DDCL-Net: Dual-stream Dense Contrastive Learning Network for Semi-supervised Medical Image Segmentation

Zheng Huang[1], Di Gai[1,2,3], Weidong Min[1,2,3]*, Qi Wang[1,2,3], Lixin Zhan[1]
[1]School of Mathematics and Computer Science, Nanchang University, Nanchang, China
[2]Institute of Metaverse, Nanchang University, Nanchang, China
[3]Jiangxi Key Laboratory of Smart City, Nanchang, China
huangzheng@email.ncu.edu.cn, gaidi@ncu.edu.cn, minweidong@ncu.edu.cn
wangqi@ncu.edu.cn, zhanlixin@email.ncu.edu.cn

Abstract

Most existing semi-supervised methods in medical image segmentation often ignore optimize feature distribution, leading to weak inter-class separability and poor decision boundaries. Additionally, the ambiguity of target contours in medical images can result in mis-segmentation due to a lack of understanding of contour features. To address these issues, we propose a novel semi-supervised segmentation framework called Dual-stream Dense Contrastive Learning Network. Our framework utilizes a prediction stream to obtain a segmentation map and Signed Distance Map, which learn from each other to explore complementary knowledge and enhance the exploitation of contour features. In our feature stream, we introduce the Dense Local Features Contrastive Learning module, which consists of Sampling by Shape Info (SSI) and Dense Local Features Contrast (DLFC). SSI combines local shape information to form the Balanced Coefficient, guiding the sampling of positive and negative pairs between dense local features. DLFC enhances intra-class compactness and inter-class separability through contrastive learning, with consistency regularization to promote learning efficiency. Our approach achieves significant improvements over existing state-of-the-art methods on benchmark datasets for medical image segmentation.

***Keywords:** Semi-supervised learning, Local features, Contrastive Learning, Image Segmentation*

1. Introduction

Medical image segmentation technology [3, 42] can provide physicians with valuable assistance in localizing and quantitatively analyzing tissue organs or lesion regions by altering the visualization process of medical images. With the current penetration of artificial intelligence technology into various fields, intelligent analysis technology has rapidly developed. This development has greatly facilitated the application of artificial intelligence in medical image segmentation, providing physicians with more efficient and accurate diagnostic tools.

In recent years, deep learning [23, 43] has demonstrated impressive success in medical image segmentation. However, these methods [1, 9, 28] typically require a large amount of annotated data for training. Obtaining labeled data can be prohibitively expensive, particularly in medical imaging, where only specialized medical experts can provide reliable labels [17]. As a result, there has been a surge of interest in methodological research under limited supervision. Fortunately, semi-supervised segmentation [19, 22, 34] attempts to apply merely a small amount of labeled data and learns from a considerable number of unlabeled data. These methods aim to reduce the amount of required labeled data, while still achieving high-quality segmentation results, making them an attractive option for medical image analysis.

Several semi-supervised segmentation methods[34, 22, 33] focus on adding a certain degree of perturbations to the original unlabeled data and enforcing consistency between the model predictions of the original and perturbed data. However, the above method solely considers the consistency of the final predicted segmentation map and fails to consider the distribution of the optimized feature space. It is beneficial for prediction if the model is allowed to learn that features of the same class are aggregated in the feature space and features of different classes are dispersed [38]. Fortunately, contrastive learning [5, 7] enables the model to involve a large amount of unlabeled data to train the model to optimize feature distribution by effectively selecting positive and negative pairs. However, the effect of contrastive learning depends on the sampling of positive and negative pairs. If similar features are adopted as negative pairs, it will destroy the distribution of features. Concretely, Zhao *et al.* [48] decomposed images into patches and proposed cross-level contrastive learning to exploit local features, but the

random sampling in negative pairs may lead to a decrease in accuracy. Wu *et al*. [40] improved the random sampling between patches based on pseudo-label, which lacks geometric awareness. Consequently, our approach embeds the predicting Signed Distance Map (SDM) [26] and the predicting segmentation map as the prediction stream, which can strengthen the exploitation of contour features. At the same time, a factor named Balanced Coefficient (BC) to determine the local variability of SDM is proposed, which guides the sampling of positive and negative pairs in local features.

In this paper, we propose a dual-stream dense contrastive learning network for semi-supervised medical image segmentation, which utilizes a large amount of unlabeled data to optimize the segmentation process of the model. Our network consists of a teacher-student model, a prediction stream, a feature stream, and a Dense Local Features Contrastive Learning (DLFCL) module. The function of the teacher-student model is to extract the features of the input image and send them into the prediction stream and the feature stream. The prediction stream simultaneously performs the predicting segmentation map task and the predicting SDM task to strengthen the exploitation of contour features through dual-task mutual learning. The DLFCL module in feture stream enhances the intra-class compactness and inter-class separability through Sampling by Shape Info (SSI) and Dense Local Features Contrast (DLFC). The SSI combines shape info to form BC, which guides local feature contrastive learning in DLFC. We also add some perturbations to the unlabeled data. For feature stream, the variability of feature pairs can be reinforced to improve contrastive learning efficiency, while for prediction streams, the model's understanding of global information can be enhanced to reduce overfitting. Extensive experiments conducted on three publicly available datasets show that our approach effectively exploits unlabeled data for learning and outperforms other methods. Our main contributions are summarized as follows:

- We propose a novel and generic dual-stream dense contrastive learning network for semi-supervised segmentation, which effectively leverages unlabeled data for learning and enhances the utilization of contour information through dual-stream.
- We propose BC in the DLFCL module to guide the sampling of positive and negative pairs between local features and further improve intra-class compactness and inter-class separability by contrastive learning.
- Extensive experiments on two challenging public datasets demonstrate more satisfactory segmentation results than other methods and achieve state-of-the-art performance.

2. Related Work

2.1. Semi-Supervised Medical Image Segmentation

The goal of semi-supervised learning [8, 12, 37] is to optimize the learning process of the network over limited labeled data by using a large amount of unlabeled data. In recent years, the semi-supervised method based on deep learning [32, 41, 44] has become a hot research direction in medical image segmentation. For example, using the idea of adversarial learning, Zhang *et al*. [47] proposed an evaluation network to distinguish the segmentation results of unlabeled images and labeled images. Li *et al*. [29]proposed a region-based attention method based on a confidence network and applied adversarial learning schemes to adaptively train the network using unlabeled data. In order to make the multi-scale prediction of unlabeled images consistent, Luo *et al*. [25] proposed a new uncertainty correction pyramid consistency regularization framework to prevent model collapse and detail loss. Li *et al*. [20] relied on extra shape information for adversarial training. Luo *et al*. [24] realized the consistency training of dual-task by inverting the predicted SDM into the segmentation map. Different from the above methods, we effectively utilize shape information for local features contrastive learning to improve the feature representation of the model.

2.2. Contrastive Learning

Contrastive learning [15, 35, 45] is a branch of self-supervised learning [27]. Its goal is to embed the features of the same category into the approximate feature space while pushing out the features of different categories. Now contrastive learning has achieved great success in the field of computer vision. Chen *et al*. [4] proposed the projection head to project the features once and map the features to a new dimension, providing a whole new framework for follow-up work. Wang *et al*. [39] has designed a dense contrastive learning method that can perform dense contrastive learning at the level of local features. The key to contrastive learning is to select positive and negative pairs. Zhao *et al*. [48] proposed cross-level contrastive learning for medical image segmentation, but chose random sampling when selecting negative pairs. Chaitanya *et al*. [2] took advantage of the structural similarity of medical images to select positive and negative pairs. Wu *et al*. [40] proposed positive and negative pair sampling based on pseudo-label, improving the process of contrastive learning. Our method provides a novel way for contrastive learning by combining shape information with positive and negative pairs sampling.

2.3. Consistency Regularization

The purpose of consistency regularization [19, 30, 34] is to effectively reduce overfitting and enhance the robustness of the model by adding perturbations [33] to the same inputs

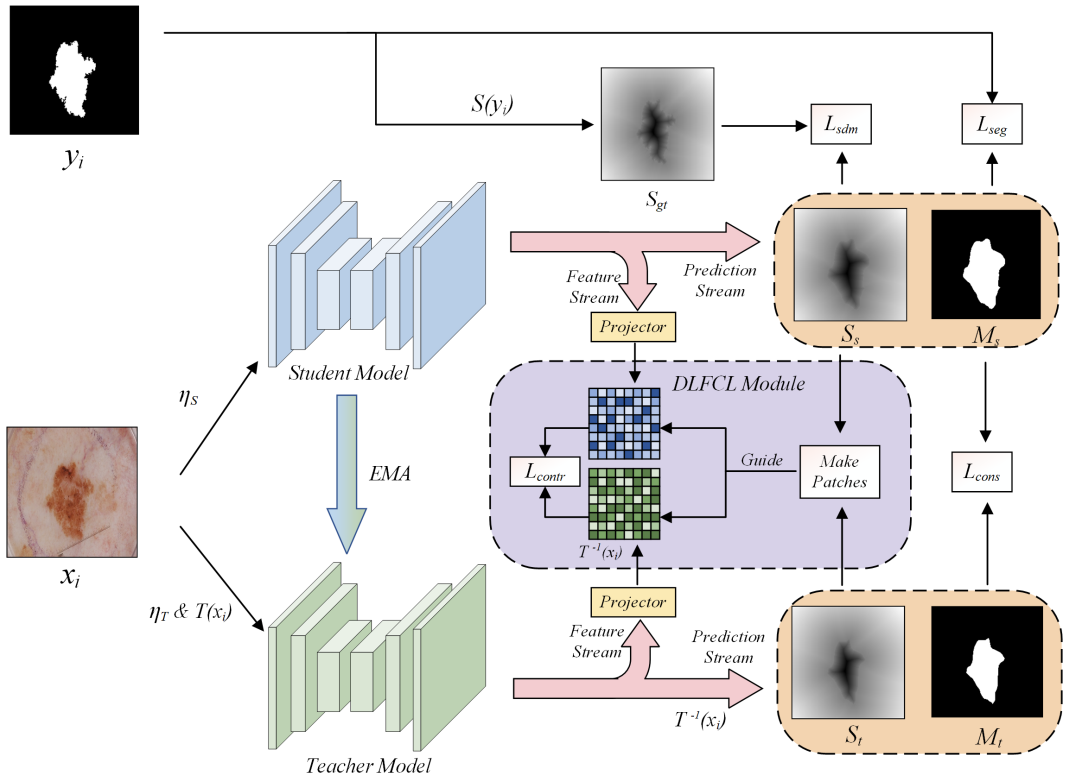

Figure 1. Overview of our framework based on the Mean-Teacher model. The teacher model and student model share the same architecture, the parameters of the student model are optimized by the weighted sum of L_{seg}, L_{sdm}, L_{cons}, and L_{contr}, teacher model is updated by Exponential Moving Average (EMA) of the student model. $T(x)$ represents the addition of perturbations to the model, including rotation and flipping, and $S(y)$ converts the real labels into real Signed Distance Map (SDM).

and allowing the model to output the same contents as much as possible. Laine *et al*. [19] proposed self-ensembling, in which the same data is enhanced twice in sequence into a model, encouraging the output of the model to be as consistent as possible. Tarvainen *et al*. [34] proposed a teacher-student consistency model to encourage different enhanced inputs to have similar outputs in the teacher-student model. Many of the existing semi-supervised medical image segmentation methods rely on consistent regularization, Yu *et al*. [46] proposed an uncertainty-aware framework to guide the calculation of consistency loss by using the uncertainty information provided by the teacher model. Li *et al*. [22] encouraged to produce similar outputs by performing a certain rotation enhancement on the data sent to the student model and teacher model. Hang *et al*. [10] introduced the entropy minimization principle to design the local consistency loss and enforced the global structural consistency by matching the weighted self-information map.Our method combines consistency regularization with contrastive learning and encourages the content of the feature layer of the teacher-student model to be as consistent as possible while adding perturbations to the images.

3. Method

3.1. Ovewview

In this section, we introduce the proposed semi-supervised segmentation network based on dual-stream dense contrastive learning, which effectively utilizes a small amount of labeled data and a large amount of unlabeled data to promote medical image segmentation results.

As shown in Figure 1, we first add various perturbations to images and send them to the student model and the teacher model respectively. Then the predicted SDM and the predicted segmentation map are obtained using the prediction stream while the feature map is obtained by the feature stream. We learn the contrastive loss by DLFCL module, and calculate the consistency loss for the predicted segmentation maps generated by the teacher model and the student model. The weight update method of the teacher model is the Exponential Moving Average (EMA) [34] of the student model, and the specific updating process is as formula 1:

$$\theta_i^t = \alpha\theta_{i-1}^t + (1-\alpha)\theta_{i-1}^s \tag{1}$$

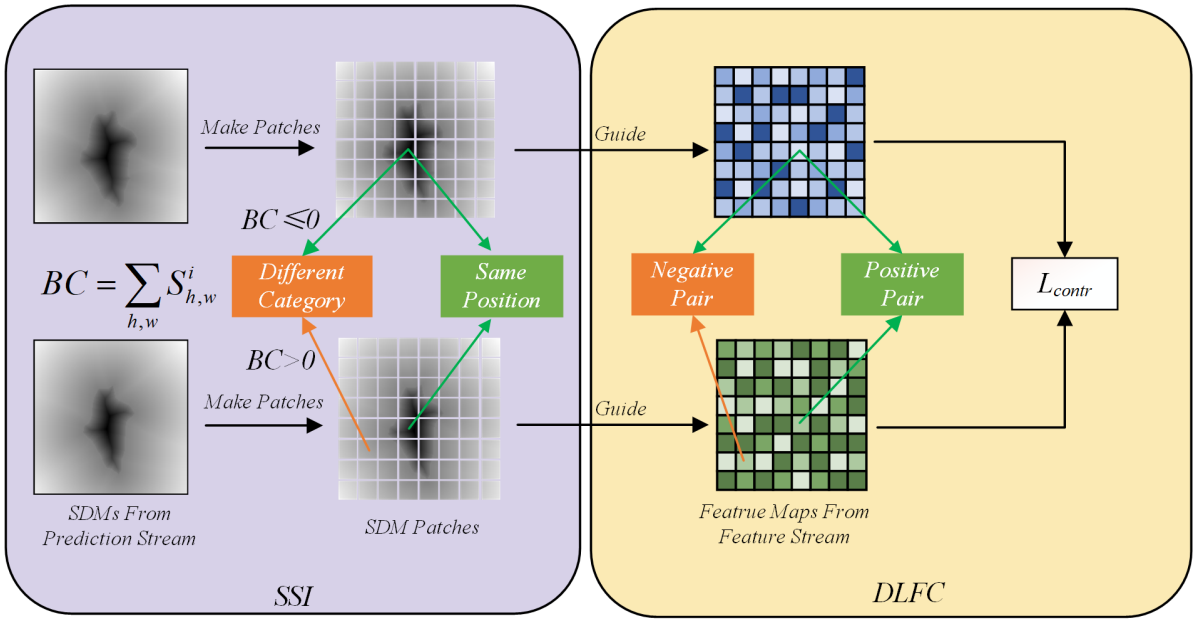

Figure 2. The details of the proposed Dense Local Features Contrastive Learning (DLFCL) module. In the Simpling by Shape Info (SSI), BC is computed for each patch of SDM and negative pairs are determined according to the category of BC. The positive pair is a patch in the same position. Dense Local Feature Contrast (DLFC) optimizes the feature distribution by contrastive learning.

where θ_i^t is the model weight of the teacher model at the ith iteration, θ_i^s is the weight of the corresponding student model, and α is the EMA update weight, which determines the dependence of the teacher model on the student model.

3.2. Base Framework

Mean-Teacher model. The proposed approach is based on the mean teacher model [34], where both the student and teacher models are DenseUnet [21]. In contrast, the Unet [31] utilizes only convolution and pooling and the ResUnet applies residual connections [11]. DenseUnet relies on DenseNet [13] as the backbone to enable feature reuse by connecting features from previous layers. Suppose that the network has l layers and each layer has a nonlinear transformation function H_l, DenseNet takes the features of all previous layers as input by formula 2:

$$x_l = H_l([x_0, x_1, ..., x_{l-1}]) \tag{2}$$

where $[x_0, x_1, ..., x_{l-1}]$ refers to the splicing of features from layer 0 to $l-1$.

Prediction Stream. The prediction stream contains the task of predicting the segmentation map and an additional task of predicting SDM. Since the SDM is an implicit representation of contour information, some existing methods [20, 24] applied SDM to segmentation tasks. Our approach introduces SDM to allow the model to encode richer contour features while guiding the positive and negative pairs sampling process of dense local features. The tasks of generating the predicted segmentation map and the predicted SDM in the prediction stream of our method can be given by formula 3:

$$\begin{aligned} M_i &= f_{seg}(d(x_i), \eta), x_i \in X \\ S_i &= f_{sdm}(d(x_i), \eta), x_i \in X \end{aligned} \tag{3}$$

where $d(\cdot)$ represents the last layer of features extracted by DenseUnet, f_{seg}, f_{sdm} represent the networks corresponding to the predicting segmentation map task and the predicting SDM task, respectively, η represents the perturbations added to the input, and X represents the set of labeled and unlabeled data.

Feature Stream: In our approach, the feature stream is only for unlabeled data. First, the unlabeled data are augmented with a variety of enhancements, including Gaussian noise as well as rotation flipping. They are then fed into the student and teacher models, respectively. The last layer of features extracted from the model is projected into the new dimension through the Projector layer to obtain the feature maps. Mathematically as shown in formula 4,

$$\begin{aligned} P_i^s &= p(d(x_i), \eta_s), x_i \in X_u \\ P_i^t &= p(d(x_i), \eta_t), x_i \in X_u \end{aligned} \tag{4}$$

where $p(\cdot)$ denotes the projector layer, P_i^s and P_i^t are the feature maps obtained by the student model and the teacher model, respectively, η_t and η_s are perturbations of the student model and the teacher model, X_u represents all the unlabeled data.

3.3. DLFCL Module

As shown in Figure 2, the DLFCL module in our approach is divided into two steps. The first step is SSI, which samples the positive and negative pairs in the feature map through the shape information in SDM. The second step is DLFC, which relies on the positive and negative pairs to calculate the contrastive loss.

SSI. Different from the way that existing contrastive learning methods select positive and negative pairs [48], we propose a novel way of sampling positive and negative pairs of dense local features based on SDM guidance, as shown in Figure 2. We perform positive and negative pairs sampling by applying the two feature maps and SDMs obtained from the student model and the teacher model.

The projected feature maps P_i are cut into $n \times n$ fixed-size patches. For the positive pairs, we select the patch at the same location in the projected feature maps of the student model and the teacher model. For negative pairs, our selection method is different from the previous method. In order to measure the difference between each patch, we introduce a metric BC based on the SDM to calculate the variability.

The calculation process of BC is as follows: After obtaining the SDM S_i from the prediction stream, S_i is also cut into multiple fixed-size patches like feature maps P_i, and the BC is obtained by summing the values in each patch with formula 5:

$$BC = \sum_{h,w} S^i_{h,w} \tag{5}$$

where $S^i_{h,w}$ is the value of the element at the current position in the SDM. By summing up each element in the patch of the SDM, it can be considered that the sum of all elements is greater than 0, then it can be considered that the impact of the segmentation target in the patch is greater, on the contrary, if the sum of all elements is less than 0, then the impact of the background in the Patch is greater. Therefore, we divide the patch into two categories according to whether BC is greater than 0, which are called Target Patch (TP) and Background Patch (BP). If a patch is TP, its negative pairs are all the BPs in the current batch. By calculating the BC of each patch, we can complete the positive and negative pairs sampling of dense local features according to the category of BC.

DLFC. After completing the sampling of positive and negative pairs of dense local features, we apply Info-NCE loss function [36] for contrastive loss calculation based on the sampling result as formula 6 and formula 7:

$$L_{contr}(q_i) = -\log \frac{s(q, k_+)}{s(q, k_+) + \sum\limits_{k_- \in \Omega_-} s(q, k_-)} \tag{6}$$

$$s(q, k) = \exp\left(\frac{q \cdot k}{|q| \cdot |k| t}\right) \tag{7}$$

where $s(q, k)$ represents the cosine similarity between features, q and k_+ denote the currently calculated patch with $n \times n$ positive pairs, Ω_- represents the set of selected negative pairs, and t is the temperature hyperparameter. Although we can classify the patches into two categories according to the sign of BC, considering the unbalanced number of pairs in these two categories, therefore we do the process of finding the mean value of L_{contr} before back propagation, and obtain the formula 8:

$$L_{contr} = \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{n*n} \frac{L_{contr}(q_i)}{n^2} \tag{8}$$

where M is the number of unlabeled data.

3.4. Loss Function

We need to define the loss for semi-supervised medical image segmentation. Total loss L_{total} is divided into supervised loss and unsupervised loss as formula 9:

$$L_{total} = L_{sup} + L_{unsup} \tag{9}$$

where L_{sup} is loss with labeled data, L_{unsup} is loss with unlabeled data.

Loss with labeled data. For labeled data, our method generates a predicted segmentation map and a predicted SDM through the prediction stream, so that the supervised loss consists of two parts, L_{seg} and L_{sdm} by formula 10:

$$L_{sup} = L_{seg} + w_{sdm} L_{sdm} \tag{10}$$

where w_{sdm} is the hyperparameters to balance the L_{sdm}.

L_{seg} denotes the segmentation loss, and the cross-entropy loss is computed from the predicted segmentation map and the real segmentation map, which can be expressed as formula 11:

$$L_{seg} = -\frac{1}{N} \sum_{i=1}^{N} y_i \log(f_{s,seg}(d(x_i))) \tag{11}$$

where N is the number of labeled data.

L_{sdm} is the signed distance regression loss, again calculated from the SDM and the true SDM, and to obtain the true SDM, we need to transform the true labels as formula 12:

$$S(x) = \begin{cases} -\inf\limits_{y \in C} \|x - y\|_2, x \in C_{in} \\ 0, \qquad\qquad x \in \partial C \\ +\inf\limits_{y \in C} \|x - y\|_2, x \in C_{out} \end{cases} \tag{12}$$

where $\|x - y\|_2$ is the Euclidean distance between different pixels and y is the pixel on the contour of the target object. $-inf$ and $+inf$ are infimum. C_{in} and C_{out} are the internal and external pixels of the target object, we further normalize each pixel in SDM to $[-1, 1]$.

After obtaining the true SDM, We calculate L_{sdm} by the predicted SDM and the real SDM with formula 13:

$$L_{sdm} = -\frac{1}{N}\sum_{i=1}^{N} \| f_{s,sdm}(d(x_i)) - S(y_i) \|^2 \quad (13)$$

where $S(y_i)$ denotes the true SDM transformed by the true labels.

Loss with unlabeled data. For unlabeled data, our method consists of L_{contr} and L_{cons}. Consequently, L_{unsup} can be expressed as formula 14:

$$L_{unsup} = w_{contr} L_{contr} + w_{cons} L_{cons} \quad (14)$$

where w_{contr} and w_{cons} are the hyperparameters to balance the respective losses, the L_{contr} is detailed in Section 3.3.

L_{cons} is the consistency loss. We enhance the distribution of images at the feature level by using a large amount of unlabeled data through contrastive learning, and we introduce consistency regularization in order to enhance the predicted segmentation map again using unlabeled data. First, we add a certain amount of different Gaussian noise to the input images into the teacher model and the student model, and randomly flip and rotate the input of the teacher model, and after the model outputs the predicted segmentation map, it is again recovered by flipping and rotating it, and the model consistency loss is defined by minimizing the KL divergence of the predicted segmentation map generated by the two models as formula 15:

$$L_{cons} = \frac{1}{M}\sum_{u=1}^{M} f_{s,seg}(d(x_u), \eta_s) \cdot \log \frac{f_{s,seg}(d(x_u), \eta_s)}{T^{-1}(f_{t,seg}(T(d(x_u)), \eta_t))} \quad (15)$$

where $T(\cdot)$ denotes the random flip and rotation of the images, η_s and η_s are the noise added into the student model and teacher model, respectively.

4. Experiments

4.1. Datasets

To validate the effectiveness of our method, we conduct a series of experiments on two publicly available medical image datasets, including a skin lesion segmentation dataset and a gastrointestinal polyp segmentation dataset.

Skin lesion segmentation dataset. The skin lesion segmentation dataset in our experiments is the ISIC 2017 dataset [6], which consists of a training set of 2000 annotated skin images, a validation set of 150 images, and a test set of 600 images, all image resolutions range from 540 × 722 to 4499 × 6748.

Gastrointestinal polyp segmentation dataset. The gastrointestinal polyp segmentation dataset is from Kvasir-SEG [14], which includes 1000 labeled images, the image size is 332×487 to 1920×1072. We randomly divided the data into a training set and a test set, where the training set contains 700 images and the test set is the remaining 300 images.

4.2. Implementation Details and Evaluation Metrics

All experiments are implemented by PyTorch, and the experimental environment is Ubuntu 20.04. We train 20000 iterations on the Tesla V100 GPU, the batch size is 8, the number of patches is set to 8×8, The projector layer is implemented by a Conv → Avgpool → ReLU architecture and the number of channels changed from 64 to 128, the stride of Avgpool is 2. w_{sdm} is set to 0.5, w_{cons} is gradually increased from 0 to 1.0 by Sigmoid curve, and w_{contr} is taken as 0.1. We resize all the images to 224×224 by bicubic interpolation. The network was trained with Adam optimizer [16] with a learning rate of 0.0001. We apply the EMA method [34] to update the teacher model weights, where α is set to 0.999 [22]. DenseUnet [21] is set as the backbone in all comparison experiments for a fair comparison.

To quantitatively evaluate the performance of the model, we adopted six evaluation metrics, including Jaccard Index (JA), Dice coefficient (DI), Mean IOU (MIOU), pixelwise Accuracy (AC), Sensitivity (SE), and Specificity (SP).

4.3. Comparison Experiments

To validate the effectiveness of the proposed method, a comparison is made with several existing state-of-the-art semi-supervised segmentation methods, including MT [34], UAMT [46], TCSMv2 [22], CAC [18], URPC [25], DTC [24] and CDCL [40].

Comparison experiments on the ISIC 2017 dataset. The proposed semi-supervised segmentation method is compared with other methods on the ISIC 2017 dataset. In the training set, we set two different proportions of labeled data to contain 5%(100) labeled data and 15%(300) labeled data, respectively.

As shown in Figure 3, the TCSMv2, URPC and CAC fail to capture the details of melanoma sufficiently, and cannot handle the cases with coarse borders. In comparison, our method can perfectly segment the detailed information and can achieve a more accurate segmentation effect in the case of unclear edges, which is suitable for real labels.

It can be seen from Table 1 that the JA and DI of each semi-supervised segmentation method are greatly improved compared with the only supervised DenseUnet, indicating that unlabeled data has been effectively utilized in each method. In particular, our method outperforms CAC by 0.34% and 0.63% in JA and DI metrics respectively for 5% of labeled data, and MIOU is higher than DTC by 0.88% for 15% of labeled data.

Comparison experiments on the Kvasir-SEG dataset. To verify the generalization of the proposed method, we conduct two sets of semi-supervised segmentation experi-

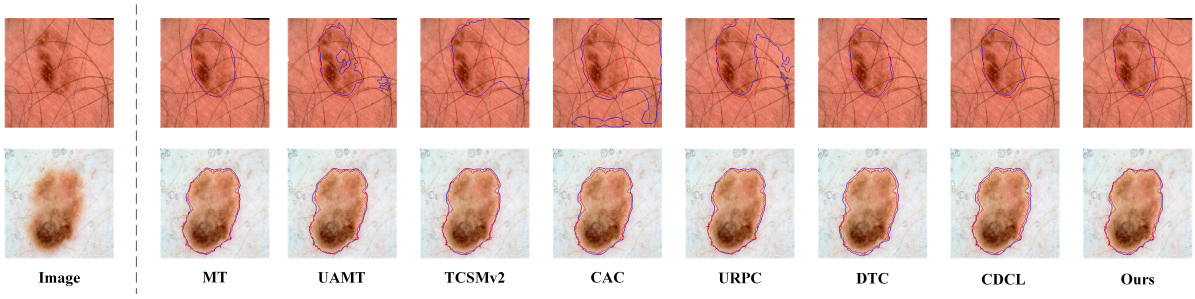

Figure 3. Visualization of different semi-supervised methods for the ISIC 2017 dataset under 15% labeled data, with the red solid line indicating the true label and the blue solid line indicating the segmentation results of each method.

Table 1. Performance of different semi-supervised segmentation methods on the test set of the ISIC 2017 dataset.

Method	Labels	Metrics					
		JA	DI	MIOU	AC	SE	SP
DenseUnet	2000	78.60	86.48	76.94	85.60	88.69	97.21
DenseUnet	100(5%)	74.86	83.22	69.95	93.82	**88.26**	94.46
MT[34]		75.19	83.38	70.35	94.36	84.05	**98.31**
UAMT[46]		75.56	83.83	70.45	93.78	83.83	97.67
TCSMv2[22]		75.77	83.91	71.11	94.30	84.83	96.52
CAC[18]		76.43	84.43	72.41	94.45	84.25	96.66
URPC[25]		75.69	83.90	72.34	92.89	85.84	95.04
DTC[24]		75.84	84.02	71.57	93.92	84.45	97.80
CDCL [40]		76.34	84.48	72.36	94.90	85.54	97.42
Ours		**76.77**	**85.06**	**72.45**	**94.93**	86.14	97.44
DenseUnet	300(15%)	76.80	84.75	71.74	94.47	86.19	96.13
MT[34]		77.05	85.09	73.44	95.04	85.06	**98.41**
UAMT[46]		77.25	83.55	72.45	94.48	85.98	96.51
TCSMv2[22]		77.43	85.47	74.16	94.82	87.97	96.31
CAC[18]		77.48	85.41	74.00	95.63	86.48	95.99
URPC[25]		77.10	85.13	74.62	93.70	84.89	96.13
DTC[24]		77.20	85.08	74.37	93.84	82.33	97.80
CDCL [40]		77.73	85.59	75.03	93.95	83.53	96.89
Ours		**78.05**	**86.31**	**76.04**	**95.45**	**88.70**	96.53

ments on the Kvasir polyp segmentation dataset, containing 10%(70) labeled data and 20%(140) labeled data, respectively.

As shown in Figure 4, three polyps with different shapes and blurred borders were present in the first sample. The MT, TCSMv2 and DTC only segment two polyps. The UAMT, URPC and CDCL produce a mis-segmentation of the region where the third polyp is located. Specially, the proposed method completely segmented the three polyps. For the second sample, the proposed method is smoother at the edges compared to other methods.

According to Figure 4 and Table 2, our method improves the contour awareness of the target by introducing SDM with implicit shape representation, thus outperforming other methods in several metrics. With 20% labeled data, our method outperforms URPC by 0.30% in the JA metric and 1.09% in the MIOU metric than UAMT.

4.4. Ablation experiments

We conduct multiple ablation experiments on the ISIC 2017 dataset to verify the effectiveness of our proposed framework. All experiments contain only 5% (100) of the labeled data, using JA and DI as evaluation metrics.

Ablation experiments on each module. Our method sets DenseUnet as the segmentation network, introduces SDM to enhance contour constraint, and applies dual-stream dense local features contrastive learning and consistency regularization to improve the utilization of unlabeled data.

The above three modules correspond to L_{sdm}, L_{contr} and L_{cons}, respectively. To verify the effectiveness of the above modules, we conduct a series of ablation experiments

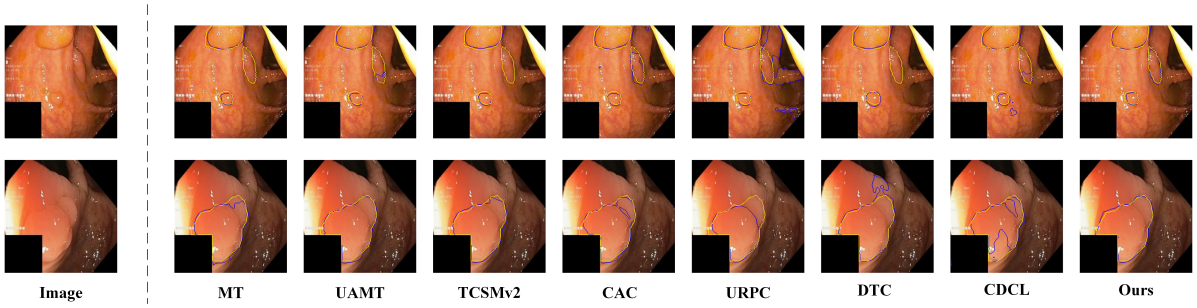

Figure 4. Visualization of different semi-supervised methods for the Kvasir-SEG dataset with 20% labeled data, the solid yellow line indicates the true label and the solid blue line indicates the segmentation results of each method.

Table 2. Performance of different semi-supervised segmentation methods on the test set of the Kvasir-SEG dataset.

Method	Labels	Metrics					
		JA	DI	MIOU	AC	SE	SP
DenseUnet	700	83.83	89.81	81.81	97.11	91.60	98.38
DenseUnet	70(10%)	78.84	86.63	77.65	96.05	93.56	96.88
MT [34]		80.61	87.40	79.76	95.95	89.62	97.64
UAMT [46]		80.56	87.48	79.89	95.95	89.81	97.74
TCSMv2 [22]		80.96	87.47	79.21	96.21	89.41	**97.93**
CAC[18]		80.98	87.42	80.16	96.06	88.83	97.58
URPC [25]		79.79	86.57	78.62	95.64	90.73	96.79
DTC[24]		80.63	87.21	78.37	95.99	90.17	97.73
CDCL [40]		79.41	86.42	78.90	95.78	90.70	96.87
Ours		**81.24**	**88.08**	**79.97**	**96.62**	**93.69**	97.29
DenseUnet	140(20%)	81.46	87.78	79.48	96.24	89.28	98.33
MT [34]		82.01	88.54	79.76	96.30	89.80	98.42
UAMT [46]		82.22	88.59	80.33	96.57	**93.02**	97.35
TCSMv2 [22]		82.40	88.79	80.18	**96.72**	92.00	97.88
CAC[18]		81.97	88.19	80.19	96.13	90.15	97.74
URPC [25]		82.59	88.87	79.37	96.40	90.60	**98.57**
DTC[24]		82.17	88.37	79.28	96.48	88.60	97.65
CDCL [40]		82.44	88.54	79.76	96.30	89.80	98.42
Ours		**82.89**	**88.98**	**81.42**	96.61	90.08	98.50

Table 3. Ablation experiments on each module.

Method	Labels	Metrics	
		JA	DI
DenseUnet	100(5%)	74.86	83.22
DenseUnet w/ L_{sdm}		75.53	84.00
Ours w/o L_{contr}		75.78	83.93
Ours w/o L_{cons}		75.95	84.16
Ours		**76.77**	**85.06**

on their losses, and the results are shown in Table 3. According to Table 3, it can be seen that after the introduction of SDM, the accuracy of the model is effectively improved, with a 0.67% improvement compared to the JA of DenseUnet. The JA with L_{cons} removed alone decreased by 0.82% compared to the original but was 0.42% higher than that with the introduction of SDM only. Finally, removing L_{contr} alone decreases JA by 0.99%, which is also 0.25% higher than introducing only SDM. The above experimental results show that our proposed modules are effective and complementary, and both are effectively using unlabeled data to improve model accuracy.

Ablation experiments on sampling methods. In order to verify the advantages of our proposed method for negative pairs sampling, we conduct ablation experiments on the sampling methods of negative pairs, which are random sampling and all sampling. Specifically, random sampling is to randomly select some patches as negative pairs, while all sampling is to consider all the patches excluding positive pairs as negative pairs, and the experimental results are shown in Table 3. If choose random sampling, JA and DI are down 0.81% compared to our method. For the case of

Table 4. Ablation experiments on the sampling method.

Method	Labels	Metrics	
		JA	DI
Ours w/o L_{contr}	100(5%)	75.78	83.93
Random		75.96	84.25
All		75.46	83.76
Ours		**76.77**	**85.06**

Table 5. Ablation experiments on the patch size.

Method	Labels	Metrics	
		JA	DI
2x2	100(5%)	75.89	84.24
4x4		76.46	84.46
8x8		**76.77**	**85.06**

choosing all sampling, JA and DI decreased by 0.32% and 0.17% respectively than those without contrastive learning. The data in Table 3 shows that the sampling method is crucial in contrastive learning and that the sampling method we propose is effective.

Ablation experiments on patch size. In order to verify the effect of patch size on our proposed method, we conduct relevant ablation experiments, including experiments for the cases of patch number under 2x2, 4x4, 8x8. The experimental results are shown in Table 5, and it can be seen that the highest accuracy is achieved with the number of Patches of 8x8, where JA is greater than 4x4 by 0.31% and DI is greater than 2x2 by 0.82%. Thus the patch size has no significant effect on the overall performance of our method, which indicates that our method works well in contrastive learning of local features at different scales.

5. Conclusion

In this work, we propose a dual-stream dense contrastive learning network for semi-supervised medical image segmentation, which effectively exploits the unlabeled data to achieve more precise segmentation. We rely on the prediction stream to obtain the predicted segmentation map and predicted SDM, and leverage their implicit mutual learning to enhance the utilization of contour features. To address the insufficient feature representation capability, we apply the DLFCL module in the feature stream to improve intra-class compactness and inter-class dispersion by contrastive learning. We also adopt consistency regularization, which effectively improves the contrastive learning efficiency and the robustness of the model. The extensive experimental analysis demonstrates the effectiveness of our proposed method achieves significant improvements against existing state-of-the-art methods. In the future, we will continue to explore the extension of our approach to 3D medical image segmentation and the multi-class case.

6. Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant No. 62076117 and 62166026, the Jiangxi Key Laboratory of Smart City under Grant No. 20192BCD40002 and the Jiangxi Provincial Natural Science Foundation under Grant No. 20224BAB212011.

References

[1] I. Avital, I. Nelkenbaum, G. Tsarfaty, E. Konen, N. Kiryati, and A. Mayer. Neural segmentation of seeding rois (srois) for pre-surgical brain tractography. *IEEE Transactions on Medical Imaging*, 39(5):1655–1667, 2020. 1

[2] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *Advances in Neural Information Processing Systems*, pages 12546–12558, 2020. 2

[3] S. Chen, C. Qiu, W. Yang, and Z. Zhang. Combining edge guidance and feature pyramid for medical image segmentation. *Biomedical Signal Processing and Control*, 78:103960, 2022. 1

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 2

[5] C. Y. Chuang, J. Robinson, Y. C. Lin, A. Torralba, and S. Jegelka. Debiased contrastive learning. In *Advances in Neural Iinformation Processing Systems*, pages 8765–8775, 2020. 1

[6] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging*, pages 168–172. IEEE, 2018. 6

[7] J. Cui, Z. Zhong, S. Liu, B. Yu, and J. Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 715–724, 2021. 1

[8] Z. Feng, Q. Zhou, Q. Gu, X. Tan, G. Cheng, X. Lu, J. Shi, and L. Ma. Dmt: Dynamic mutual training for semi-supervised learning. *Pattern Recognition*, page 108777, 2022. 2

[9] D. Gai, J. Zhang, Y. Xiao, W. Min, Y. Zhong, and Y. Zhong. Rmtf-net: Residual mix transformer fusion net for 2d brain tumor segmentation. *Brain Sciences*, 12(9), 2022. 1

[10] W. Hang, W. Feng, S. Liang, L. Yu, Q. Wang, K.-S. Choi, and J. Qin. Local and global structure-aware entropy regularized mean teacher model for 3d left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 562–571. Springer, 2020. 3

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[12] Z. Hu, Z. Yang, X. Hu, and R. Nevatia. Simple: similar pseudo label exploitation for semi-supervised classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15099–15108, 2021. 2

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. 4

[14] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, pages 451–462. Springer, 2020. 6

[15] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, pages 18661–18673, 2020. 2

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 6

[17] M. D. Kohli, R. M. Summers, and J. R. Geis. Medical image data and datasets in the era of machine learning—whitepaper from the 2016 c-mimi meeting dataset session. *Journal of Digital Imaging*, 30(4):392–399, 2017. 1

[18] X. Lai, Z. Tian, L. Jiang, S. Liu, H. Zhao, L. Wang, and J. Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1205–1214, 2021. 6, 7, 8

[19] S. Laine and T. Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2016. 1, 2, 3

[20] S. Li, C. Zhang, and X. He. Shape-aware semi-supervised 3d semantic segmentation for medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 552–561. Springer, 2020. 2, 4

[21] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE Transactions on Medical Imaging*, 37(12):2663–2674, 2018. 4, 6

[22] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):523–534, 2020. 1, 3, 6, 7, 8

[23] Y. Liu, H. Wang, Z. Chen, K. Huangliang, and H. Zhang. Transunet+: Redesigning the skip connection to enhance features in medical image segmentation. *Knowledge-Based Systems*, 256:109859, 2022. 1

[24] X. Luo, J. Chen, T. Song, and G. Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8801–8809, 2021. 2, 4, 6, 7, 8

[25] X. Luo, W. Liao, J. Chen, T. Song, Y. Chen, S. Zhang, N. Chen, G. Wang, and S. Zhang. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 318–329. Springer, 2021. 2, 6, 7, 8

[26] J. Ma, Z. Wei, Y. Zhang, Y. Wang, R. Lv, C. Zhu, C. Gaoxiang, J. Liu, C. Peng, L. Wang, et al. How distance transform maps boost segmentation cnns: an empirical study. In *Medical Imaging with Deep Learning*, pages 479–492. PMLR, 2020. 2

[27] Y. Ma, Y. Hua, H. Deng, T. Song, H. Wang, Z. Xue, H. Cao, R. Ma, and H. Guan. Self-supervised vessel segmentation via adversarial learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7536–7545, 2021. 2

[28] W. Min, M. Fan, X. Guo, and Q. Han. A new approach to track multiple vehicles with the combination of robust detection and two classifiers. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):174–186, 2018. 1

[29] D. Nie, Y. Gao, L. Wang, and D. Shen. Asdnet: attention based semi-supervised deep networks for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 370–378. Springer, 2018. 2

[30] Y. Ouali, C. Hudelot, and M. Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 2

[31] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 4

[32] Y. Shi, J. Zhang, T. Ling, J. Lu, Y. Zheng, Q. Yu, L. Qi, and Y. Gao. Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(3):608–620, 2021. 2

[33] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, pages 596–608, 2020. 1, 2

[34] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, 2017. 1, 2, 3, 4, 6, 7, 8

[35] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*, pages 6827–6839, 2020. 2

[36] A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. 5

[37] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, A. Solin, Y. Bengio, and D. Lopez-Paz. Interpolation consistency training for semi-supervised learning. *Neural Networks*, 145:90–106, 2022. 2

[38] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. V. Gool. Exploring cross-image pixel contrast for semantic segmentation. In *2021 IEEE/CVF International Conference on Computer Vision*, pages 7283–7293, 2021. 1

[39] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 2

[40] H. Wu, Z. Wang, Y. Song, L. Yang, and J. Qin. Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11666–11675, 2022. 2, 6, 7, 8

[41] Y. Xie, J. Zhang, Z. Liao, J. Verjans, C. Shen, and Y. Xia. Intra-and inter-pair consistency for semi-supervised gland segmentation. *IEEE Transactions on Image Processing*, 31:894–905, 2021. 2

[42] F. Yang, F. Liang, L. Lu, and M. Yin. Dual attention-guided and learnable spatial transformation data augmentation multi-modal unsupervised medical image segmentation. *Biomedical Signal Processing and Control*, 78:103849, 2022. 1

[43] Y. Yang, T. Yan, X. Jiang, R. Xie, C. Li, and T. Zhou. Mh-net: Model-data-driven hybrid-fusion network for medical image segmentation. *Knowledge-Based Systems*, 248:108795, 2022. 1

[44] C. You, Y. Zhou, R. Zhao, L. Staib, and J. S. Duncan. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 2022. 2

[45] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems*, pages 5812–5823, 2020. 2

[46] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–613. Springer, 2019. 3, 6, 7, 8

[47] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 408–416. Springer, 2017. 2

[48] X. Zhao, C. Fang, D.-J. Fan, X. Lin, F. Gao, and G. Li. Cross-level contrastive learning and consistency constraint for semi-supervised medical image segmentation. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022. 1, 2, 5