SemiCTrans: Semi-Supervised Medical Image Segmentation Framework combining CNN and Transformer

Fei Xu, Zihan Li, Qingqi Hong, Qingqiang Wu, Qingde Li, and Jie Tian, *Fellow, IEEE*

Abstract—In medical image segmentation, both local details and global information are important. CNN has demonstrated good detail capture ability in many visual tasks, but due to the limitation of the receptive field, its ability to capture global information is insufficient. Transformer is just the opposite, which can model long-range dependency well, but it is insufficient for capturing local details. Therefore, we construct a dual-branch medical image segmentation model based on CNN and Transformer to take advantage of their complementary strengths. Furthermore, for the scarcity of labeled data in medical images, we incorporate semi-supervised learning to further improve the accuracy and robustness of the model by leveraging the information of a large amount of unlabeled data. Experiments on two datasets show that our model can achieve excellent results in both fully-supervised and semi-supervised configurations.

Index Terms—Semi-Supervised, Medical Image Segmentation, CNN, Transformer, Heart Segmentation.

1 Introduction

Medical image segmentation is a type of semantic segmentation, which aims to segment various organs, tissues, or lesions contained in medical images. Accurate segmentation can be of great help in disease diagnosis and research. However, medical image segmentation faces two major challenges. Firstly, the shape of organs or tissues is often irregular and varies greatly in different samples. Therefore, it is difficult to accurately segment them, and it is easy to have inaccurate boundaries or to miss some small objects. Secondly, as the labeling of medical images requires expertise and a lot of time, it is difficult to provide a large number of labeled samples for model training like natural images. These challenges restrict the improvement of model performance, and the network model is prone to overfitting.

For the first problem, we intend to combine the great local feature extraction ability of convolutional neural networks (CNN) and the strong long-range dependency modeling ability of Transformer to accurately capture the boundary details and global context of the image. As a traditional image segmentation framework, CNN has shown its strong segmentation ability and detail capture ability on a large number of medical image segmentation tasks, such as skin lesion segmentation[1], carotid artery segmentation[2], bladder segmentation[3], etc.. With end-to-end network training, it is shown to be powerful in building hierarchical feature representations. However, despite the great success of CNN-based methods, CNN still falls short in capturing global contextual information. Specifically, CNN-based methods generally expand the receptive field by performing continuous downsampling to obtain global information. Multiple downsampling will lead to the loss of spatial location information and low-level features[4].

In contrast, Transformer[5] emerged in recent years has an innate global receptive field. It was first proposed in the field of natural language processing, and its good long-range dependency modeling ability has been widely used in various vision tasks, such as classification, segmentation, object detection, etc.[6][7][8][9]. Transformer treats image segmentation as a sequence-to-sequence prediction task, which computes the patch-to-patch relationship by segmenting the image into disjoint patches. This allows the global context to be modeled at the beginning without the need to downsample the image. But this operation also has problems in image segmentation. Due to computational overhead, it is not suitable to set the patch size too small (the patch size in classic Vision Transformer[6] is 16×16). This makes it easy for Transformer to miss some small targets, that is, it cannot capture local details very accurately. However, in medical image segmentation tasks, the segmentation of small objects is very common, such as nucleus segmentation, retina segmentation, etc., which requires the model to have good local detail capture ability.

Given the complementary characteristics of CNN and Transformer, we hope to combine respective advantages of Transformer and CNN to deal with the problems of variable target shape and blurred boundaries in medical image segmentation. We construct a dual-branch model called SemiCTrans, which contains a CNN branch and a Transformer branch. The CNN branch is used to capture spatial location details and ensure the segmentation accuracy of boundaries. The Transformer branch is used to extract the global context to adequately identify the overall shape of the target.

For the second problem, we hope to combine semi-

- *Fei Xu, Qingqi Hong, and Qingqiang Wu are with Xiamen University, Xiamen 361005, China, e-mail:hongqq@xmu.edu.cn.*
- *Zihan Li is with Xiamen University and the Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA, e-mail: lizihanlizh@foxmail.com.*
- *Qingde Li is with the Department of Computer Science, University of Hull, Hull, HU6 7RX, UK, e-mail:q.li@hull,ac.uk.*
- *Jie Tian is with the Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China, e-mail:tian@ieee.org.*

supervised learning to exploit information from a large number of unlabeled medical images through pseudo labels and consistency constraints. It can further improve the segmentation accuracy while alleviating the scarcity of labeled data in medical image segmentation. In a large number of semi-supervised and fully-supervised comparative experiments, it has been proved that when the amount of available labeled data is small, the performance and generalization of the model can be effectively improved with a large amount of unlabeled data, which is suitable for medical image segmentation where image labeling is expensive[10][11][12][13]. Our dual-branch model is suitable for consistency constraints. By imposing consistency constraints on the prediction results of the two branches, not only can the training set be expanded with unlabeled data, but also the two branches can learn from each other, enhancing the final fusion results while improving the prediction results of the two branches.

In summary, the main contributions of this paper are as follows:

- We propose a dual-branch framework based on CNN and Transformer, which utilizes the great local detail extraction ability of CNN and the powerful long-range context modeling ability of Transformer to accurately segment medical images.
- We improve the Transformer branch by constructing an encoder-decoder structure, which obtains feature maps of different scales in the encoder and utilizes dense connections in the decoder to aggregate different scale feature maps to obtain more accurate segmentation results.
- We combine semi-supervised learning to exploit information from a large number of unlabeled medical images through pseudo-labels and consistency constraints. At the same time, deep supervision and deep consistency supervision are used to further improve the segmentation accuracy.
- Experimental results show that our semi-supervised model that using a small amount of labeled data can achieve accuracy close to that of fully supervised method.

2 Related work

2.1 Convolutional Neural Networks in Medical Image Segmentation

Convolutional neural networks have become the mainstream framework for medical image segmentation, and commonly used networks include fully connected network (FCN)[14], U-Net[15], V-Net[16], and their various variants. FCN is the first end-to-end fully convolutional network for pixel-level prediction, and its applications in medical image segmentation include cancer segmentation[17], brain segmentation[18], carotid artery segmentation[19], etc.. U-Net is a network specially proposed for medical image segmentation in 2015. It adopts an encoder-decoder structure shaped like 'U' and preserves details by adding skip connections between the corresponding layers of the encoder and decoder. Based on this approach, U-Net has achieved great success in a wide range of medical image segmentation tasks, such as rectal tumor segmentation[20], organ segmentation[21], bladder segmentation[22], etc. On top of the basic U-net architecture, researchers have also proposed several variants to suit different tasks in medical image segmentation. For example, V-Net[16] is the 3D version of U-Net. Milletari et al. extended the 2D U-Net to a 3D structure by replacing all 2D operations in the original U-Net with 3D operations. Zhang et al. added attention mechanism to U-Net to segment cervical cells[23]. Zhang et al. added Inception[24] structure to U-Net for brain tumor detection[25]. Baldeon-Calisto et al. added residual block[26] in U-Net for medical image analysis[27]. In addition to using a single convolutional neural network for medical image segmentation, some works combine multiple networks for segmentation. For example, Huang et al. combined VGG-16[28] and U-Net for automatic segmentation of median nerve in ultrasound images, they used VGG-16 as encoder and U-net as decoder[29].

Although the above methods have introduced various improvements based on convolutional neural networks, the core of these networks still uses convolution and pooling operations to extract features at different levels of abstraction. Convolutional networks are limited by the receptive field and cannot model the global context well. At present, many works have been devoted to increasing the receptive field in convolutional networks, including dilated convolution, attention mechanism, etc, and applying Transformer to medical image segmentation is one of them.

2.2 Transformer in Medical Image Segmentation

Transformer[5] was first proposed in the field of natural language processing and later applied to vision tasks. Its excellent performance in a large number of vision tasks including image classification[6] [7], object detection[8] [30], and semantic segmentation[31] [9] has attracted researchers to explore its application in the field of medical image segmentation. The application of Transformer in medical image segmentation can be roughly divided into two categories: segmentation model combining Transformer and CNN and segmentation model with pure Transformer structure.

Given the complementary characteristics of Transformer and CNN, many researchers try to combine Transformer or its key structure with convolution. For example, TransUNet[32] proposed by Chen et al. uses Transformer to take the serialized feature extracted by CNN as input to extract the global context. The features extracted by the encoder are then upsampled in the decoder and combined with high-resolution CNN features for precise localization. The UTNet[33] proposed by Gao et al. combines the advantages of convolution and self-attention, uses convolutional layers to extract local features, avoids large-scale pretraining of Transformer, and uses self-attention to obtain long-range dependencies. They follow the standard design of U-Net but replace the last convolutional layer at each resolution (except the highest resolution) with the proposed Transformer module. The TransBTS[34] constructed by Wang et al. extends the network structure to 3D to better capture volume information in MRI data. It first uses 3D CNN to generate compact feature maps to capture spatial and depth information, then uses the Transformer

encoder to model long-range dependencies in global space, and finally repeatedly stacks upsampling and convolutional layers to gradually generate high-resolution segmentation results.

In addition to the combination of Transformer and CNN, some researchers have begun to use the Transformer as the main body of the segmentation model. For example, Karimi et al. propose a convolution-free medical image segmentation network, which is based on self-attention between adjacent patches and obtains better segmentation results than CNN[35]. UNETR[36] proposed by Hatamizadeh et al. employs a pure Transformer as an encoder to learn sequential representations of input data and efficiently acquire global multi-scale information. The representations extracted from the encoder are merged with the features of the decoder through skip connections to predict segmentation results. Cao et al. use a pure Transformer to construct a U-Net-like network called Swin U-Net[37]. They use a hierarchical Swin Transformer[38] with sliding windows as an encoder to extract contextual features and implement local-to-global self-attention in the encoder. Then a symmetric Swin Transformer-based decoder with patch expansion layers is designed to perform upsampling operations to restore the spatial resolution of feature maps.

Although Transformer has been successfully applied in medical image segmentation, due to the setting of patch size, it is easy to miss the small targets, that is, it is not as good as the convolutional neural network in capturing local details. Therefore, in order to better preserve the spatial features and global context of medical images, we construct a hybrid model based on CNN and Transformer, which contains two branches in parallel to exploit the respective strengths of Transformer and CNN, and generate hierarchical feature representations.

2.3 Semi-Supervised Learning

Since the labeling of medical images requires professional knowledge and the labeling often takes a long time, there is a problem of scarcity of labeled data in medical image segmentation. Although labeled data is scarce, a large number of unlabeled medical images exist. Therefore, how to utilize a small number of labeled images and a large number of unlabeled images is the focus of research on medical image segmentation today. Semi-supervised learning is a way to solve this problems. According to the utilization of unlabeled data, semi-supervised learning can be roughly divided into consistency-based, pseudo-label-based, adversarial training-based, and so on. These methods are often used in combination. Our model incorporates the idea of pseudo-labels and consistency constraints.

Among them, consistency-based methods assume that the model's predictions are invariant when subjected to various perturbations or across different tasks. Perturbations include data-level perturbations, network-level perturbations, and so on. For example, Li et al. enhance the regularization of pixel-level predictions by introducing transformations such as rotation and flip into the self-ensemble model. They exploit unlabeled data by perturbing the original image at the data level while constraining the consistency of predictions after different perturbations[39]. Chen et al. simultaneously input images into two networks with the same structure but different initialization, which is a perturbation at the network level to encourage different initialization networks to make consistent predictions for the same input images[40]. Lei et al. believe that in semi-supervised learning, due to the lack of labeled data, the model will overrely on the context of the training data, resulting in its lack of generalization to unseen data. Therefore, they propose a directional contrastive loss, which aims to constrain the consistency of the same patch in different contexts, while requiring lower-quality features to align with corresponding high-quality features[41]. Luo et al. construct a dual-task deep network to simultaneously predict the segmentation results and level set representations of targets. The authors then convert the level set function into a segmentation probability map, so that the consistency loss of the two tasks can be calculated[42]. Xu et al. further propose a three-task framework, called AugSegNet, which utilizes saliency detection and multi-label image classification as auxiliary tasks to improve semantic segmentation using only image-level labels[43].

Pseudo label-based methods usually use labeled data to train the model first, and then use the pre-trained model to generate the prediction results of unlabeled images as pseudo labels. Usually, the generated pseudo labels are selected with confidence, and only those with high confidence are used to further fine tune the model. Wang et al. propose a 2.5D CNN to solve the problem of inconsistent receptive fields in different axial in 3D medical images, while adopting an iterative approach to improve the confidence of pseudo labels. The authors use the Monte Carlo Dropout method to obtain the uncertainty maps of the initial pseudo labels. Then adjust pseudo labels of uncertain voxels according to their uncertainty maps and context[44].

The main idea of adversarial training-based methods is to use two networks, one as a generator to synthesize artificial medical images, and the other as a discriminator to determine whether the input image is a real image or synthesized by the generator. The two networks are improved by playing against each other. For example, Qin et al. use 3D CNNs as the generator and the discriminator to build an adversarial generative network for pulmonary nodule segmentation[45].

In addition to semi-supervised learning, there are also some works to deal with the scarcity of labeled data through transfer learning[46], data augmentation[47], introducing prior knowledge[48] [49], weak supervision[50], self-supervision[51], etc..

3 METHODS

The overall architecture of our model is shown in Fig. 1, which consists of three parts: the CNN branch, the Transformer branch, and the fusion module. The CNN branch is responsible for extracting input spatial features, while the Transformer branch is used to extract global context. The fusion module is responsible for the deep fusion of different semantic features extracted by the two branches to combine the respective advantages of CNN and Transformer. Both the CNN branch and the Transformer branch follow the encoder-decoder structure. The encoder extracts multi-scale features from the image. The decoder aggregates feature at

different scales and recovers spatial information to generate predicted segmentation results. For labeled data, we use ground truths to supervise the prediction results of two branches and the fusion module. For unlabeled data, we use the fusion results as pseudo labels to supervise the prediction results of the two branches. In addition, for all data, we also use consistent supervision. By constraining the consistency of the two branches' predictions, the two branches can learn from each other. The specific introduction of each part of the framework is as follows.

3.1 CNN Branch

The encoder of the CNN branch adopts ResNet-34[26] as the backbone, which expands the receptive field through multiple downsampling, and obtains features with different levels of abstraction from local to global. The decoder gradually restores the spatial resolution of the image and reduces the number of channels by upsampling. At the same time, high-level features and low-level features are combined to preserve spatial information to the greatest extent. The structure of this branch is shown in Fig. 2. In the encoder, from Layer0 to Layer4, each Layer halves the width and height of the input feature map and doubles the number of channels. Among them, Layer1 to Layer4 use the same structure, as shown in Fig. 2(c), and the structure of Layer0 is shown in Fig. 2(b). The decoder starts upsampling from the top-level feature map, and Up4 to Up1 use the same structure, as shown in Fig. 2(d). In the process from Up4 to Up1, each Up operation first performs bilinear interpolation on the low-resolution feature map, doubling its width and height, and halving the number of channels at the same time. Then, the upsampled feature map and the feature map of the same scale in the encoder are concatenated on the channel, and the doubled number of channels is adjusted through a 1×1 convolution. This is to preserve the spatial information lost due to downsampling in the encoder as much as possible. Up0 first uses bilinear interpolation to upsample the spatial resolution of the output feature map back to the spatial resolution of the original input image. Then it uses a 1×1 convolution to adjust the number of channels to equal the number of categories to get the final prediction output. The multi-scale feature maps obtained at the decoder will be sent to the fusion module to be fused with the features from the Transformer branch.

3.2 Transformer Branch

Different from the classic Vision Transformer (ViT)[6] that only outputs single-scale feature maps, we refer to the idea of Segformer[52] and improve ViT so that it can output multi-scale feature maps. This is because high-level features are sufficiently discriminative in classification tasks. But in the segmentation task, if the low-resolution high-level features are directly upsampled back to the original resolution, a lot of details will be lost. Therefore, we let the Transformer branch output multi-scale features to better combine information of different abstraction levels. Specifically, we let the Transformer branch follow the encoder-decoder structure, as shown in Fig. 3(a). In the encoder, the size of the input image is gradually reduced by four Layers, instead of directly using a large patch size to quickly reduce the image size to obtain a low-resolution high-level feature map. Layer1 to Layer4 have the same structure. In each Layer, we first use a convolutional layer to segment the input image into multiple patches, this step is called Patch Embedding. We make the stride of the convolution smaller than the size of the convolution kernel. So that there are overlapping parts between the patches, which can better allow the information exchange between different patches. The divided patch will go through multiple stacked Transformer blocks. The composition of Transformer blocks follows the classic ViT architecture, as shown in Fig. 3(b). Each Transformer block consists of two Layer Normalization, a Multi-Head Self-Attention, two Dropouts, an MLP block, and two residual connections.

The Patch Embedding layer in Layer1 uses a convolution with a size of 7×7 and a stride of 4 to divide the image. After division, the size of the feature map is reduced to $1/4$ of the input. In the process of Layer2 to Layer4, the Patch Embedding part of each layer is a convolutional layer with a convolution kernel of 3×3 and a stride of 2. The size of the feature map after convolution becomes $1/2$ of the input. These Patch Embedding layers can be used to model global information and reduce the size of feature maps to obtain features at different levels of abstraction without downsampling. Here we use a small patch size for three reasons: First, in the segmentation task, using a large patch size, such as the classic 16×16, is easy to miss some small targets, while small patches can better capture details[9]. The second is to control the size of the feature map so that it can gradually generate multi-scale features instead of directly obtaining the highest-level abstract features. The third is to keep the Transformer's feature map in sync with the CNN's feature map, making it easier to combine them.

It should be noted that we feed the multi-scale feature maps produced by the decoder into the fusion module, not the feature maps produced by the encoder. The feature map output by each abstraction level is formed by the dense fusion of the feature map of the same level on the encoder and the feature maps of higher levels on the decoder, as shown in Eq.(1) and Eq.(2), where f_n represents the feature of the highest level. In this experiment, the value of i ranges from 1 to n, and $n = 4$. The larger i is, the deeper the feature of f_i is. This dense fusion approach can improve the representation ability of each level feature map to better adapt to the segmentation task. Our experimental results show that using dense fusion in the decoder part gives better results than using simple skip connections (see 4.4). The structure of Up0 is the same as that in the CNN branch, that is, the spatial resolution of the feature map is first upsampled back to the spatial resolution of the input image by bilinear interpolation, and then the number of channels is adjusted using 1×1 convolution to make it equal to the number of categories to get the final prediction output.

$$f_{cat} = concatenate[f_{i_{encoder}}, Up(Conv(f_{i_{deocder+1}})), \ldots, Up(Conv(f_n))] \quad (1)$$

$$f_{i_{decoder}} = ReLU(BN(Conv(f_{cat})) \quad (2)$$

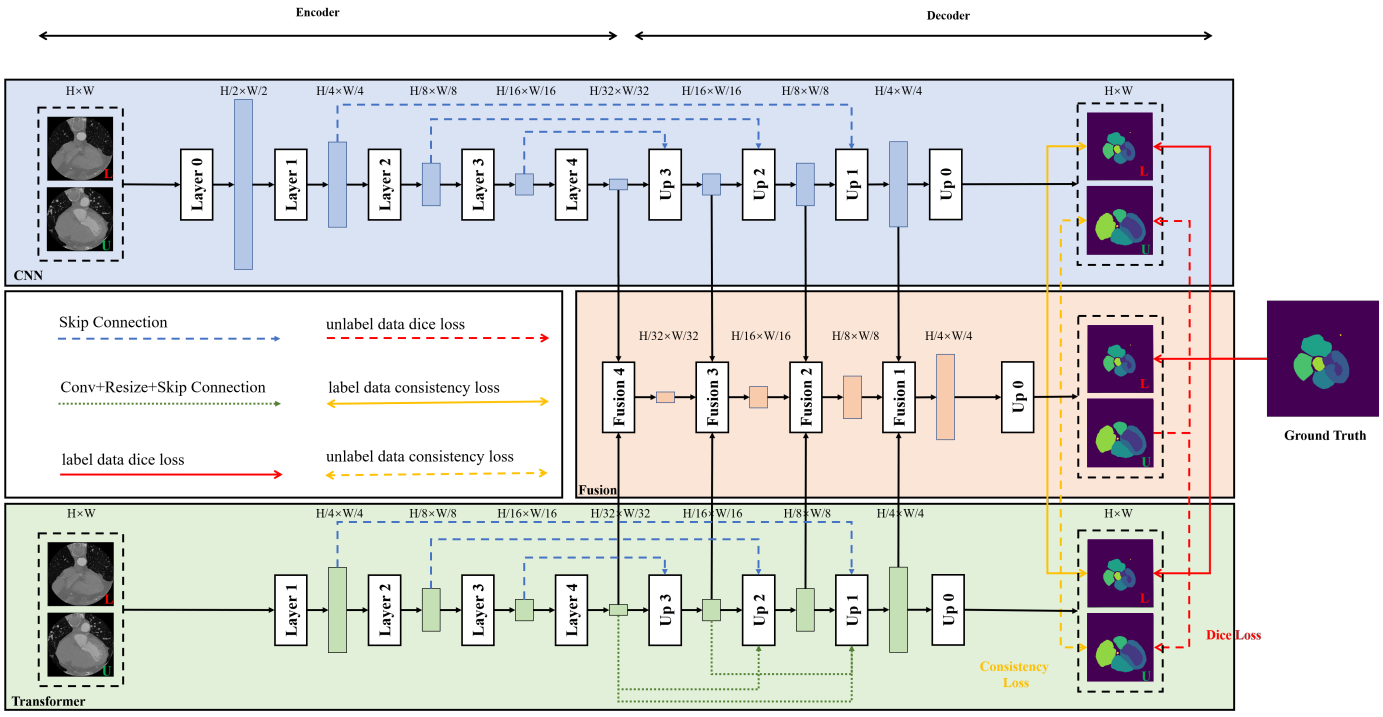

Fig. 1: The overall structure of SemiCTrans. The overall framework includes a CNN branch, a Transformer branch, and a fusion module. The CNN branch is responsible for extracting spatial information, and the Transformer branch is responsible for extracting global context, both of which follow the encoder-decoder structure. The two branches send the extracted multi-scale features to the fusion module, and the output of the fusion module is the final prediction. Both labeled and unlabeled images are fed during training. For labeled images, we use ground truths to supervise the outputs and fusion results of the two branches. For unlabeled images, we combine the ideas of pseudo labels and consistency constraints to exploit them.

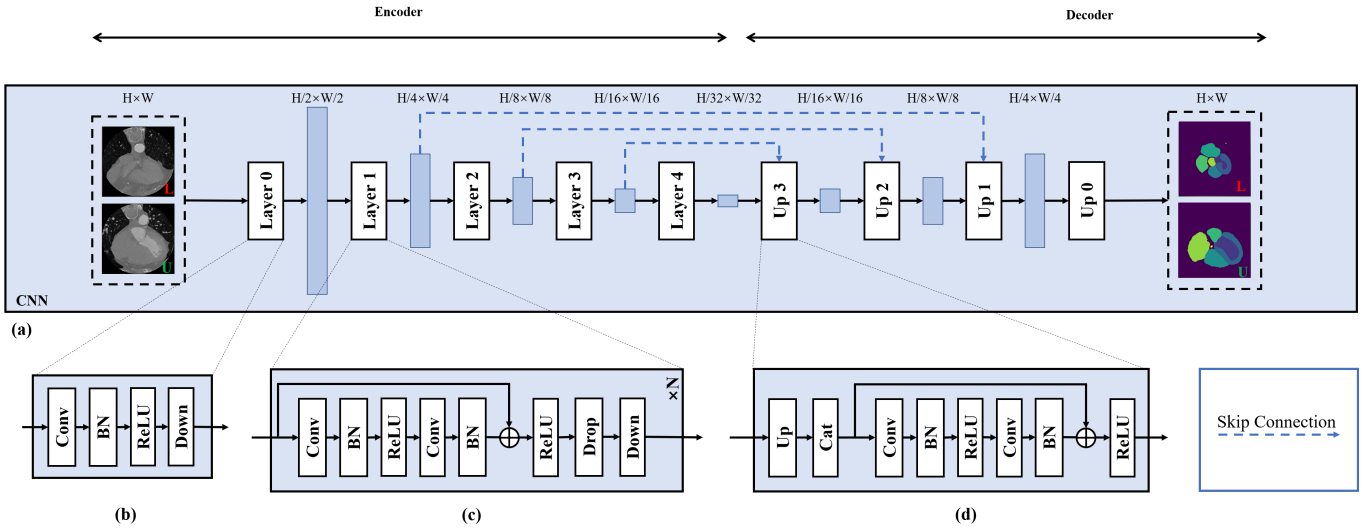

Fig. 2: CNN branch. (a) is the overall structure of the CNN branch, (b) is the structure of Layer0, (c) is the structure shared by Layer1 to Layer4, and (d) is the structure shared by Up4 to Up1.

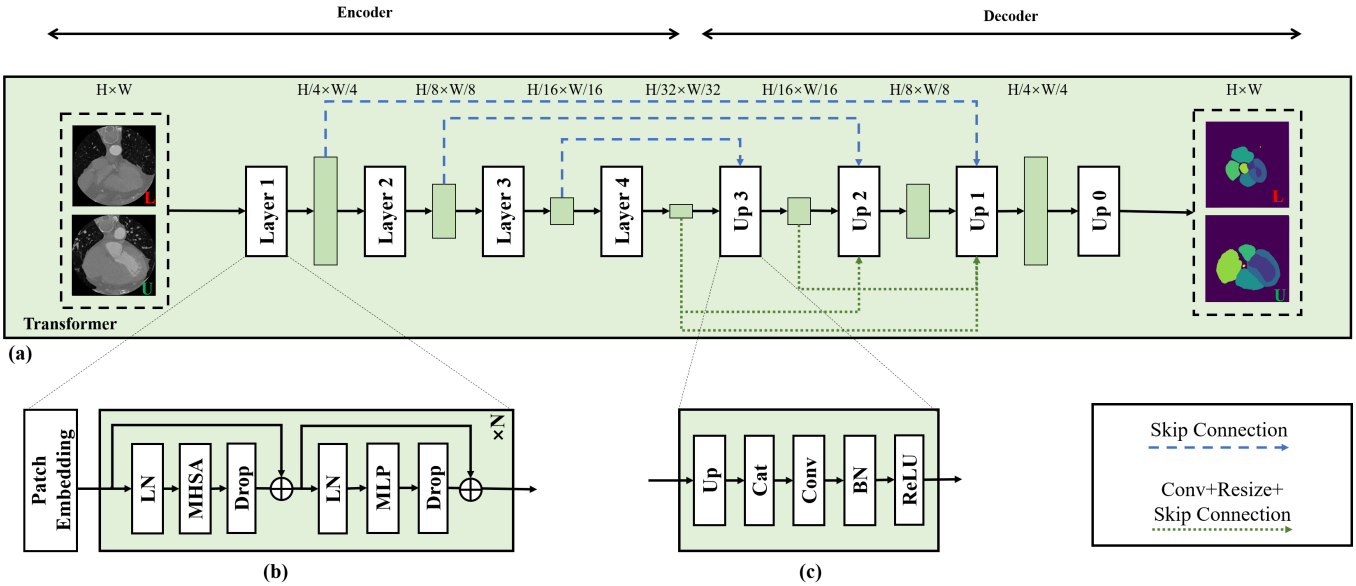

Fig. 3: Transformer branch. (a) is the overall structure of the Transformer branch, (b) is the structure shared by Layer1 to Layer4, (c) is the structure shared by Up3 to Up1.

3.3 Fusion Module

As shown in Fig. 4, the fusion module is used to fuse the features from the CNN branch and the Transformer branch, extract different semantic information from the two kinds of features, and combine them effectively. The fused features at different abstraction levels are successively concatenated and fused to produce the final pixel-level prediction output. In order to better combine different semantic information, we refer to the TransFuse[4] and use the Attention mechanism in the fusion module to weight the important information of CNN and Transformer features while suppressing irrelevant information. Specifically, the fusion module first performs Channel Attention[53] on Transformer features to extract global information. Afterward, Spatial Attention[54] is performed on the CNN features to obtain local details. Then the two processed features are concatenated, and the channel is adjusted by convolution to obtain the fusion feature, as shown in Eq.(3) and Eq.(4). Except for the highest layer, the fusion features of other layers are concatenated with the fusion features of higher layers. That is, first upsampling the low-resolution fusion feature, then concatenating it with the high-resolution feature, and then adjusting the channel through the convolution layer to obtain the new fusion feature. As shown in Eq.(5) and Eq.(6). Finally, Up0 upsamples the shallowest fusion features back to the spatial resolution of the input image using bilinear interpolation. After that, it uses 1×1 convolution to adjust the number of feature map channels to the number of categories, and obtains the final output of the fusion module.

$$f_{cat_i} = concatenate[SpatialAttention(f_{CNN_i}), ChannelAttention(f_{Transformer_i})] \quad (3)$$

$$f_{Fusion_i} = Drop(Conv(ReLU(BN(Conv(ReLU(BN(f_{cat_i})))))))) \quad (4)$$

$$f_{Fusion_i} = concatenate[Up(f_{Fusion_{i-1}}), f_{Fusion_i}] \quad (5)$$

$$f_{Fusion_i} = ReLU(f_{Fusion_i} + BN(Conv(ReLU(BN(Conv(f_{Fusion_i})))))) \quad (6)$$

3.4 Semi-Supervised Learning and Loss Functions

Because the labeling of medical images is time-consuming and labor-intensive, labeled data available for training is scarce, but a large number of unlabeled images exist. Therefore, we hope to combine semi-supervised learning to further improve the performance of the model with unlabeled data. We combine the ideas of consistency constraints and pseudo-labels to exploit unlabeled data. Specifically, we feed the model both labeled and unlabeled images during training, as shown in Fig. 1, where L represents labeled data and U represents unlabeled data. For labeled images, we use ground truths to supervise the outputs of both branches and fusion modules. In addition, since the input of the two branches is the same image, their prediction results are expected to be consistent, so we also calculate the consistency loss between the two branches' prediction results for consistency constraints. Although the input is the same image, since the two branches focus on extracting different semantic information from the image, the segmentation results they output are different. CNN can better segment details, while Transformer has better results for region segmentation. Therefore, by imposing consistency constraints on the two branches, they can learn from each other. Thereby, the effect of improving two branch predictions is achieved. At the same time, in order to allow the network to produce reliable predictions earlier, we use deep supervision for the supervised loss. That is, at each level of the network, we pass the predicted probability map through Softmax layer to obtain the predicted segmentation

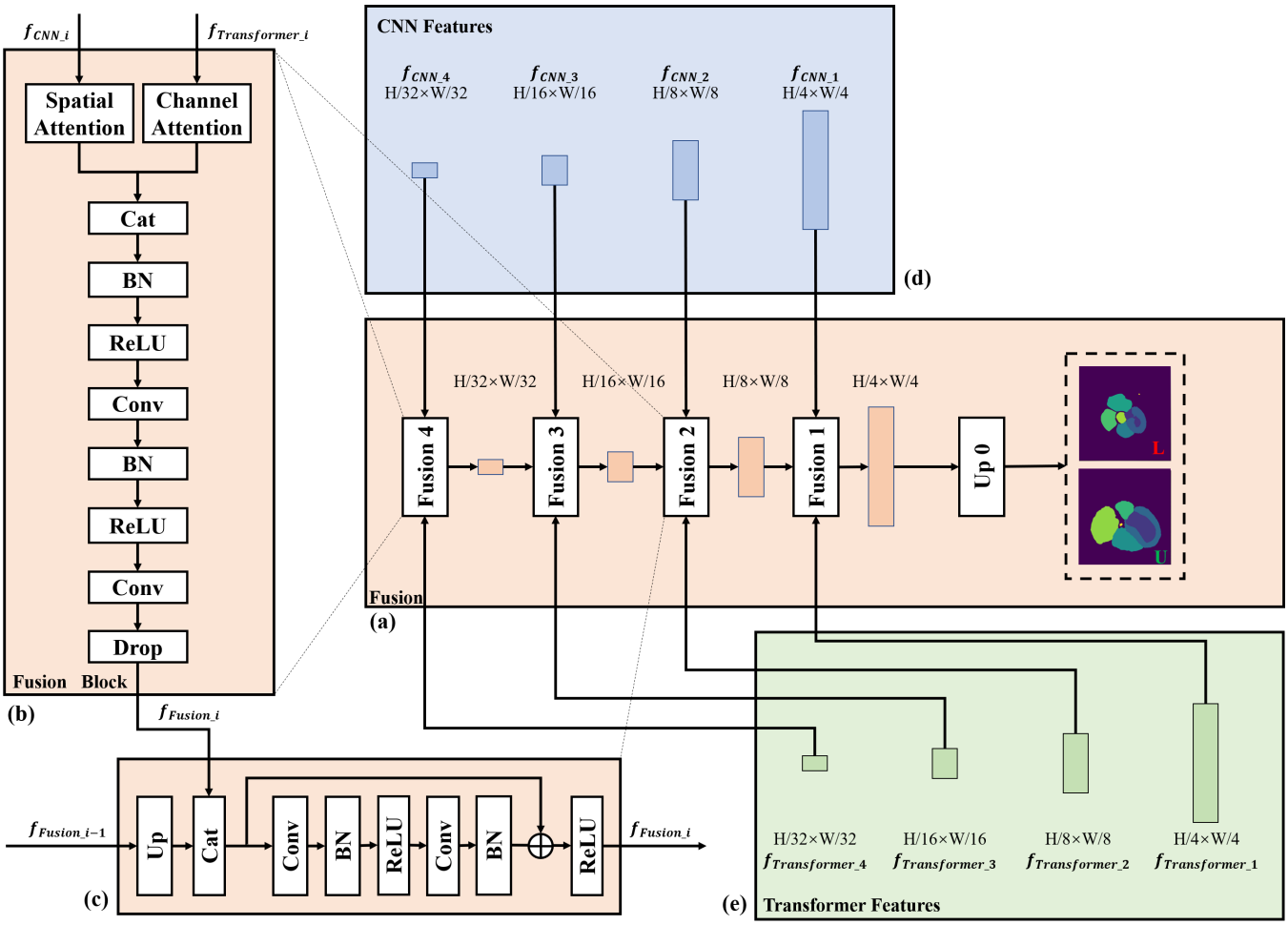

Fig. 4: Fusion module. (a) is the overall structure of the fusion module, (b) is the structure of the Fusion Block, which is also the structure of Fusion4, (b)+(c) are the shared structures of Fusion3 to Fusion1. They first fuse the features from different branches and then concatenate them with the higher-level fusion features. (d) and (e) represent the features from the CNN branch and the Transformer branch, which are sent to the fusion module according to the corresponding way on this Figure.

map and calculate the loss, including segmentation loss and consistency loss. Note that to avoid inaccuracies due to upsampling, our high-level segmentation loss is computed by downsampling the ground truths to the low resolution, rather than upsampling the predictions back to the original resolution. The supervised loss is shown in Eq. (7), Eq.(8) and Eq.(9), where L_{seg} represents the segmentation loss, which consists of the Dice Loss of each layer. L_{Dice_i} represents the Dice Loss of the layer i, and L_{Dice_1} represents the final segmentation output. $L_{consistency}$ represents the consistency loss, which consists of the mean squared error between the CNN output and the Transformer output in each layer, and L_{2_i} represents the mean squared error of the layer i.

$$L_{label} = L_{seg} + L_{consistency} \tag{7}$$

$$L_{seg} = 0.2L_{Dice_4} + 0.2L_{Dice_3} + 0.2L_{Dice_2} + 0.4L_{Dice_1} \tag{8}$$

$$L_{consistency} = 0.2L_{2_4} + 0.2L_{2_3} + 0.2L_{2_2} + 0.4L_{2_1} \tag{9}$$

For unlabeled images, we first use the prediction outputs of the fusion module as pseudo labels to supervise the prediction results of CNN and Transformer. In addition, we compute the difference between the predictions of the two branches as part of the unsupervised loss, and impose a consistency constraint on the two networks by minimizing this term. Because the predictions of the pseudo labels and the two branches are not very accurate at high levels, we do not use deep supervision for the unsupervised loss, and only calculate the loss for the final outputs. The unsupervised loss is shown in Eq.(10). The overall loss of the model consists of two parts: supervised loss and unsupervised loss, which are calculated as shown in Eq.(11).

$$L_{unlabel} = L_{Dice_{unlabel}} + L_{2_{unlabel}} \tag{10}$$

$$L = L_{label} + L_{unlabel} \tag{11}$$

4 EXPERIMENTS

4.1 Data Set

We used the dataset from the Multi-Modality Whole Heart Segmentation(MM-WHS) 2017 Challenge([55]), which includes 20 CT examples. Each CT sample is 3D data, which can be converted into 200 to 400 2D slices, and the spatial resolution of each slice is 512×512. Each CT sample has eight categories, left ventricle, right ventricle, left atrium, right atrium, myocardium, aorta, pulmonary artery, and background. Consistent with the comparative method, in the experiments, we use 5 CT examples as the test data, 5 CT examples as the validation data, 10 CT examples as the training data.

In addition to the MMWHS dataset, we also used MICCAI 2017 Automated Cardiac Diagnosis Challenge (ACDC) dataset[56]. We selected the training set from the original competition as the experimental data in our experiments, which contains 100 MR samples. Each MR sample is 3D data and can be converted into 100-300 slices. Since the spatial resolution of different MR samples is different, we uniformly scaled them to 256×256 in the experiment. Each MR sample contains four categories: right ventricle (RV), left ventricle (LV), left ventricle myocardium (Myo) and background. In the experiment, we use 70 samples as the training data, 20 samples as the test data, and 10 samples as the validation data.

4.2 Experimental Configuration

Our experiments are performed on an NVIDIA GeForce GTX 1080Ti GPU. We use the Adam optimizer with a learning rate of 8e-5, a batch size of 2 for fully supervised, and a batch size of 4 (2 labels + 2 unlabels) for semi-supervised. For the MMWHS dataset, we set the number of training epochs to 20 and for the ACDC dataset, we set the number of training epochs to 60. In the experiment, we use the $Dice$ to measure the result of the segmentation, which is calculated by (12), where Y is ground truth and X is segmentation result. The larger the $Dice$, the more accurate the segmentation results.

$$Dice = \frac{2 \times |X \bigcap Y|}{|X| + |Y|} \tag{12}$$

4.3 Compare with Other Methods

To validate the effectiveness of SemiCTrans, we compare SemiCTrans with other advanced methods in both semi-supervised and fully supervised configurations.

4.3.1 Comparison of semi-supervised methods

We compared our method with other semi-supervised methods in ACDC dataset, including deep adversarial network (DAN)[57], mean teacher (MT)[58], uncertainty-aware mean teacher (UAMT)[59], entropy minimization (EM)[60], cross pseudo supervision (CPS)[61], cross-teaching between CNN and transformer (CTCT)[62] and cross consistency training (CCT)[63]. Following the data division way of dual-teacher uncertainty-aware guided combination of CNN and Transformer (DTACT) [64], we tested three different semi-supervised settings, they were using 5%, 10% and 20% of the training data as labelled data. The results are reported in Table 1, some of which refer to [64]. The n/m in the table indicates that there are a total of $n + m$ samples in the training set, of which n are labeled and m are unlabeled. As can be seen from the results, our method achieves the best results among all compared methods in all three configurations. In particular, in settings with less labelled data, our method can achieve a large lead, indicating that SemiCTrans can make full use of the information in the unlabelled data. Fig. **??** shows the segmentation results with different scales of labeled training data. As can be seen from this, the segmentation results are progressively finer as more labeled data is available.

TABLE 1: Comparison with other semi-supervised methods on ACDC dataset.

Labeled data	Methods	Mean $Dice$
3/67 (5%)	DAN	0.528
	MT	0.566
	UAMT	0.610
	EM	0.602
	CPS	0.603
	CTCT	0.656
	CCT	0.586
	DTACT	0.694
	SemiCTrans	**0.809**
7/63 (10%)	DAN	0.795
	MT	0.810
	UAMT	0.815
	EM	0.791
	CPS	0.833
	CTCT	0.864
	CCT	0.816
	DTACT	0.878
	SemiCTrans	**0.897**
14/56 (20%)	DAN	0.847
	MT	0.849
	UAMT	0.861
	EM	0.845
	CPS	0.863
	CTCT	0.873
	CCT	0.860
	DTACT	0.886
	SemiCTrans	**0.904**

Following the data division way of Semi-supervised Contrastive Learning for Label-Efficient Medical Image Segmentation (SSCL)[65], we compared our method with other semi-supervised methods in MMWHS dataset, including a data augmentation based method Mixup[66], global and local contrastive learning method (GLCL)[67], self-supervised global and local contrastive learning method (SSGLCL) [68], and transformation-consistent self-ensembling model (TCSM)[69]. We also tested three different semi-supervised settings, they were using 10%, 20% and 40% of the training data as labelled data. The results are shown in Table 2, some of which refer to [65]. As can be seen from the results, our method clearly outperforms the other comparison methods. Even when using only one labeled sample, our method still achieves good segmentation accuracy, which validates the effectiveness of our semi-supervised framework.

4.3.2 Comparison of fully-supervised methods

In addition to comparisons with semi-supervised methods, we also compare with some advanced fully supervised

TABLE 2: Comparison with other semi-supervised methods on MMWHS dataset.

Labeled data	Methods	Mean *Dice*
1/9 (10%)	Mixup	0.365
	GLCL	0.359
	SSGLCL	0.367
	TCSM	0.347
	SSCL	0.384
	SemiCTrans	**0.842**
2/8 (20%)	Mixup	0.541
	GLCL	0.487
	SSGLCL	0.490
	TCSM	0.489
	SSCL	0.553
	SemiCTrans	**0.908**
4/6 (40%)	Mixup	0.755
	GLCL	0.724
	SSGLCL	0.730
	TCSM	0.733
	SSCL	0.764
	SemiCTrans	**0.911**

methods on the two datasets. Following the data partitioning way mentioned in 4.1, we use all training data as the labeled data, that is, 70 training samples for ACDC and 10 training samples for MMWHS.

For ACDC dataset, we compared our method with four fully supervised methods, including UNet[15], Attention UNet[70], TransUNet[32], and SwinUnet[37]. The results are shown in Table 3, some of which refer to [37]. For the MMWHS dataset, we compared our method with four fully supervised methods, including UNet[15], two stage UNet (Two Stage)[71], nested u-structure GAN (Nu-Net-GAN)[72] and nested recurrent residual UNet on GAN (NRRU-GAN)[73]. The results are shown in Table 4.

TABLE 3: Comparison with other fully supervised methods on ACDC dataset.

Methods	Mean *Dice*
UNet	0.876
Attention UNet	0.866
TransUNet	0.897
SwimUet	0.900
SemiCTrans	0.914

TABLE 4: Comparison with other fully supervised methods on MMWHS dataset.

Methods	Mean *Dice*
UNet	0.785
Two-Stage	0.793
NU-Net-Gan	0.899
NRRU-GAN	0.908
SemiCTrans	0.913

4.4 Ablation Study

To verify the effectiveness of the proposed dual-branch framework, we test the results of using the CNN branch alone, the Transformer branch alone, and using the fusion framework on the MMWHS dataset in the supervised manner. The results are shown in Table 5. As can be seen from Table 5, the two branches can achieve better results in the framework than if they were trained separately, because they can learn complementary information from each other. Furthermore, the fusion result of SemiCTrans can be further improved on the results of the two branches. In addition, it can be seen from the Fig. 5, in the training process, the loss of fusion module is lower than that of the two branches, which can converge better.

TABLE 5: Ablation study on the effectiveness of the proposed dual-branch framework.

Models	Test *Dice*
CNN alone	0.882
Transformer alone	0.879
CNN in SemiCTrans	0.906
Transformer in SemiCTrans	0.894
SemiCTrans	0.913

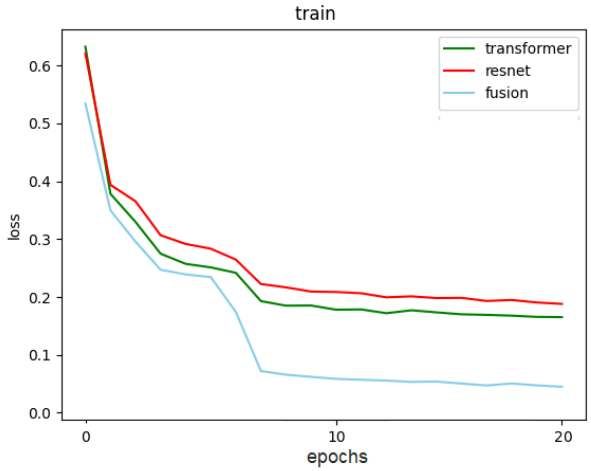

Fig. 5: Illustration of the train loss curves of CNN branch, Transformer branch, and fusion module during the training process.

In addition, in order to show that our improvement to the Transformer branch is effective, we conducted some comparative experiments on different changes in the Transformer branch using MMWHS dataset in supervised manner, and the results are shown in Table 6. The simple decoder refers to the residual connection of the features of different scales to obtain the final prediction output. In the configuration of single-scale features, we construct features of different scales by upsampling the top-level features multiple times in the decoder. In the configuration of multi-scale features, the features of different scales are obtained by the encoder, and they are different. The dense decoder, as we introduced in 3.2, makes each layer of feature maps in the decoder combine the feature maps of all previous layers to deepen the use of these feature maps through dense connections. As can be seen from the Table 6, both

the construction of multi-scale features and the use of dense decoders can significantly improve the prediction results of the Transformer branch. Compared to the original configuration, using multi-scale features and a dense decoder can improve the segmentation dice by more than 0.6.

TABLE 6: Comparison of different configurations for the Transformer branch.

Models	Test *Dice*
Single-scale features + Simple decoder	0.810
Multi-scale features + Simple decoder	0.852
Multi-scale features + Dense decoder	0.879

To verify the effectiveness of semi-supervision, we compared different semi-supervised loss components in MMWHS dataset, and the results are shown in Table 7. The fully supervised configuration in Table 7 uses 2 labeled samples for training and the semi-supervised configuration uses 2 labeled samples and 8 unlabeled samples for training. The validation and test data are the same for both, following the division way in 4.1. It can be seen from the results that both pseudo labels and consistency constraints can achieve better results than full supervision. After combining them, the segmentation results can be further improved. It should be noted that for full supervision and consistency constraints, using deep supervision can improve the effect and make the model get better segmentation results at deeper layers. However, for pseudo labels, the reliability of deep pseudo labels is not high, so deep supervision will reduce the final segmentation effect. Therefore, we only use deep supervision for consistency constraints in fully supervised and semi-supervised losses.

TABLE 7: Comparison of fully supervised model and semi-supervised models with different loss components.

Models	Mean *Dice*
Fully supervised model	0.679
Semi-supervised model with pseudo labels	0.902
Semi-supervised model with consistency constraints	0.904
Semi-supervised model with pseudo labels + consistency constraints	0.908

5 Conclusion

Aiming to solve the problem of the variable shape of target regions in medical image segmentation, we propose a dual-branch medical image segmentation model based on CNN and Transformer, which combines the good local detail extraction ability of CNN and the good long-range dependency modeling ability of Transformer. At the same time, the Transformer branch is improved so that it can generate multi-scale features and better integrate multi-scale information through dense connections in the decoder. Furthermore, we extend the model to semi-supervised. By combining the ideas of pseudo-labels and consistency constraints, the information in unlabeled images is used to further improve the segmentation effect of the model.

References

[1] Ran Gu, Guotai Wang, Tao Song, Rui Huang, Michael Aertsen, Jan Deprest, Sébastien Ourselin, Tom Vercauteren, and Shaoting Zhang. Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE Transactions on Medical Imaging*, 40(2):699–711, 2021.

[2] S Sudha, K B Jayanthi, C Rajasekaran, and T Sunder. Segmentation of roi in medical images using cnn- a comparative study. In *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, pages 767–771, 2019.

[3] Kaustav Brahma, Viksit Kumar, Anthony E. Samir, Anantha P. Chandrakasan, and Yonina C. Eldar. Efficient binary cnn for medical image segmentation. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 817–821, 2021.

[4] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. 2021.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *arXiv*, 2017.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 2020.

[7] H. Touvron, M. Cord, M. Douze, F. Massa, and H Jégou. Training data-efficient image transformers distillation through attention. 2020.

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing.

[9] R. Strudel, R. Garcia, I. Laptev, and C. Schmid. Segmenter: Transformer for semantic segmentation. 2021.

[10] A. K. Chaudhary, P. K. Gyawali, L. Wang, and J. B. Pelz. Semi-supervised learning for eye image segmentation. 2021.

[11] Shuai Chen, Gerda Bortsova, Garcia Uceda Juarez, Gijs Van Tulder, and Marleen De Bruijne. Multi-task attention-based semi-supervised learning for medical image segmentation. 2019.

[12] Yuexiang Li, Jiawei Chen, Xinpeng Xie, Kai Ma, and Yefeng Zheng. Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation. *arXiv e-prints*, 2020.

[13] Xuyang Cao, Houjin Chen, Yanfeng Li, Yahui Peng, and Lin Cheng. Uncertainty aware temporal-ensembling model for semi-supervised abus mass segmentation. *IEEE Transactions on Medical Imaging*, PP(99), 2020.

[14] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*, pages 234–241. Springer International Pub-

lishing, Cham, 2015.

[16] F. Milletari, N. Navab, and S. A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. 2016.

[17] Yi-Jie Huang, Qi Dou, Zi-Xian Wang, Li-Zhi Liu, Li-Sheng Wang, Hao Chen, Pheng-Ann Heng, and Rui-Hua Xu. Hl-fcn: Hybrid loss guided fcn for colorectal cancer segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 195–198, 2018.

[18] Dong Nie, Li Wang, Ehsan Adeli, Cuijin Lao, Weili Lin, and Dinggang Shen. 3-d fully convolutional networks for multimodal isointense infant brain image segmentation. *IEEE Transactions on Cybernetics*, 49(3):1123–1136, 2019.

[19] Ran Zhou, Fumin Guo, M. Reza Azarpazhooh, J. David Spence, Eranga Ukwatta, Mingyue Ding, and Aaron Fenster. A voxel-based fully convolution network and continuous max-flow for carotid vessel-wall-volume segmentation from 3d ultrasound images. *IEEE Transactions on Medical Imaging*, 39(9):2844–2855, 2020.

[20] Yu-xue Wang and Chao Xu. Ct image segmentation of rectal tumor based on u-net network. In *2021 2nd International Conference on Big Data and Informatization Education (ICBDIE)*, pages 69–72, 2021.

[21] Nuh Hatipoglu and Gokhan Bilgin. Histopathological image segmentation using u-net based models. In *2021 Medical Technologies Congress (TIPTEKNO)*, pages 1–4, 2021.

[22] Naomi Yagi, Manabu Nii, and Syoji Kobashi. Abdominal organ area segmentation using u-net for cancer radiotherapy support. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 1210–1214, 2019.

[23] H. Zhang, H. Zhu, and X. Ling. Polar coordinate sampling-based segmentation of overlapping cervical cells using attention u-net and random walk. *Neurocomputing*, 383:212–223, 2020.

[24] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[25] Z. Zhang, C. Wu, S. Coleman, and D. Kerr. Dense-inception u-net for medical image segmentation. *Computer Methods and Programs in Biomedicine*, 192:105395, 2020.

[26] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *IEEE*, 2016.

[27] M. Baldeon-Calisto and S. K. Lai-Yuen. Adaresu-net: Multiobjective adaptive convolutional neural network for medical image segmentation. *Neurocomputing*, 2020.

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014.

[29] Aiyue Huang, Qing Wang, Li Jiang, and Jiangshan Zhang. Automatic segmentation of median nerve in ultrasound image by a combined use of u-net and vgg16. In *2021 IEEE International Ultrasonics Symposium (IUS)*, pages 1–4, 2021.

[30] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

[31] L. Ye, M. Rochan, Z. Liu, and Y. Wang. Cross-modal self-attention network for referring image segmentation. 2019.

[32] J. Chen, Y. Lu, Q. Yu, X. Luo, and Y. Zhou. Transunet: Transformers make strong encoders for medical image segmentation. 2021.

[33] Yunhe Gao, Mu Zhou, and Dimitris Metaxas. Utnet: A hybrid transformer architecture for medical image segmentation. 2021.

[34] Wenxuan Wang, Chen Chen, Meng Ding, Jiangyun Li, and Sen Zha. Transbts: Multimodal brain tumor segmentation using transformer. 2021.

[35] Davood Karimi, Serge Didenko Vasylechko, and Ali Gholipour. Convolution-free medical image segmentation using transformers. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 78–88, Cham, 2021. Springer International Publishing.

[36] A. Hatamizadeh, D. Yang, H. Roth, and D. Xu. Unetr: Transformers for 3d medical image segmentation. 2021.

[37] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. 2021.

[38] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. 2021.

[39] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, and Pheng-Ann Heng. Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model, 08 2018.

[40] X. Chen, Y. Yuan, G. Zeng, and J. Wang. Semi-supervised semantic segmentation with cross pseudo supervision. 2021.

[41] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. pages 1205–1214, 06 2021.

[42] Xiangde Luo, Jieneng Chen, Tao Song, Yinan Chen, Guotai Wang, and Shaoting Zhang. Semi-supervised medical image segmentation through dual-task consistency, 09 2020.

[43] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, Ferdous Sohel, and Dan Xu. Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation. pages 6964–6973, 10 2021.

[44] Guotai Wang, Shuwei Zhai, Giovanni Lasio, Baoshe Zhang, Byong Yi, Shifeng Chen, Thomas Macvittie, Dimitris Metaxas, Jinghao Zhou, and Shaoting Zhang. Semi-supervised segmentation of radiation-induced pulmonary fibrosis from lung ct scans with multi-scale guided dense attention. *IEEE Transactions on Medical Imaging*, PP:1–1, 10 2021.

[45] Yulei Qin, Hao Zheng, Xiaolin Huang, Jie Yang, and Yuemin Zhu. Pulmonary nodule segmentation with ct

sample synthesis using adversarial networks. *Medical Physics*, 46, 12 2018.

[46] Hongming Shan, Yi Zhang, Qingsong Yang, Uwe Kruger, Wenxiang Cong, and Ge Wang. 3-d convolutional encoder-decoder network for low-dose ct via transfer learning from a 2-d trained network. *IEEE Transactions on Medical Imaging*, PP, 02 2018.

[47] Daphne Schlesinger and Collin Stultz. Deep learning for cardiovascular risk stratification. *Current Treatment Options in Cardiovascular Medicine*, 22, 06 2020.

[48] Rosana el Jurdi, Caroline Petitjean, Paul Honeine, and Fahed Abdallah. Bb-unet: U-net with bounding box prior. *IEEE Journal of Selected Topics in Signal Processing*, PP:1–1, 06 2020.

[49] Shuxin Wang, Shilei Cao, Dong Wei, Wang Renzhen, Kai Ma, Liansheng Wang, Deyu Meng, and Yefeng Zheng. Lt-net: Label transfer by learning reversible voxel-wise correspondence for one-shot medical image segmentation. pages 9159–9168, 06 2020.

[50] Hsien-Tzu Cheng, Chun-Fu Yeh, Po-Chen Kuo, Andy Wei, Keng-Chi Liu, Mong-Chi Ko, Kuan-Hua Chao, Yu-Ching Peng, and Tyng-Luh Liu. *Self-similarity Student for Partial Label Histopathology Image Segmentation*, pages 117–132. 11 2020.

[51] Yuexiang Li, Jiawei Chen, Xinpeng Xie, Kai Ma, and Yefeng Zheng. *Self-Loop Uncertainty: A Novel Pseudo-Label for Semi-supervised Medical Image Segmentation*, pages 614–623. 09 2020.

[52] Zhiqi Li, Wenhai Wang, Enze Xie, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, Tong Lu, and Ping Luo. Panoptic segformer, 09 2021.

[53] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):2011–2023, 2020.

[54] Sanghyun Woo, Jongchan Park, Joon Young Lee, and In So Kweon. Cbam: Convolutional block attention module. *Springer, Cham*, 2018.

[55] X. Zhuang, L. Li, C. Payer, D. Stern, M. Urschler, M. P. Heinrich, J. Oster, C. Wang, O. Smedby, C. Bian, X. Yang, P. A. Heng, A. Mortazi, U. Bagci, G. Yang, C. Sun, G. Galisot, J. Y. Ramel, T. Brouard, Q. Tong, W. Si, X. Liao, G. Zeng, Z. Shi, G. Zheng, C. Wang, T. MacGillivray, D. Newby, K. Rhode, S. Ourselin, R. Mohiaddin, J. Keegan, D. Firmin, and G. Yang. Evaluation of algorithms for multi-modality whole heart segmentation: An open-access grand challenge. *Medical Image Analysis*, 58:101537, 2019.

[56] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohé, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jäger, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Christian F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Išgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, and Pierre-Marc Jodoin. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018.

[57] Yizhe Zhang, Lin Yang, Jianxu Chen, Maridel Fredericksen, David Hughes, and Danny Chen. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. pages 408–416, 09 2017.

[58] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. 2017.

[59] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. pages 605–613, 2019.

[60] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2512–2521, 2019.

[61] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. pages 2613–2622, 2021.

[62] X. Luo, M. Hu, T. Song, G. Wang, and S. Zhang. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. *arXiv e-prints*, 2021.

[63] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. pages 12671–12681, 2020.

[64] Zhiyong Xiao, Yixin Su, Zhaohong Deng, and Weidong Zhang. Efficient combination of cnn and transformer for dual-teacher uncertainty-aware guided semi-supervised medical image segmentation. *COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE*, 2022.

[65] Xinrong Hu, Dewen Zeng, Xiaowei Xu, and Yiyu Shi. Semi-supervised contrastive learning for label-efficient medical image segmentation. pages 481–490, 2021.

[66] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[67] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. *International conference on machine learning, PMLR*, page 1597–1607, 2020.

[68] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *arXiv preprint arXiv:2006.10511*, 2020.

[69] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):523–534, 2021.

[70] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. Mcdonagh, N. Y. Hammerla, and B. Kainz. Attention u-net: Learning where to look for the pancreas. *IMIDL Conference*, 2018.

[71] Tao Liu, Yun Tian, Shifeng Zhao, Xiaoying Huang, and Qingjun Wang. Automatic whole heart segmentation

using a two-stage u-net framework and an adaptive threshold window. *IEEE Access*, 7:1–1, 06 2019.

[72] Zeyu Lou, Kening Le, and Xiaolin Tian. Nu-net based gan: Using nested u-structure for whole heart auto segmentation. *021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pages 601–604, 06 2021.

[73] Le Kening, Lou Zeyu, and Tian Xiaolin. Nested recurrent residual unet (nrru) on gan (nrrg) for cardiac ct images segmentation task. *2021 2nd International Conference on Artificial Intelligence and Information Systems*, pages 1–5, 05 2021.