1-\beta_{pl}\right)\cdot Var^i_{s,n}\right] \tag{17}$$

where w_s^i is the correction weight parameter $w_s^i = e^{-Var_s^i}$, and p_s^i is PBM value for i_{th} pixel at the s_{th} scale; $L_{pl,n}$ is the pyramid consistency loss of network n; a weight to balance the impact of the two terms is written as β_{pl}, which is set as 0.5 according to Ref. [30].

2.3.4. Confidence learning strategy

Some parameters are dropped in the forward pass due to the Dropout structure of the network, yet the final output results should be consistent. T copies of the unlabeled data are replicated, and fed into the network separately, until the finally result of the average output softmax PBM is acquired. The predictive entropy of the mean is used as a measure of uncertainty U and is normalized afterward, *i.e.*,

$$p_{avg} = \frac{1}{T}\sum_t p_t \tag{18}$$

$$u = -\frac{\sum_{c=1}^{C} p_{avg}\ \ln\left(p_{avg}\right)}{\ln 2} \tag{19}$$

where p_{avg} is the average of PBM of repeated T copies; u is the normalized uncertainty, one item of the matrix U; ln 2 is the maximum entropy.

Besides the direct use of pseudo labels, it's important to filter the noise of pseudo label to get a better performance. A confidence-learning based self-denoising process is proposed to mitigate the possible misleading effects of pseudo-label noise. Inspired by consulting third parties in arbitration proceedings, we let the output of one network as the third part. Based on the classification noise process (CNP) assumption [31], label noise should be minimized according to class probability. Then a work about image-level classification task, called Confidence Learning (CL) [27], further reveals that the Confidence Joint Matrix [32] is effective in dividing and counting labeling errors. Therefore, we remold the Confidence Learning for one network output, and use it to guide the other. Specifically, on one hand, we input the unlabeled data into the Unet and use its predicted segmentation as the weak label Y_l (where noise need to be corrected), noted as $y_l = i$. On the other hand, the PBM $\widehat{p}_i(x)$ (belongs to class i at pixel x) of ISUnet is used as a third part to correct Y_l. The specific adjustment is as follows: If the PBM of pixel x is greater than a certain threshold, it belongs to this class with a high confidence, otherwise, this pixel category should be adjusted to the class with the highest confidence. We define the threshold of j_{th} class as: $t_j = \frac{1}{|X_{y_l=j}|}\sum_{x\in X_{y_l=j}} \widehat{p}_j(x)$. For example, when $\widehat{p}_j(x) > t_j$, then the latent category y_l^* of pixel x should be j instead of i. First, according to this confidence threshold strategy, the confidence joint matrix is defined as

$$C_{y_l,y_l^*}[i][j] := \left|\widehat{X}_{y_l=i,y_l^*=j}\right| \tag{20}$$

where

$$\widehat{X}_{y_l=i,y_l^*=j} := \left\{ x \in X_{y_l=i} : \widehat{p}_j(x) > t_j, j = \underset{c\in C:\widehat{p}_c(x)>t_c}{\operatorname{argmax}}\ \widehat{p}_c(x) \right\} \tag{21}$$

Then, we get calibrated confidence joint matrix $\widetilde{C}_{y_l,y_l^*}$ and joint distribution matrix $\widetilde{Q}_{y_l,y_l^*}$. They are defined as:

$$\widetilde{C}_{y_l,y_l^*}[i][j] = \frac{C_{y_l,y_l^*}[i][j]}{\sum_{j\in C} C_{y_l,y_l^*}[i][j]} \cdot \left|\widehat{X}_{y_l=i,y_l^*=j}\right| \tag{22}$$

$$\widetilde{Q}_{y_l,y_l^*}[i][j] = \frac{\widetilde{C}_{y_l,y_l^*}[i][j]}{\sum_{i\in C,j\in C} \widetilde{C}_{y_l,y_l^*}[i][j]} \tag{23}$$

Next, we can select the noise X_{err} with lowest self-confidence $\widehat{p}_i(x)$ from weak label by using the joint distribution matrix and the prune-by-class (PBC) strategy:

$$X_{err} = N\cdot \sum_{j\in C\,:\,j\neq i} \left(\widetilde{Q}_{y_l,y_l^*}[i][j]\right) \tag{24}$$

where N means the number of pixels in the image X; '1' (the value of X_{err}) means this pixel is misclassified, vice versa. By using error map X_{err} and normalized uncertainty matrix U calculated before, the rectified weak label $\dot{Y}_l$ is required by

$$\dot{Y}_l = Y_l + X_{err}\cdot(-1)^{Y_l}\cdot(1-U) \tag{25}$$

Finally, the cross entropy and focal loss in between the rectified weak label of Unet and the output of ISUnet are adopted as the confidence learning loss, *i.e.*

$$L_{cl} = 0.5^*\left(L_{focal}(P_{ISUnet}, \dot{Y}_l) + L_{ce}(P_{ISUnet}, \dot{Y}_l)\right) \tag{26}$$

2.3.5. Total loss function

The total loss is a weighted combination of supervised loss and semi-supervised loss for labeled data and unlabeled data respectively. The latter consists of pseudo-label loss, pyramid consistency learning loss, and confidence learning loss. The total loss is calculated by:

$$L_{total} = L_{super,n=0} + L_{supervise,n=1} + \lambda\left(L_{pseudo,n=0} + L_{pseudo,n=1} + L_{pl,n=0} + L_{pl,n=1} + \beta_{cl}\cdot L_{cl}\right) \tag{27}$$

Note, λ is a time-dependent ramp-up trade-off weight function following previous work [16,33], here it is defined by $\lambda(t) = w_{max}\cdot\exp\left[-5^*\left(1-\frac{t}{t_{max}}\right)^2\right]$, where the parameters w_{max}, t, t_{max} are the maximum weight to be set, the current iteration number during training, and the iteration time of maximum ramp up training, respectively. Such a weight change function reduces the misguidance in the initial training phase. Besides, as the weak label confidence learning loss weight, β_{cl} is empirically set to 5 according to the ablation study [34].

2.4. Training algorithm

Two principal iteration frameworks were applied in the training session. First, the labeled enhanced gray images are fed into two backbone networks, and 8 feature maps from 4 different scales & 2 networks are predicted. Feature maps & labels of original images are used to calculate the supervised loss (including dice loss and cross-entropy loss). Second, unlabeled data was added to the loss calculation. The enhanced labeled & unlabeled gray images are fed into two backbone networks, and 8 feature maps from 4 different scales & 2 networks are predicted. Then softmax and argmax operations are performed with 8 soft images and 8 predicted pseudo labels obtained. Pyramid consistency was computed with the average and KL divergence. The confidence learning loss was obtained based on the calibrated confidence joint matrixes. The soft image from net 0 & pseudo label from net 1 are used to calculate the pseudo loss; the soft image, the average and KL divergence from same net are used calculate PL loss; the feature maps from net 0 & corrected

label by calibrated confidence joint matrixes from net 1 are used calculate CL loss. The details for loss calculation were summarized in Algorithm 1.

In addition to the rationality of the algorithm, the computational complexity of the algorithm is also an important factor for generality. Here, we list the computational complexity and number of parameters in different networks in Table 1. The time and space complexity of Unet are $O(\sum_{l=1}^{D}((M_l^2K_l^2C_{l-1}C_l) \bullet (M_l^2K_l^2C_lC_l)))$ and $O(\sum_{l=1}^{D}(K_l^2C_{l-1}C_l + K_l^2C_lC_l + M_l^2C_l))$, where D is the depth of network, l is the index of layer, M is the height or width, K is the convolutional kernel size, C is the channel counts. Each layer contains two convolution operations, so there are two similar factors in the complexity calculation. The time complexity of MSA and W-MSA are $O(\sum_{l=1}^{D}4hwC^2 + 2(hw)^2C)$ & $O(\sum_{l=1}^{D}4hwC^2 + 2M^2hwC)$, where h and w indicate the height and width, respectively; C is the channel number; M is the window size of W-MSA. The time and space complexity will be cut in half during the training due to the average pooling in low-frequency of Inception Mixer.

Algorithm 1. Total Loss Function (the second iteration we picked above)

Algorithm 1 Total Loss Function (the second iteration we picked above)

Input: Feature maps output from network 0 and 1: $M_n \in R^{S\times H\times W}$, where $n \in \{0,1\}$, $N = 2$, $s \in \{0,1,2,3\}$, $S = 4$. Ground truth $GT \in R^{S\times H\times W}$. Labeled batch size $lb \in R$.

Output: Total loss

In one iteration, initialize PBMs $Soft \in R^{S\times H\times W}$ and pseudo-labels $Pseudo \in R^{S\times H\times W}$ for different scale outputs.
for n in {0,1} **do**
 # 1. Calculate the soft images & pseudo labels using feature maps (M) output by net
 # 1. Calculate the supervised loss of each scale for labeled data
for s in S **do**
$Soft_s^n \leftarrow$ Softmax($M_n[s]$, dim=channel);
$Pseudo_s^n \leftarrow$ Argmax($Soft_s^n$, dim=channel);
$Loss_{super}^s \leftarrow 0.5 * (Loss_{dice} + Loss_{ce})$ for $(M_n[s][:lb], GT[:lb])$;
end
$Avg^n \leftarrow$ Mean($Soft_s^n$);
Calculate the Kullback-Leibler divergence KL;
Calculate the corrected the weak label as $\dot{Y}_l$;
 # 2. Calculate the total supervised loss for labeled data & 3 losses for unlabeled data
for labeled data **do**
$Loss_{super}^n \leftarrow w^s \cdot Loss_{super}^s$;
end
for unlabeled data **do**
$Loss_{pseudo}^n \leftarrow L_{dice}(Soft^n, Pseudo^{1-n})$;
$Loss_{pl}^n \leftarrow Mean(\sum_s(\beta_{pl} \cdot MSE(Soft_s^n, Avg^n) + (1-\beta_{pl}) \cdot KL))$;
$Loss_{cl} \leftarrow 0.5 * (Loss_{focal} + Loss_{ce})$ for $(M_n[lb:], \dot{Y}_l[lb:])$
end
end
Calculate the ramp up consistency weight λ;
$Loss_{total} \leftarrow \sum_n(Loss_{super}^n + \lambda(Loss_{pseudo}^n + Loss_{pl}^n)) + \lambda\beta_{cl}Loss_{cl}$

3. Experiments and results

3.1. Dataset introduction

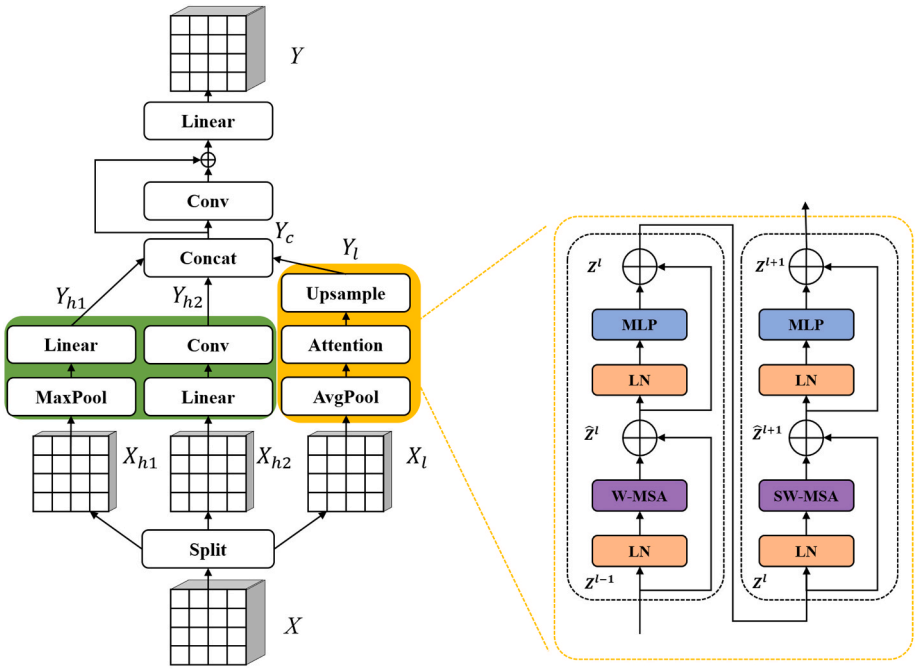
The DSA data for this experiment was obtained from the General Hospital of Southern Theatre Command. It contains 300 sets of coronary X-ray angiographic sequences from 50 patients, including two imaging sessions for the left and right coronaries of each patient, respectively. The videos recorded the arrival and disappearance of the contrast agent through the catheter to the coronary artery. The most clinically significant frames from the video set were selected by the clinician as experimental data, thus a total of 300 sets of 2D DSA images are acquired. Uniformly there are 150 images for each person's left and right coronary angiography with a common resolution of 512 × 512 pixels per image. The corresponding labels were generated by Photoshop software under the guidance of the professional physician.

For the FSL experiments on left coronary angiography, 100 images were selected from the left coronary angiography for training and validation, and the remaining 50 images for testing. For the SSL experiments, we divided the 100 training images into 20 and 80 according to Ref. [35]. During training, 20 original images and their labels were input to the network at the same time. As the rest 80 images, we inputted only the original images without labels. The above-mentioned experiment arrangement was equally applied to the right coronary angiography. Regarding the proportion of labeled data, additional experiments were also conducted and we will discuss them in the next section.

3.2. Implementation details

3.2.1. Preprocessing

Since the original DSA images have quite a few artifacts, a histogram-based contrast adjustment should be used to preprocess the original images before feeding the images into the network. Adaptive histogram equalization (AHE) is a good method of contrast enhancement, which uses the gray values of local windows to construct mapping functions so that the gray values are more uniformly distributed between 0 and 255. To avoid the discontinuity and over-enhancement disadvantages in the result by using AHE, Contrast Limited Adaptive Histogram Equalization (CLAHE) [36] was introduced by Ref. [36]. Different from AHE, the CLAHE-based two superiorities are improved by two tricks: For the first trick, strong capability to restrict the histogram distribution. If the number of a certain gray value in the histogram exceeds the threshold, it is cropped, and then the parts that exceed the threshold are equally distributed to each gray level. Usually, the threshold is set either as the



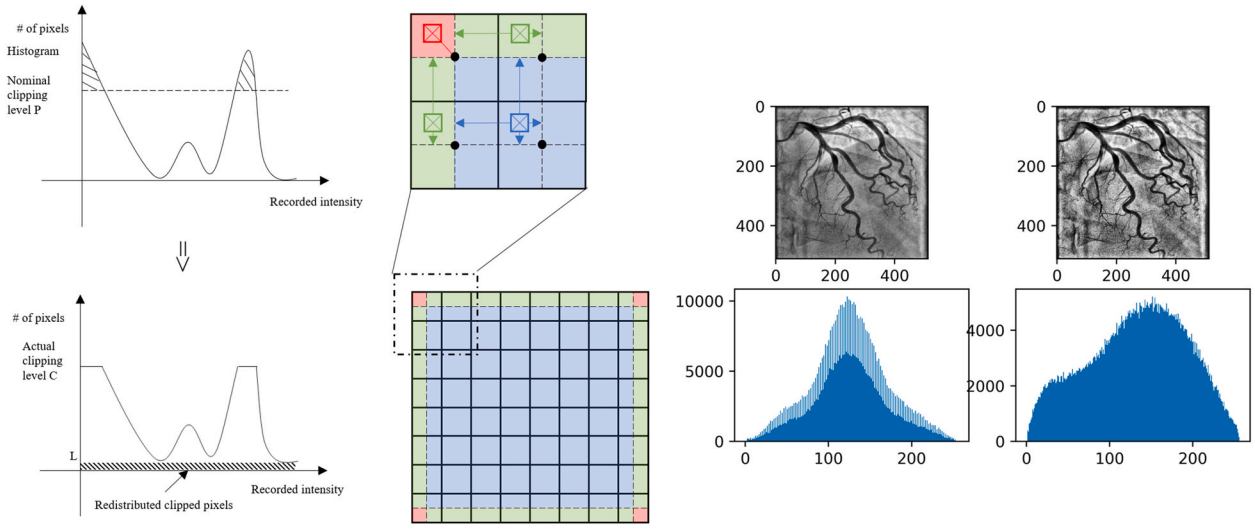
(A) Inception Mixer (B) Attention(W-MSA, SW-MSA)

Fig. 3. (A): The details of Inception Mixer. (B): two successive base blocks, as shown in Eq. (5). W-MSA and SW-MSA are multi-head self-attention modules, respectively with regular and shifted windowing configurations.

Table 1
Computational complexity and number of parameters in different networks.

Network	Computational complexity (GMac)	Number of parameters (M)
Unet	218.73	31.04
Vnet	57.01	8.95
SwinUnet	40.89	41.39
ISUnet (4layers)	23.5	35.98
ISUnet (3layers)	14.91	9.92

occurrence times of gray levels directly, or as a percentage of total pixels. The cumulative distribution function (CDF) map of the resultant image can't be changed too drastically, therefore the noise points with excessive gray value are eliminated. The first trick of CLAHE can be seen in Fig. 4 (A). For the second trick, an interpolation method is proposed to accelerate the histogram equalization. Given the image is chunked into 8×8 and a histogram CDF is computed for each chunk (here named a window), four adjacent windows of each image pixel point are found and their mapping values of the histogram CDF are acquired, in terms of the upper left, upper right, lower left, and lower right windows for a



(A) Distribution Histogram (B) Interpolation (C) Real Image

Fig. 4. Illustration for preprocessing the DSA images. (A): Distribution histogram of CLAHE. (B): Interpolation operation in CLAHE. (C): DSA image and histogram before and after CLAHE.

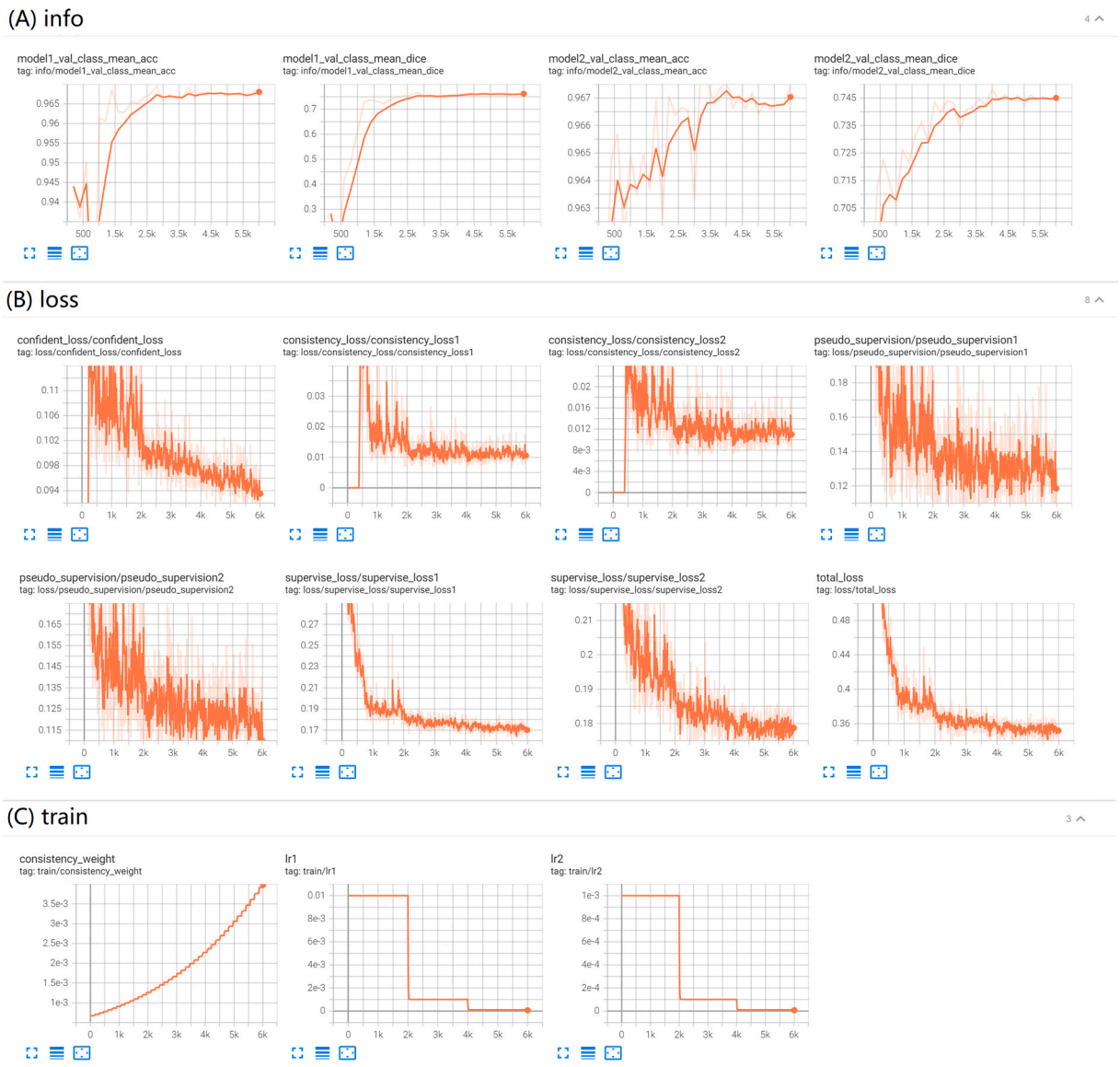


Fig. 5. Training Curves. Subfigures in (A) showed the validation mean accuracy and dice score using model 1 & 2. Subfigures in (B) described all the losses: confidence loss, pyramid consistency loss 1 & 2, pseudo supervision loss 1 & 2, supervise loss 1 & 2 and total loss. Subfigures in (C) showed the variation of the consistency weight and the learning rate of the two models over time.

pixel point. The final mapping value of each pixel point is obtained by imposing the such computational steps as bilinear interpolation (for blue area), linear interpolation (for green area), or the transformation function of the corner tile (for red corner area). The second trick is shown in Fig. 4 (B). Before and after CLAHE processing for the DSA image, both the images and their gray value distribution histograms are shown in Fig. 4 (C). Besides of the CLAHE, the data loading stage involves random transformations like random horizontal flip, random vertical flip, and random rotation, which are used for data enhancement to generalize the applicability of deep learning models, by which the morphological diversity and sample richness of the input images are enriched.

3.2.2. *Training setup and evaluation metrics*

All experiments were implemented on the Windows 10 system with the PyTorch 1.11.0+cu113 framework, 3 NVIDIA GTX 1080Ti GPUs, and the backbone networks of 2D Unet, Vnet, SwinUnet, and ISUnet. For all networks, the outputs of decoder at each layer were upsampled to the original image size for comparison. During the training, the maximum iteration time was set to 6000 [35]. The Adam optimizer was adopted with its weight decay parameter setting of 10^{-4}. We set base learning rate to 10^{-2} for of Unet and Vnet, while 10^{-3} for that of SwinUnet and ISUnet, considering that the convolution operation's convergence rate is faster than Transformer's. The training time scales and images are shown in Figs. 5 and 6, including the validation accuracy/dice, all losses, consistency weight, learning rate, and respective results (transformed original image, label, predictions of model 1 & 2). As shown in figures, the accuracy & dice get higher, the losses get smaller and the prediction image noise gets less as the iterations increase, which is logical. We also did some ablation experiments on the base learning rate. Under the above parameter setting principle, the learning rate becomes one-tenth of the previous one after every 2000 iterations. The batch size of training was set to 4 * 3 (3 GPUs), where half of 4 are labeled & unlabeled data. The random seed was set as 1337 according to Ref. [35]. The base parameters w_{max} and t_{max} were set to 0.1 and 200 for $\lambda(t) = w_{max} \cdot \exp\left[-5^{*}\left(1-\frac{t}{t_{max}}\right)^{2}\right]$.

To digitally present the results of the experiment, average Dice score coefficient (DSC), sensitivity (SE), and specificity (SP) were adopted as

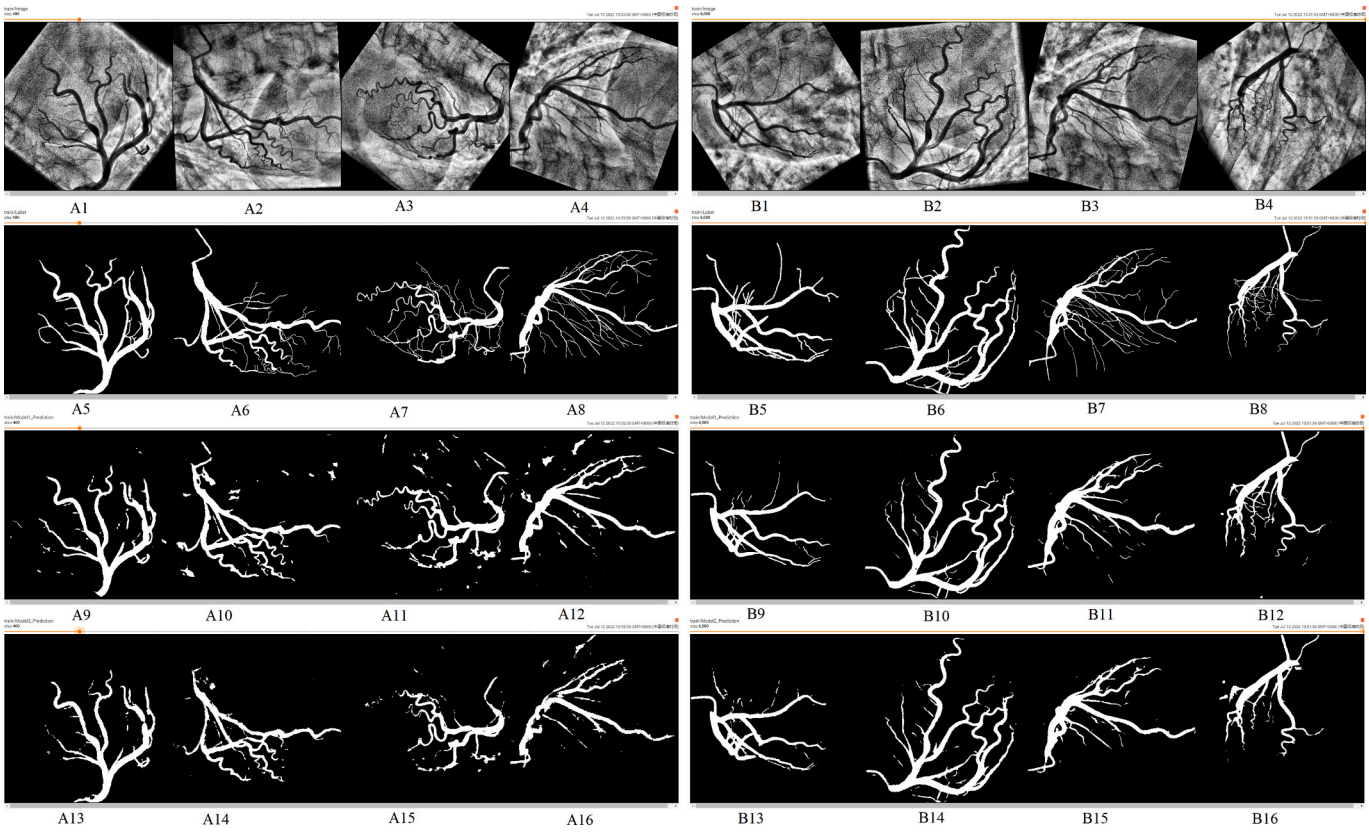


Fig. 6. Representative predicted images during training. The left 16 images (A1-A16) are sampled in 400 iterations after training, and the right (B1-B16) are sampled in 6000 iterations after beginning. The 4 rows indicate transformed original image, label, predictions of model 1 & 2, respectively.

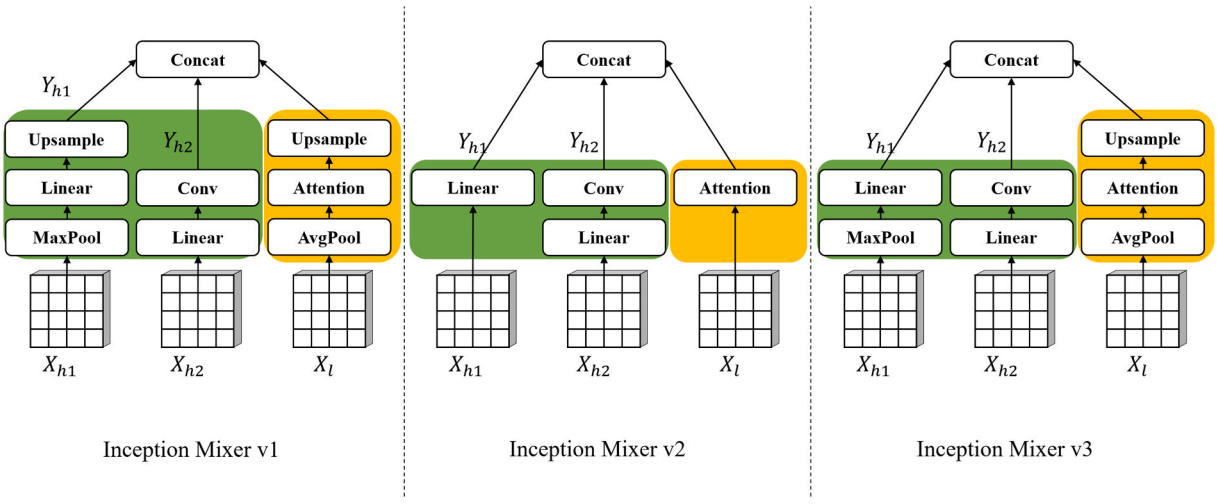


Fig. 7. Structures of inception mixer for ablation experiments.

the evaluation metrics. Among them, DSC reflects the overlap between ground truth (Y) and prediction results (X); SE indicates the proportion of true positive (TP) pixel points in all positive cases (TP + FN); and SP denotes the proportion of true negative (TN) pixel points in all negative cases (FP + TN). The evaluation metrics are formulized as follows.

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} \tag{28}$$

$$SE = \frac{TP}{TP + FN} \tag{29}$$

$$SP = \frac{TN}{FP + TN} \tag{30}$$

3.3. Ablation experiments

In this study, we proposed a new segmentation network ISUnet combined with CNN and SW-MSA. To verify the validity of its structure, some experiments were performed to test the effectiveness of its components and parameters, including the max pooling structure (*i.e.*, v1, v2, v3), the layer number (*i.e.*, $S = 3, 4, 5$), and the base learning rate (lr) and ramp up maximum iteration (*i.e.*, t_{max}). As the key to our study,

Table 2
Comparison of the performance using different structures v1, v2, and v3. The number of samples used for FSL are 20 and 100, respectively. The top two best results of each metric are indicated in bold.

Method	left-coronary-artery (LCA)			right-coronary-artery (RCA)			Params
	DSC(%)↑	SE(%)↑	SP(%)↑	DSC(%)↑	SE(%)↑	SP(%)↑	
ISUnetv1_20	73.68	76.34	97.83	79.48	76.44	**99.22**	9.91 M
ISUnetv1_100	74.01	72.41	**98.37**	**81.19**	**80.47**	99.11	
ISUnetv2_20	74.29	**77.58**	97.79	79.74	76.35	**99.26**	9.92 M
ISUnetv2_100	**75.19**	72.71	**98.54**	79.87	78.73	99.09	
ISUnetv3_20	72.27	**77.4**	97.42	77.41	75.7	99.01	9.91 M
ISUnetv3_100	**74.92**	73.96	98.35	**80.24**	**80.54**	98.99	

Table 3
Comparison of the performance using the different numbers of layers *S* from 2 to 5.20 and 100 denote the number of samples used for FSL, respectively. The top two best results of each metric are indicated in bold.

Method	LCA			RCA			Params
	DSC (%)↑	SE (%)↑	SP (%)↑	DSC (%)↑	SE (%)↑	SP (%)↑	
S = 2_20	72.04	76.96	96.54	77.02	74.50	98.85	1.55 M
S = 2_100	73.15	71.25	97.65	79.12	77.95	98.11	
S = 3_20	72.27	**77.40**	97.42	77.41	75.70	**99.01**	9.91 M
S = 3_100	**74.92**	73.96	**98.35**	**80.24**	**80.54**	98.99	
S = 4_20	74.01	**77.57**	97.78	78.43	76.35	**99.25**	35.96 M
S = 4_100	**75.21**	72.71	**98.54**	**81.66**	**78.42**	98.98	
S = 5_20	72.07	76.85	96.74	77.06	74.49	98.88	139.48
S = 5_100	73.21	71.65	97.85	79.17	78.03	98.22	M

Table 4
Comparison of the performance using different base learning rate *lr* from $\{10^{-2}, 10^{-3}, 10^{-4}\}$ and t_{max} from {40, 200, 300}. The number of samples used as labeled data and unlabeled data are 20 and 80, respectively. The top two best results of each metric are indicated in bold.

Method	LCA			RCA		
	DSC (%)↑	SE (%)↑	SP (%)↑	DSC (%)↑	SE (%)↑	SP (%)↑
UAMT_lr1e-2_tm40	71.19	72.74	96.92	61.13	60.29	98.87
UAMT_lr1e-2_tm200	70.00	72.70	96.74	68.70	61.02	99.12
UAMT_lr1e-2_tm300	68.05	72.61	97.32	72.22	63.91	99.21
UAMT_lr1e-3_tm40	**74.54**	**72.98**	**98.32**	**75.92**	**67.51**	99.24
UAMT_lr1e-3_tm200	**74.31**	**72.94**	**98.35**	**76.51**	**67.60**	**99.57**
UAMT_lr1e-3_tm300	74.19	72.88	98.28	75.79	67.23	99.23
UAMT_lr1e-4_tm40	73.26	71.53	98.04	75.29	66.54	99.40
UAMT_lr1e-4_tm200	72.95	71.19	98.09	74.99	65.46	99.34
UAMT_lr1e-4_tm300	73.26	71.31	98.01	74.46	63.89	**99.48**

Note: lr1e-2, lr1e-3, and lr1e-4 mean that the base learning rate are $10^{-2}, 10^{-3}, 10^{-4}$.

additional two SSL strategies (*i.e.*, PL and CL) are also performed along with stepwise ablation experiments.

3.3.1. Components & parameters about ISUnet

The Inception Mixer plays an important role in the ISUnet. Three different structures are proposed to get the best performance: the first one consists of a maximum pooling to downsample and an upsampling after a linear operation in the h1 channel; the second one eliminates the pooling and upsampling operations of the h1 and l channels, which can improve the resolution but increase computational effort; and the third one adopts maximum pooling with constant resolution in the h1 channel so that no up-sampling module is added after the linear transformation. The related experiments were performed as follows. Among them, Inception Mixer structures are shown in Fig. 7. The FSL results of different structures using 20 or 100 pairs of data are shown in Table 2. By comparing the performance of different structures, the third structure shows a balance in both between the left and right coronary artery angiographic images, and in between the performance and the number of parameters. So, the third structure v3 are adopted in the subsequent experiments.

There are a lot of parameters in the ISUnet because of the embedded modules of CNN and Transformer. In order to trade off the performance and the number of network parameters, an ablation experiment about the number of layers (S from 2 to 5) was conducted. When $S = 2$, the down- and up-sample are used only once, and when $S = 3$, they are used twice, and so forth. The results are shown in Table 3. By comparing the performance and parameters of the different S, the better results came from $S = 3$. Its reason lies in: the smaller the number of downsampling, the less high-dimensional information is extracted; while the more the number of downsampling, the lower the resolution of the feature map. The network performance is not very good when $S = 2$ and $S = 5$. In fact, $S = 3$ and $S = 4$ are comparable, if desired with fewer network parameters, $S = 3$.was adopted in the subsequent experiments.

In the SSL training, it is important to compute the consistent weight λ by using the learning rate (lr) of the optimizer and the ramp-up maximum iteration (t_{max}). Therefore, some relevant experiments are performed with the controlled variable method. The choice of the values lr and t_{max} is referred to the UAMT [35], some adjustments were made, e.g., replace Unet with ISUnet as the backbone network. According to Table 4, the best performance of the network training is achieved when $lr = 10^{-3}$ and $t_{max} = 200$. The results indicate that: on one hand, compared with the convolutional network, the ISUnet containing the SW-MSA converges more slowly, and the base learning rate of ISUnet

Table 5
Comparison of the performance using different SSL strategies (four cases, *i.e.*, none, use PL, use CL, use both). PL means pyramid consistency learning and CL means confidence learning. The top two best results are in bold.

Method	LCA			RCA		
	DSC (%)↑	SE (%)↑	SP (%)↑	DSC (%)↑	SE (%)↑	SP (%)↑
Vnet_ISUnet	74.58	73.00	98.40	78.30	72.76	99.36
Vnet_ISUnet_PL	75.65	74.24	**98.44**	80.11	74.90	99.41
Vnet_ISUnet_CL	74.86	74.56	98.24	79.01	73.00	**99.43**
Vnet_ISUnet_PLCL	**76.70**	**76.63**	98.35	**80.99**	**76.69**	99.38
Unet_ISUnet	75.11	75.24	98.21	78.84	73.40	99.38
Unet_ISUnet_PL	76.22	**77.55**	98.17	80.68	76.41	99.36
Unet_ISUnet_CL	75.41	76.10	98.10	78.90	72.75	**99.42**
Unet_ISUnet_PLCL	**76.71**	75.51	**98.50**	**81.48**	**77.81**	99.35

Table 6
Comparisons of the proposed method with other methods including three supervised learning methods (same labeled data 20, or same total data 100), and two SSL methods (single same networks as student and teacher like UAMT, and two different networks as teachers cross teaching like Vnet_SwinU).

Method		LCA			RCA		
		DSC(%)↑	SE(%)↑	SP(%)↑	DSC(%)↑	SE(%)↑	SP(%)↑
supervised	Vnet_20	72.51	74.92	97.75	78.04	75.65	99.10
	Vnet_100	75.87	73.10	**98.60**	81.37	**81.70**	99.04
	Unet_20	70.46	**78.05**	96.92	78.96	78.03	99.03
	Unet_100	**76.48**	75.17	98.48	**82.15**	**81.23**	99.17
	SwinUnet_20	69.49	72.69	97.41	70.87	63.72	99.21
	SwinUnet_100	71.40	69.16	98.27	77.17	78.18	98.78
	ISUnet_20	72.27	77.40	97.42	77.41	75.70	99.01
	ISUnet_100	74.92	73.96	98.35	80.24	80.54	98.99
semi-supervised	UAMT_Vnet_20/80	73.81	75.80	97.88	78.72	74.25	99.30
	UAMT_Unet_20/80	73.78	**77.90**	97.62	79.81	77.30	99.19
	UAMT_Swinunet_20/80	71.82	71.20	98.09	73.24	63.89	**99.49**
	UAMT_ISU_20/80	74.75	74.27	98.27	77.20	70.12	99.44
	Vnet_SwinU_20/80	74.61	74.33	98.24	78.09	72.12	99.39
	Unet_SwinU_20/80	75.66	75.79	98.26	79.20	73.05	**99.45**
	Vnet_ISU_20/80	74.58	73.00	98.40	78.30	72.76	99.36
	Unet_ISU_20/80	75.65	74.24	98.44	80.11	74.90	99.41
	Vnet_ISU_PLCL_20/80(ours)	75.04	76.10	98.10	78.70	72.75	99.42
	Unet_ISU_PLCL_20/80(ours)	**76.71**	75.51	**98.50**	**81.48**	77.81	99.35

should be set ten times smaller than it in the convolution at the beginning of the training. On the other hand, a larger t_{max} make smoother change of the weight of consistency loss. Finally, $lr = 10^{-3}$ and $t_{max} = 200$ are adopted as the uniform parameters for the next experiments.

3.3.2. Two semi-supervised training strategies

To verify the superiority of two proposed strategies (PL and CL), two pairs of networks were used for mutual teaching (Vnet & ISUnet, or Unet & ISUnet, respectively). First, neither PL nor CL was used, and a pure SSL approach was employed. It included guidance by labeled data and mutual guidance by pseudo labels. Next, we added the filtered pseudo label loss function after pyramid consistency correction or confident learning correction for the experiments, respectively. Finally, both consistency correction strategies were employed in the experiments. The results of these experiments are shown in Table 5. By comparison, no matter which strategy we added, the performance of the model was improved. The final SSL method with both correction strategies had a large performance improvement.

In addition to the ablation experiments of the above-mentioned cases with respect to structures or parameters, the weight parameters β_{pl} and β_{cl} also need to be considered during the loss calculation, the best results corresponding to $\beta_{pl} = 0.5$ and $\beta_{cl} = 5$ are adopted, which had been tested by ablation experiments in Refs. [30,34].

3.4. Comparison experiments

To further test the reliability of our approach, we compared our model with different methods through quantitative experiments. Current three FSL segmentation networks named Unet [7], Vnet [37] and SwinUnet [24] were used for the comparison of the coronary DSA segmentation. The following two novel semi-supervised methods are also compared with our method. (1) Uncertainty-aware mean teacher UAMT [35] (a method using two same networks as the student and teacher, which extends the mean teacher [38] method by uncertainty-aware consistency). (2) CNN_SwinUnet [39] , *i.e.*, an approach used two kinds of networks as teachers to guide each other, which improves other methods using a single convolutional or Transformer network with pseudo-label consistency. For the FSL, we performed two experiments with the labeled data at different scales: one has the same number of labeled data as the SSL (*i.e.*, 20 pairs of labeled data), and the other has the same total data as the SSL (*i.e.*, 100 pairs of labeled data). For the SSL, only 20 pairs of original and ground truth images were used as labeled data, and the remaining 80 images without labels were used as unlabeled data. The results of the testing data are shown in Table 6, Figs. 8 and 9.

The table and figures showed the exciting results. In general, in the FSL experiments, the test results using the models pretrained with more labeled data from the same network outperformed those with fewer labels; among the four networks (*i.e.*, Vnet, Unet, SwinUnet, ISUnet), SwinUnet had the worst effect, and the difference between the effects of Vnet, Unet, and ISUnet was not large. In the SSL experiment, it is roughly divided into the same network guidance and different networks guiding each other, the latter is better than the former, and the idea is consistent with ISUnet (*i.e.*, different operations can extract different effective features).

The specific data are compared as follows. Firstly, compared with the semi-supervised single network approach (e.g., UAMT_Vnet_20/80) on the LCA and RCA segmentation, our method has about 3% of performance improvement on DSC, respectively, thanks to the dual network structure where different networks are able to extract high-dimensional features respectively. Secondly, compared with the semi-supervised dual network approach (e.g., Vnet_SwinU_20/80), the proposed two pseudo-label correction strategies (PL & CL) have about 1–3% of improvement w.r.t DSC. Thirdly, for the FSL using the same amount of labeled data, SwinUnet is the worst model because Transformer paid more attention in the whole receptive field, and Unet, Vnet, and ISUnet is comparable in ±1%. Excitingly, our semi-supervised method significantly outperforms them, due to the fact that the SSL method introduces weakly supervised information of pseudo-labeling. Last but not least, compared with the FSL methods using the same total amount of labeled data, our SSL method is comparable to them. In other words, our approach has been able to achieve the performance of FSL only using fewer labeled images, which is thanks to two pseudo-label correction strategies.

Besides the numerical comparisons, two figures present the visual results. As is shown in Fig. 8, among all the methods, the FSL method of Unet achieved the best performance by using the labels of the entire training data, and the SwinUnet hasn't achieved idiot performance. By comparing the task of the SwinUnet paper with ours, the former is a block multi-organ segmentation and the latter is a tubular object segmentation, which have different shapes. The global feature capture capability of Transformer is a bit underpowered in the feature extraction of slender vessels. However, our proposed SSL method in the last column of boxplots used a smaller percentage of labels, achieved the same performance in comparison with the FSL method using the all-labeled data. Besides, what we can see is that the prediction of RCA is slightly

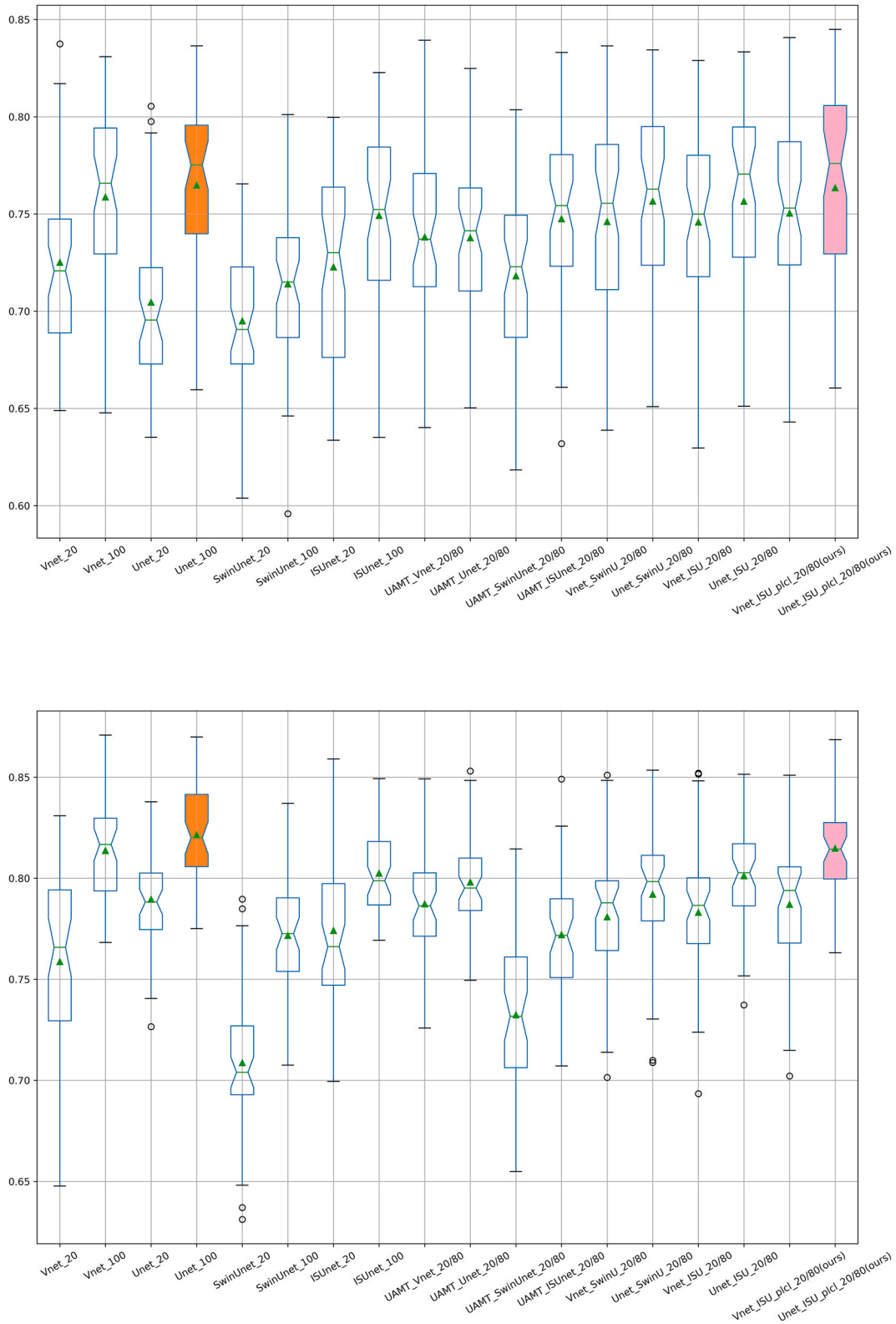


Fig. 8. Box plots of DSC results using different methods. Top: the results for the LCA. Bottom: the results for the RCA. The center line indicates the median. The box indicates the interquartile range. The upper and lower bars indicate the maximum and minimum values. ▲ indicates the mean and indicates the outlier. Top two are colored.

better than LCA and has a lower degree of dispersion. That's because of the less complex structure of the RCA. As depict in Fig. 9, with the addition of our two SSL strategies (PL & CL) separately and simultaneously, the resultant segmentation noise gradually decreases and the vascular continuity gradually becomes better, especially in the yellow and green boxed parts.

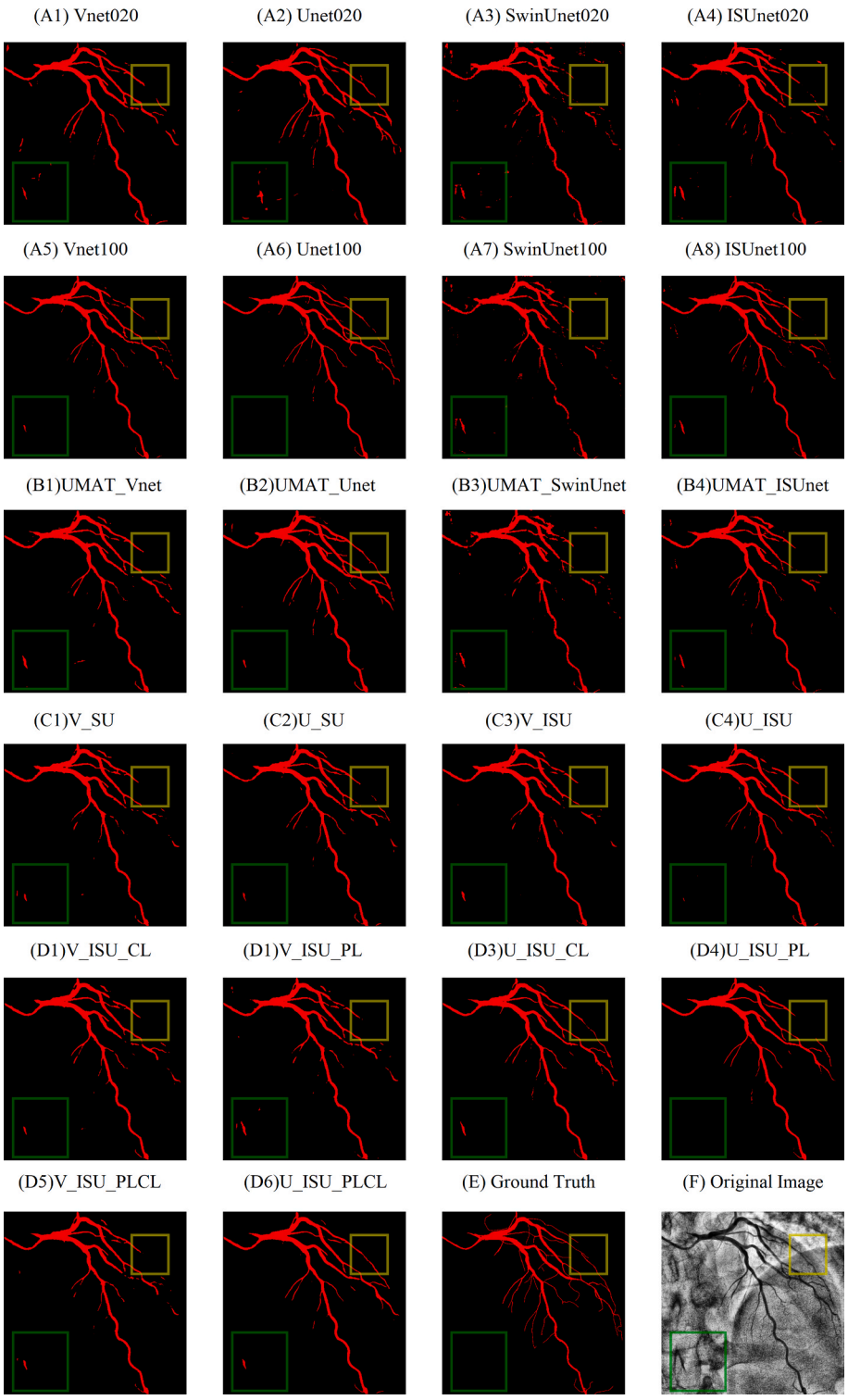


Fig. 9. The segmentation results of the test set by using different methods. (F) and (E) are the original image and the manually marked image, respectively. The 1st and 2nd rows (A1∼A8) are the results of FSL methods relying on 20 and 100 pairs of labeled data, respectively. The 3rd and 4th rows are the results of SSL methods relying on two same networks (B1∼B4) or two different networks (C1∼C4). The 5th (D1∼D4) row shows the ablation experiment adding CL or PL strategies to the dual network SSL method, respectively. The first two in the 6th row (D5 & D6) shows the results of the SSL methods combining both CL and PL strategies using Vnet and ISUnet or Unet and ISUnet. For comparison, in all semi-supervised methods, we used 20 pairs of labeled data and 80 unlabeled images. The differences within the green and yellow boxes showed the resultant advantages such as better vascular continuity and less noise respectively by our method.

4. Discussion

This work presented an SSL approach and showed its application in the DSA data using a little percentage of labeled images. Now, we will discuss the choice about the overall framework, rectifying strategies, solutions for the gradient disappearance, and the ratios of labeled data, where the scalability and novelty are the science ideology of our work.

The Mean Teacher [38] model is the dominant framework in the SSL. However, its limitation is obvious because of using the same network as teacher and student simultaneously. The extracted features are more homogeneous, which is not friendly for the small labeled data. This leads to vascular discontinuity and many noises in the result background region. To solve this problem, we not only adopted a structure combining CNN and Transformer in the overall framework, but also proposed a novel network that takes different channel operations for features extraction. Experiments show that the overall architecture is superior to the network with single feature extraction.

About the strategies for SSL, the most common ideas are pseudo-

Table 7
Comparison of the segmentation results of the test data using models trained by different ratios of labeled data.

Method	Label ratios	LCA			RCA		
		DSC(%)↑	SE(%)↑	SP(%)↑	DSC(%)↑	SE(%)↑	SP(%)↑
Vnet_ISUnet_PLCL	10%	70.66	63.22	**98.90**	76.57	67.58	**99.57**
Vnet_ISUnet_PLCL	20%	75.04	**76.10**	98.10	78.70	72.75	99.42
Vnet_ISUnet_PLCL	30%	**75.30**	74.55	98.34	**79.66**	**75.20**	99.34
Unet_ISUnet_PLCL	10%	75.29	72.42	**98.60**	80.49	75.39	99.41
Unet_ISUnet_PLCL	20%	76.71	75.51	98.50	81.48	77.81	99.35
Unet_ISUnet_PLCL	30%	**76.75**	**77.12**	98.30	**81.72**	**77.96**	**99.48**

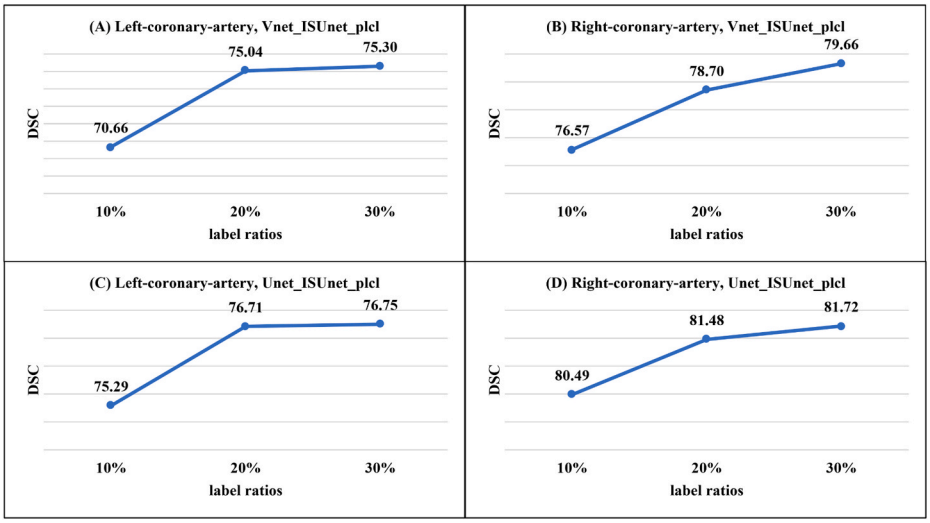


Fig. 10. Test results by training with different ratios of labeled data, where subfigures (A∼D) are corresponding to LCA and RCA respectively with Vnet_ISUnet_plcl and Unet_ISUnet_plcl.

label-based and consistency-based learning. Based on the ideas, we newly proposed two strategies for pseudo label correction and consistency loss calculation. The first one is to calculate the uncertainty penalty term by KL divergence in between the average output of one network and the pseudo label generated by the other network. We then calculate the gap between the two above mentioned items with a penalty term through the MSE. KL divergence has a higher penalty for outliers than L2 and is more suitable as a correction loss function, which was approved in Ref. [30]. The second one is to calculate the uncertainty as the penalty weight term for confidence learning by repeating the input T times. By combining confidence learning and uncertainty suppression, the noise of pseudo label is reduced effectively, as depict in Fig. 9.

In addition to these, experiments show that premature loss computation of pseudo label frustrates gradient backpropagation. At the beginning of training, the pseudo label causes high error rate and misleading, for which we adopt the following two approaches to solve the problem. One is to implement ramp-up trade-off weights for the pseudo label loss function, which is a Gaussian distribution function depending on the number of manually designed maximum iterations. The other is to defer the loss calculation of PL and CL, *i.e.*, computation is performed after a certain number of iterations.

To investigate the efficiency of the proposed SSL method for data utilization, we conducted labeled data ratios experiments, where the same architecture was used and the proportion of labeled data was changed. As shown in Table 7 and Fig. 10, when the ratio of labeled data increases, the performance of our model is improved with first faster and then slower, which is accord with common sense. Thus, it can be seen that our method can use a smaller ratio of labels and get a better segmentation result. It further demonstrates the value to reduce the reliance on manual labeling.

The scalability and novelty of proposed method can be summarized as follows. First, the proposed deep learning networks allows various operations in different channels, such as pooling, activation, dilated convolution, sliding window attention, and other functions. Second, the dual different networks in this study interact with each other during semi-supervised learning, and benefit the multiple networks parameter sharing, joint training, and co-polling. Furthermore, the semi-supervised learning strategies in the paper, such as pyramid consistency and confidence learning, can improve image filtering and noise removal. The novelty of this study is the proposed deep learning framework innovates the vessel segmentation using small amount of labeled training data and incorporating the semi-supervised strategy.

There are several limitations of this study. First, 2D data instead of 3D imaging data was used in our research. The continuity information of vessels in Z direction is not accounted for in the segmentation, possibly resulting in mistaken vessel assignment. Second, the DSA images used in this study were from a single center, for which overfitting is inevitable with the pretrained model. Future work with 3D data across multiple centers and facilities are expected to further improve the generalization of the proposed semi-supervised learning model with few labeled data.

5. Conclusion

In this paper, we address the problem of coronary angiographic image segmentation for small amount of labeled DSA data. For this purpose, a new segmentation network ISUnet combining the advantage of CNN and SW-MSA was proposed allowing various operations in different channels. The proposed semi-supervised segmentation framework outperforms a single CNN or Transformer. The advantages of the framework are due to: the proposed new network, the dual different networks guiding each other and the use of two semi-supervised optimization strategies. Nevertheless, the limitations of ignoring 3D

information and using only single-center 2D data make the generalization performance of the model worthy of examination, and semi-supervised vessel segmentation of multicenter 3D medical images will be the next development direction of this research.

Funding

This work was supported in part by the National Key Research and Development Program of China (Grant No. 2018YFA0704102), in part by the National Natural Science Foundation of China (Grant No. 81827805), in part by Natural Science Foundation of Guangdong Province (Grant No. 2023A1515010673), and in part by Shenzhen Technology Innovation Commission (Grants No. JCYJ20200109114610201, JCYJ20200109114812361, JSGG20220831110400001, and KCXFZ20201221173202007), and in part by the Shenzhen Engineering Laboratory for Diagnosis & Treatment Key Technologies of Interventional Surgical Robots, and in part by the Discipline Construction Project of Guangdong Medical University with Grant No. 4SG21017G.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Mackay, G.A. Mensah, K. Greenlund, The Atlas of Heart Disease and Stroke, World Health Organization, 2004.
[2] I.I. Abubakar, T. Tillmann, A. Banerjee, Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013, Lancet 385 (9963) (2015) 117–171.
[3] W.G. Members, D. Mozaffarian, E.J. Benjamin, et al., Executive summary: heart disease and stroke statistics–2016 update: a report from the American heart association, Circulation 133 (4) (2016) 447–454.
[4] Antiplatelet Trialists' Collaboration, Collaborative overview of randomised trials of antiplatelet therapy Prevention of death, myocardial infarction, and stroke by prolonged antiplatelet therapy in various categories of patients, BMJ 308 (6921) (1994) 81–106.
[5] S. Masood, M. Sharif, A. Masood, et al., A survey on medical image segmentation, Curr. Med. Imag. 11 (1) (2015) 3–14.
[6] G. Wang, M.A. Zuluaga, W. Li, et al., DeepIGeoS: a deep interactive geodesic framework for medical image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 41 (7) (2018) 1559–1572.
[7] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2015, pp. 234–241.
[8] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., An Image Is Worth 16x16 Words: Transformers for Image Recognition at scale[J], 2020 arXiv preprint arXiv: 2010.11929.
[10] C. Si, W. Yu, P. Zhou, et al., Inception Transformer, 2022 arXiv preprint arXiv: 2205.12956.
[11] R. Gharleghi, N. Chen, A. Sowmya, et al., Towards automated coronary artery segmentation: a systematic review, Comput. Methods Progr. Biomed. (2022), 107015.
[12] O. El Ogri, H. Karmouni, M. Sayyouri, et al., 3D image recognition using new set of fractional-order Legendre moments and deep neural networks, Signal Process. Image Commun. 98 (2021), 116410.
[13] H. Karmouni, M. Sayyouri, H. Qjidaa, A novel image encryption method based on fractional discrete Meixner moments, Opt Laser. Eng. 137 (2021), 106346.
[14] J. Long, E. Shelhamer, T. Darrell, Fully Convolutional Networks for Semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[15] D.H. Lee, Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks[C]//Workshop on challenges in representation learning, ICML 3 (2) (2013) 896.
[16] A. Tarvainen, H. Valpola, Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results, Adv. Neural Inf. Process. Syst. (2017) 30.
[17] S.P. Liu, J.M. Hong, J.P. Liang, X.P. Jia, J. Ouyang, J. Yin, Medical image segmentation using semi-supervised conditional generative adversarial nets, Ruan Jian Xue Bao/Journal of Software 31 (8) (2020) 2588–2602 (in Chinese).
[18] X. Cao, H. Chen, Y. Li, et al., Uncertainty aware temporal-ensembling model for semi-supervised abus mass segmentation, IEEE Trans. Med. Imag. 40 (1) (2020) 431–443.
[19] A. Mehrtash, W.M. Wells, C.M. Tempany, et al., Confidence calibration and predictive uncertainty estimation for deep medical image segmentation, IEEE Trans. Med. Imag. 39 (12) (2020) 3868–3878.
[20] X. Li, L. Yu, H. Chen, et al., Transformation-consistent self-ensembling model for semisupervised medical image segmentation, IEEE Transact. Neural Networks Learn. Syst. 32 (2) (2020) 523–534.
[21] X. Xu, T. Sanford, B. Turkbey, et al., Shadow-consistent semi-supervised learning for prostate ultrasound segmentation, IEEE Trans. Med. Imag. 41 (6) (2021) 1331–1345.
[22] Y. Wang, X. Wei, F. Liu, et al., Deep distance transform for tubular structure segmentation in ct scans, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3833–3842.
[23] S. Li, C. Zhang, X. He, Shape-aware Semi-supervised 3D Semantic Segmentation for Medical images[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2020, pp. 552–561.
[24] H. Cao, Y. Wang, J. Chen, et al., Swin-unet: Unet-like Pure Transformer for Medical Image Segmentation, 2021 arXiv preprint arXiv:2105.05537.
[25] C.Y. Lee, S. Xie, P. Gallagher, et al., Deeply-supervised nets[C]//Artificial intelligence and statistics, PMLR (2015) 562–570.
[26] L. Wang, C.Y. Lee, Z. Tu, et al., Training Deeper Convolutional Networks with Deep Supervision, 2015 arXiv preprint arXiv:1505.02496.
[27] C. Northcutt, L. Jiang, I. Chuang, Confident learning: estimating uncertainty in dataset labels, J. Artif. Intell. Res. 70 (2021) 1373–1411.
[28] S. Azadifar, M. Rostami, K. Berahmand, et al., Graph-based relevancy-redundancy gene selection method for cancer diagnosis[J], Comput. Biol. Med. 147 (2022), 105766.
[29] F. Saberi-Movahed, M. Rostami, K. Berahmand, et al., Dual regularized unsupervised feature selection based on matrix factorization and minimum redundancy with application in gene selection, Knowl. Base Syst. 256 (2022), 109884.
[30] X. Luo, G. Wang, W. Liao, et al., Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency, Med. Image Anal. 80 (2022), 102517.
[31] D. Angluin, P. Laird, Learning from noisy examples, Mach. Learn. 2 (4) (1988) 343–370.
[32] C. Elkan, The foundations of cost-sensitive learning[C]//International joint conference on artificial intelligence, 1 17, Lawrence Erlbaum Associates Ltd, 2001, pp. 973–978.
[33] S. Laine, T. Aila, Temporal Ensembling for Semi-supervised Learning, 2016 arXiv preprint arXiv:1610.02242.
[34] Z. Xu, D. Lu, J. Luo, et al., Anti-interference from Noisy Labels: Mean-Teacher-Assisted Confident Learning for Medical Image Segmentation, IEEE Transactions on Medical Imaging, 2022.
[35] L. Yu, S. Wang, X. Li, et al., Uncertainty-aware Self-Ensembling Model for Semi-supervised 3D Left Atrium segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2019, pp. 605–613.
[36] S.M. Pizer, E.P. Amburn, J.D. Austin, et al., Adaptive histogram equalization and its variations, Comput. Vis. Graph Image Process 39 (3) (1987) 355–368.
[37] F. Milletari, N. Navab, S.A. Ahmadi, V-net: Fully Convolutional Neural Networks for Volumetric Medical Image segmentation[C]//2016 Fourth International Conference on 3D Vision (3DV), IEEE, 2016, pp. 565–571.
[38] A. Tarvainen, H. Valpola, Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results, Adv. Neural Inf. Process. Syst. (2017) 30.
[39] X. Luo, M. Hu, T. Song, et al., Semi-Supervised Medical Image Segmentation via Cross Teaching between CNN and Transformer, 2021 arXiv preprint arXiv: 2112.04894.

Abbreviations (sorted by order of appearance)

DSA: digital subtraction angiography
ISUnet: Inception-SwinUnet
FSL: fully supervised learning
SSL: semi-supervised learning
CL: confidence learning
PL: pyramid-consistency learning
CAD: coronary artery disease
CNN: convolutional neural network
NLP: natural language processing
CV: computer vision
MSA: multi-head self-attention
iFormer: Inception Transformer
PBM: probability map
IST: Inception Swin Transformer
ITM: Inception Token Mixer
FFN: feedforward network
LN: Layer Normalization
W-MSA: window multi-head self-attention
SW-MSA: shifted window multi-head self-attention
MSE: mean square error
CNP: classification noise process
PBC: prune-by-class
AHE: Adaptive histogram equalization
CLAHE: Contrast Limited Adaptive Histogram Equalization
CDF: cumulative distribution function

DSC: dice score coefficient
SE: sensitivity
SP: specificity
TP: true positive
TN: true negative
FN: false negative
FP: false positive
lr: learning rate
LCA: left-coronary-artery
RCA: right-coronary-artery