Graphical Abstract

CCL-MPC: Semi-Supervised Medical Image Segmentation via Collaborative intra-inter Contrastive Learning and Multi-Perspective Consistency

Xiaogang Du, Yibin Zou, Tao Lei, Weichuan Zhang, Yingbo Wang, Asoke K. Nandi

Highlights

CCL-MPC: Semi-Supervised Medical Image Segmentation via Collaborative intra-inter Contrastive Learning and Multi-Perspective Consistency

Xiaogang Du, Yibin Zou, Tao Lei, Weichuan Zhang, Yingbo Wang, Asoke K. Nandi

- Compare the pixels within the same category across various images to discern their analogous intrinsic characteristics
- Enrich positive and negative samples by cross sampling from labeled and unlabeled images
- Leverage certainty pseudo-labels for enhancing the reliability of consistency supervision
- Use the feature differences inside and outside the image can enable more comprehensive representation learning

CCL-MPC: Semi-Supervised Medical Image Segmentation via Collaborative intra-inter Contrastive Learning and Multi-Perspective Consistency

Xiaogang Du[a], Yibin Zou[a], Tao Lei[a,*], Weichuan Zhang[a], Yingbo Wang[a], Asoke K. Nandi[b]

[a]*Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, 710021, China*
[b]*Department of Electronic and Electrical Engineering, Brunel University of London, London, UB8 3PH, United Kingdom*

Abstract

Semi-supervised image segmentation extracts specific regions and tissues by utilizing extensive unlabeled images and limited labeled images, which can alleviate the dependence on plenty of accurately labeled data. However, it is difficult to learn robust feature representations due to the potential noise in pseudo-labels caused by inefficient consistency learning, and poor class diversity in feature spaces. To address this issue, we propose a semi-supervised medical image segmentation method using Collaborative intra-inter Contrastive Learning and Multi-Perspective Consistency (CCL-MPC). First, we propose a collaborative intra-inter contrastive learning strategy that includes symmetric bidirectional contrastive learning and certainty-guided contrastive learning, to exploit the intrinsic differences between inter-image and intra-image feature representations. Second, we design a multi-perspective consistency learning strategy to improve the class diversity by utilizing a dual-branch network and two augmented views. Additionally, we dynamically partition the pseudo-label certainty area for auxiliary consistency learning to reduce the potential noise during the training process. Experimental re-

*Corresponding author

Email addresses: duxiaogang@sust.edu.cn (Xiaogang Du), emarkzou@gmail.com (Yibin Zou), leitao@sust.edu.cn (Tao Lei), zwc2003@163.com (Weichuan Zhang), wangyingbo@sust.edu.cn (Yingbo Wang), asoke.nandi@brunel.ac.uk (Asoke K. Nandi)

sults on the publicly available datasets demonstrate that CCL-MPC can achieve better segmentation performance than the state-of-the-art methods for semi-supervised medical image segmentation tasks. The code will be made available upon publications.

Keywords: Deep learning, medical image segmentation, consistency regularization, contrastive learning

---

1. Introduction

Medical image segmentation is the process of analyzing specific regions or tissues in medical images and extracting their pathological morphology. Medical image segmentation can provide a professional basis for clinical doctors to analyze and diagnose diseases, and make medical treatment plans. Therefore, it plays a very crucial role in clinical healthcare.

1.1. Motivations

In the past decade, with the development of deep learning techniques, scholars have proposed numerous supervised learning-based methods for medical image segmentation [1, 2, 3, 4, 5]. These methods typically rely on a large number of precisely labeled data for model training. However, due to patient privacy protection and expensive costs for medical image annotations, it is usually difficult to collect a large amount of dataset with accurate annotations in practical clinical scenarios. Consequently, it is important to effectively mine the potential knowledge of unlabeled data, particularly in scenarios with limited labeled data.

The semi-supervised medical image segmentation methods [6, 7, 8, 9] usually utilize limited labeled data and extensive unlabeled data for joint training, alleviating the dependence on a large number of precisely labeled samples. Especially, by introducing consistency learning [10, 19, 12, 13] and contrastive learning [14, 15, 16], several semi-supervised learning methods have been widely applied for medical image segmentation. However, due to limited annotations, the current segmentation methods based on semi-supervised learning have poor learning capacity in the early stage of training, resulting in low-quality pseudo-labels and inefficient representation learning. Therefore, we aim to reduce the potential noise during the training process to enhance pseudo-labels and improve the efficiency of representation learning, thereby achieving better segmentation performance.

In complicated medical scenarios, organs and tissues often exhibit a variety of shapes and sizes, along with indistinct boundaries with surrounding tissues, leading to potential noise in pseudo-labels. To address this challenge, medical image segmentation methods often employ contrastive learning [17, 18, 19] to capture more accurate boundary and category features. However, the insufficient number of positive and negative samples in a semi-supervised situation hinders contrastive learning from adequately bringing positive samples closer and pushing negative samples farther, leading to difficulties in capturing robust features. Moreover, many existing methods train labeled and unlabeled data separately [20, 21, 22], thus neglecting the potential relationship between them. In semi-supervised learning, the potential relationship between labeled and unlabeled data is that although different images have different details and segmentation targets, pixels of the same category have similar abstract features in the hidden space. This similarity is based on semantic information, which is essentially consistent even if there are different appearance between different images. Therefore, we propose the Symmetric Bidirectional Contrastive Learning (SBCL) module to exploit this potential relationship.

Also, due to the low quality of pseudo-labels during the early stages of training [23, 24, 25], it becomes more likely for contrastive learning-based methods to sample false negative samples, leading to incorrect feature representations. Therefore, we propose the Certainty-Guided Contrastive Learning (CGCL) module to improve the quality of pseudo-labels.

In addition, medical image segmentation methods often use consistency learning [26, 29, 30, 31] to extract critical features in complicated medical segmentation scenarios. Consistency learning methods introduce various perturbations and regularizations to obtain consistent predictions. This process relies on the high quality and diversity of training data. However, in semi-supervised scenarios, due to the potential noise of pseudo-labels, it is difficult to ensure high-quality and diverse data, resulting in low representation learning and segmentation performance. To overcome these, we propose a Multi-Perspective Consistency (MPC) strategy to utilize network architecture and differences between various views, with auxiliary supervision from certainty pseudo-labels, aiming to improve the effectiveness of representation learning.

1.2. Contributions

In this paper, we propose a semi-supervised medical image segmentation method based on Collaborative intra-inter Contrastive Learning and Multi-Perspective Consistency, namely CCL-MPC. The main contributions are summarized as follows:

- We propose the SBCL module to focus on the potential of inter-image feature learning. In this module, we design a novel symmetric bidirectional sampling mechanism for cross-sampling both labeled and unlabeled data, aiming to learn inherent features across images.
- We propose the CGCL module to focus on the potential of intra-image feature learning. The CGCL module uses certainty pseudo-labels and top-K methods to guide more reliable positive and negative sampling, thereby preventing learning incorrect features caused by potential noise, and improving the quality of positive and negative samples.
- We propose the MPC strategy to improve data diversity and to suppress potential noise that may mislead training. The MPC utilizes a dynamically adaptive threshold to categorize high certainty regions as certainty pseudo-labels, ensuring consistent predictions from two augmented views under certainty pseudo-labels auxiliary supervision.

The rest of this paper is organized as follows. In Section II, the related works of contrastive learning and consistency learning are investigated. In Section III, the proposed CCL-MPC are described. In Section IV, experimental results are presented and discussed. Finally, Section V concludes the paper.

2. Related Work

2.1. Consistency Learning

Consistency learning assumes that an output of network should remain similar before and after samples are transformed or perturbed. Therefore, consistency learning utilizes various perturbations to train data and applies regularization to ensure consistent predictions. According to different perturbation patterns, consistency learning can be coarsely categorized into data perturbation-based and network perturbation-based consistency learning.

Data perturbation-based consistency learning introduces data diversity to enable the model to adapt to different input samples and improve its generalization capacity. Mean Teacher (MT) [10] is one of the most representative data perturbation-based consistency learning methods. This method applies different noise to unlabeled data and constrains prediction results, and introduces the supervision loss on labeled data for joint learning to prevent overfitting on a small amount of labeled data. Furthermore, Li et al. [26] proposed different data perturbation ways (noise, rotation, scaling, etc.) and effectively utilized unlabeled data by introducing a transformation-consistent strategy. Liu et al. [27] designed a Perturbed and Strict Mean Teachers (PS-MT) framework to improve segmentation accuracy by extending the MT model, which includes a new auxiliary teacher, diverse loss functions, and adversarial perturbation.

Network perturbation-based consistency learning can make the model more robust and less susceptible to small changes or noise in input samples by introducing different perturbations, such as changing weights, and architecture, or introducing regularization. Chen et al. [28] proposed a Cross Perturbation Supervision (CPS) method based on network perturbation to encourage high consistency between the predictions of two perturbed networks. However, it may lead to unreliable guidance and unstable training by calculating consistency between different predictions on unlabeled data. To address this issue, Yu et al. [29] designed an Uncertainty-Aware Mean Teacher (UA-MT) framework, which enables the student model to gradually learn more reliable targets by using uncertainty estimations. To reduce time and memory overhead, Wu et al. [30] devised a Mutual Consistency Network (MC-Net). MC-Net encourages two decoders to output consistent predictions, which can gradually capture generalized features from unlabeled challenging regions.

However, current consistency learning often imposes high requirements on the quality and diversity of training data. If the training data is biased or insufficiently diverse, consistency learning may not significantly improve the segmentation performance. To effectively enhance data diversity, we employ a dual-branch heterogeneous predictor architecture and two augmented views to design richer feature sources, and utilize certainty pseudo-labels for auxiliary supervision to achieve more reliable consistency learning.

2.2. Contrastive Learning

Contrastive learning aims to learn distinctive feature representations to enlarge the margin between different classes in the feature space, thereby distinguishing between positive and negative samples to enhance intra-class cohesion and inter-class separation. It requires a sufficient number of negative samples to learn the discriminative feature representations from different classes. Therefore, Zhao et al. [6] extracted local features and treated different pixel positions of an image as negative samples, effectively expanding the number of negative samples. Wu et al. [14] proposed a dual contrastive learning method based on anatomical structure to enhance the discriminability of learned features and the data utilization for few-shot segmentation. Lou et al. [20] input unlabeled data after data augmentation into a dual-branch network to increase the number of negative samples for model training. This method focuses on the pixel position relationship among unlabeled data, but ignores the potential for sampling rich features within an image.

Meanwhile, it is also necessary to pay attention to the false negative sample rate during sampling to avoid affecting the quality of negative samples and learning misleading feature representations. Therefore, to improve the quality of negative samples, Zhao et al. [23] proposed Rectified Contrastive Pseudo Supervision (RCPS), which employs various intensity transformation techniques and optimizes pseudo-labels using KL divergence. Wu et al. [15] used the teacher-student model to construct affinity maps of different branches for affinity contrastive learning, improving the quality of pseudo-labels. Wang et al. [8] proposed Uncertainty Guided Pixel Comparative Learning (UGPCL), which utilizes uncertainty maps to optimize pseudo-labels and guide contrastive learning. Basak et al. [16] generated multi-category pseudo-labels through teacher-student model, then obtained the class-perception matrix to guide patch-level contrastive learning. However, the above methods focus more on improving the confidence of pseudo-labels, ignoring the equally important impact of sample quantity on improving the performance of contrastive learning.

Different from these existing works, we conduct pixel-level contrastive learning on limited data from both inter-images and intra-images, and implement a new symmetric bidirectional sampling mechanism based on the symmetry of the dual-branch network, effectively increasing the number of positive and negative samples. Meanwhile, we dynamically identify high-certainty areas to guide more accurate sampling, which improves the quality

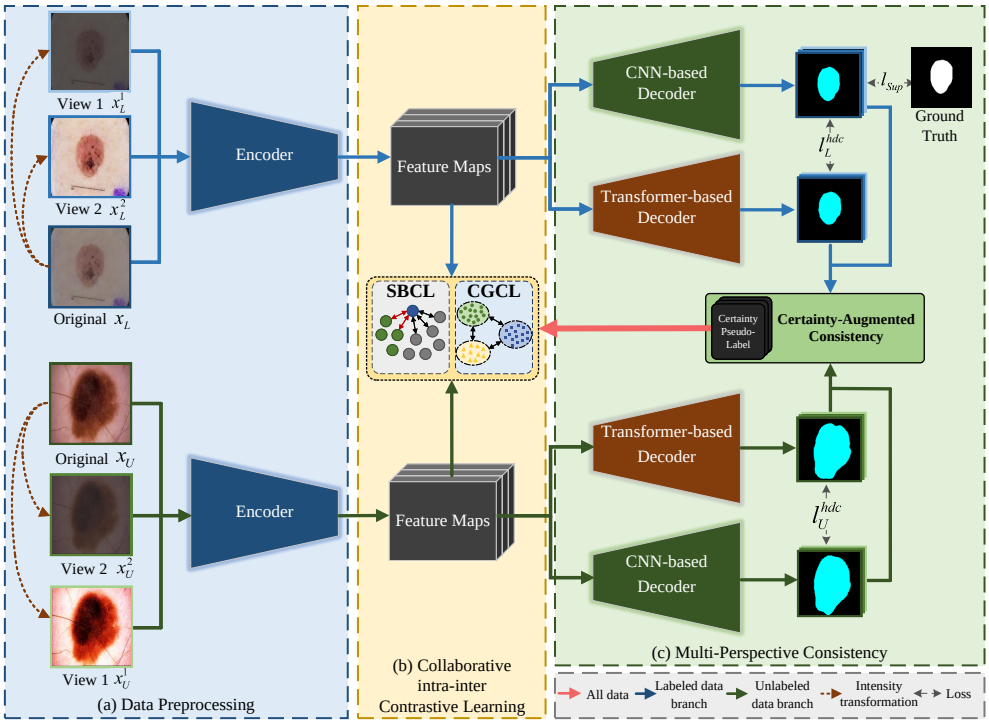

Figure 1: The framework of the proposed CCL-MPC, which builds on the baseline U-Net and Swin-UNet and applies CCL and MPC strategies. (a) Data preprocessing. Random intensity transformations are utilized to generate two augmented views, which can enrich sample diversity. (b) CCL. The encoder features are input into the SBCL and CGCL modules to exploit the inherent differences from inter-images and intra-images. (c) MPC. We apply consistency constraints to the results of the three views through the heterogeneous predictor, to leverage the diversity of different perspectives. Simultaneously, the two results of the heterogeneous predictor are input into CAC.

of positive and negative samples and obtains more reliable feature representations.

3. Method

3.1. Overall Structure

The overall architecture of the proposed CCL-MPC is illustrated in Fig. 1. CCL-MPC is a dual-branch network, which includes the labeled and unlabeled data branches. These two branches adopt encoder-decoder structures.

The encoder utilizes the same CNN-based encoder. The decoder employs a Transformer-based decoder and a CNN-based decoder, respectively.

Initially, the labeled data x_L and unlabeled data x_U apply different random intensity transformations to generate two augmented views, denoted as x_L^1, x_L^2 and x_U^1, x_U^2, respectively. Subsequently, these augmented views and the original images are input into the dual-branch network, proceeding through encoders E_L and E_U to extract deep features. The extracted deep features are input into both MPC and CCL.

In the MPC, we design two consistency learning strategies, including Heterogeneous Decoding Consistency (HDC) and Certainty-Augmented Consistency (CAC), to achieve consistency predictions from multiple perspectives, thereby enriching the quality and quantity of features. For CAC, we dynamically generate certainty pseudo-labels and establish consistency constraints between certainty pseudo-labels and the augmented views and between the augmented views, to facilitate consistency learning and achieve more robust feature representations. For HDC, we design a heterogeneous predictor that predicts feature maps of three different augmented views from the same encoder. Thereafter, we apply consistency constraints between the prediction results, aiming to leverage the inherent differences between Transformer and CNN structures to capture more essential features.

In the CCL, we introduce two contrastive learning modules at different levels to improve the quantity and quality of samples, aiming to achieve better intra-class cohesion and inter-class separation. Firstly, we establish the SBCL module. SBCL takes the same spatial location of features maps Z_L^1, Z_L^2 or Z_U^1, Z_U^2 as positive pairs, and obtains higher quality negative samples by the symmetric bidirectional sampling mechanism for pixel-level contrastive learning. Secondly, we establish the CGCL module. CGCL utilizes the certainty pseudo-labels to perform high-confidence positive and negative sample sampling in the embedding space for pixel-level contrastive learning.

3.2. Multi-Perspective Consistency

At present, consistency learning relies heavily on the high quality and diversity of training data, but there are often two problems of data limitations and potential noise in semi-supervised learning. Therefore, to improve data quality and diversity, we propose the MPC, including the HDC and CAC, which utilizes data from multiple perspectives for consistency learning.

3.2.1. Heterogeneous Decoding Consistency

We design a heterogeneous predictor by incorporating a Transformer-based decoder and a CNN-based decoder. This predictor obtains feature representations from different perspectives by leveraging the local capturing capacity of CNNs and the global context awareness of Transformers. This heterogeneous structure could improve the feature representation capacity of the model, especially to improve the feature diversity learning of the model.

In this predictor, the feature maps captured by the encoder are input into the Transformer-based decoder and CNN-based decoder, respectively to obtain different prediction results. Consistency constraints are applied to the prediction results to generate more similar predictions. Due to the inherent complementary in architecture and processing ways between the Transformer-based and CNN-based decoders, heterogeneous decoding consistency learning can extract more comprehensive and robust feature representations, leading to better segmentation performance.

For the labeled data, prediction probability distributions $\{p_t^i, i = 0, 1, 2\}$ and $\{p_c^i, i = 0, 1, 2\}$ are obtained through the Transformer-based and CNN-based decoder, respectively. We calculate the consistency loss of the probability distributions obtained from the two decoders:

$$l_{hdc}^L = -\frac{1}{3}\sum_{i=1}^{3} L_{dis}(p_c^i, p_t^i), \tag{1}$$

where L_{dis} represents the distance measurement between the two probability distributions. Mean Squared Error (MSE) is used as the distance measure. Similarly, for the unlabeled data, we generate two probability distributions $\{p_t^j, j = 0, 1, 2\}$ and $\{p_c^j, j = 0, 1, 2\}$, and calculate the consistency loss l_{hdc}^U of the unlabeled data branch.

Due to the identical consistency of the labeled and unlabeled data branches, we calculate the total consistency loss as:

$$l_{hdc} = \frac{1}{2}(l_{hdc}^L + l_{hdc}^U). \tag{2}$$

3.2.2. Certainty-Augmented Consistency

During the data processing stage, intensity transformations are applied to labeled and unlabeled data to generate augmented views. The intensity transformations include intensity scaling, intensity offset, and Gaussian noise.

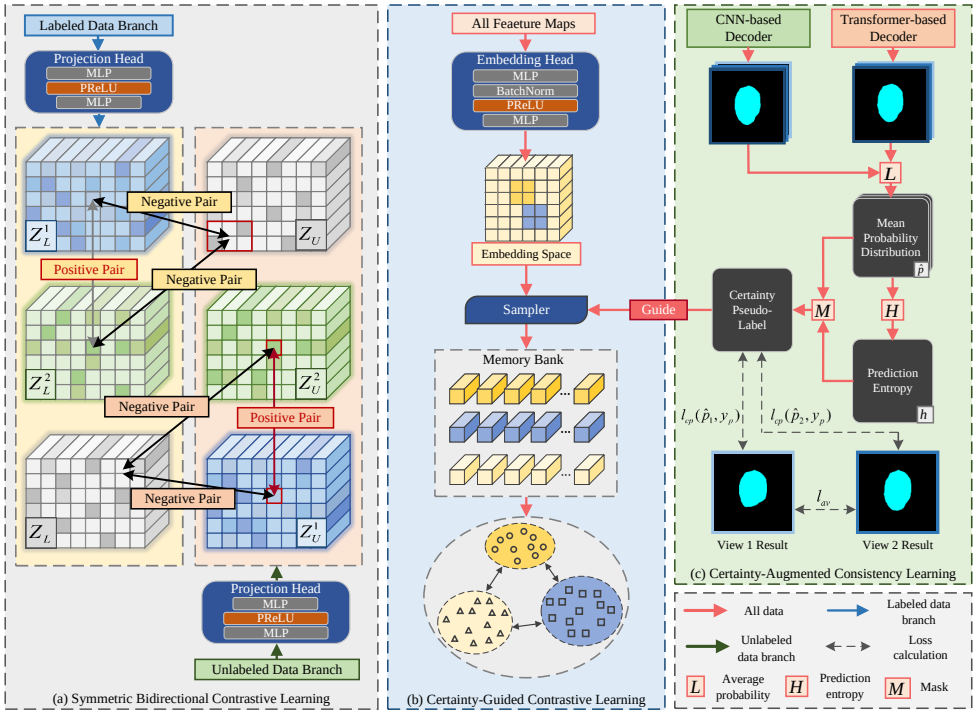

Figure 2: The diagram illustrates the CCL and Certainty-Augmented Consistency. (a) SBCL: Pixel-level contrastive learning for inter-images. We cross-sample three views of labeled and unlabeled data to capture more intrinsic features. (b) CGCL: Pixel-level contrastive learning for intra-images. We use certainty pseudo-labels for auxiliary sampling and the top-K method to reduce potential noise during the training process. (c) The workflow of Certainty-Augmented Consistency Learning. We use certainty pseudo-labels for auxiliary supervision and perform consistency regularization between two views to improve feature quality in consistency learning.

These transformations are designed to create the augmented views of the original image, enhancing the robustness of our model to intensity variations. Although these transformations diversify appearances, predictions from the original input and augmented views should remain similar. Therefore, we impose a consistency constraint on the predictions of the two augmented views to capture robust feature representations. Certainty-augmented consistency learning is illustrated in Fig. 2(c).

Cosine distance is employed to measure the similarity between the predictions of the two augmented views. The predictions of the two augmented views are denoted as $\hat{y}_1$ and $\hat{y}_2$, respectively. The loss function for the labeled

data branch can be denoted as:

$$l_{av}^{L}(\hat{y}_1, \hat{y}_2) = 1 - \frac{\hat{y}_1 \cdot \hat{y}_2}{\|\hat{y}_1\|_2 \cdot \|\hat{y}_2\|_2}. \tag{3}$$

Similarly, the loss $l_{av}^{U}(\hat{y}_1, \hat{y}_2)$ is calculated for the unlabeled data branch. Additionally, consistency constraints are applied to the pseudo-labels of the original image and the prediction results of the two augmented views. However, due to limitations of learning capacity of the model during the early training stages, the pseudo-labels may introduce noise. To achieve more robust feature representations, we apply uncertainty-guided correction to the pseudo-labels.

Predictive entropy is adopted as a metric to approximate the uncertainty. Firstly, the mean probability distribution of the prediction results $\hat{p} = (p_c + p_t)/2$ is computed, where p_c and p_t represent the prediction results obtained by the CNN-based decoder and Transformer-based decoder, respectively. Next, the predictive entropy h of the probability distribution for each pixel along the channel dimension is calculated. The greater value of the entropy means the stronger uncertainty of the predicted category. Therefore, a threshold T is set and regions are treated as uncertainty regions if $h \leq T$. We dynamically adjust the uncertainty threshold T and initially only retain high certainty regions of pseudo-labels for training. As the number of iterations increases, T is gradually decreased to include broader pseudo-label coverage. It not only avoids the impact of early high noise pseudo-labels, but also fully utilizes the potential of unlabeled data.

Finally, we mask the uncertain regions and obtain the certainty pseudo-labels y_p:

$$y_p = Arg\ max(\hat{p})_{|h \leq T}, \tag{4}$$

where h is calculated as:

$$h = -\sum_{n} \hat{p}_n log(\hat{p}_n + \varepsilon), \tag{5}$$

where ε is a very small constant to avoid singularity.

We use the certainty pseudo-labels y_p to constrain consistency with the mean probability distribution $\hat{p}_1$ and $\hat{p}_2$ of the predicted results from two views. The consistency constraints implicitly reduce the cosine distance between predictions from different views, yielding more robust feature representations. In MPC, the temperature scaling technique is employed. This

technique divides the certainty pseudo-labels by a temperature coefficient τ before applying the Softmax operation. This technique smoothens the output probability distribution, which makes the model more cautious in dealing with uncertain situations. The consistency constraint loss for the labeled data branch is formulated as:

$$l_{cp}^{L}(\hat{p}_1, y_p) = L_{CE}(\sigma(\hat{p}_1), \sigma(y_p/\tau)), \tag{6}$$

where L_{CE} denotes the cross-entropy loss, σ represents the Softmax operation. Similarly, the loss $l_{cp}^{L}(\hat{p}_2, y_p)$ is calculated for another augmented view. In addition, these two losses should also be calculated for the unlabeled data branch, denoted as $l_{cp}^{U}(\hat{p}_1, y_p)$ and $l_{cp}^{U}(\hat{p}_2, y_p)$.

Finally, the total consistency loss is a combination of the consistency losses between the three different views of the labeled and unlabeled branches:

$$\begin{aligned} l_{cac} = {} & l_{av}^{L}(\hat{y}_1, \hat{y}_2) + l_{av}^{U}(\hat{y}_1, \hat{y}_2) + l_{cp}^{L}(\hat{p}_1, y_p) \\ & + l_{cp}^{L}(\hat{p}_2, y_p) + l_{cp}^{U}(\hat{p}_1, y_p) + l_{cp}^{U}(\hat{p}_2, y_p). \end{aligned} \tag{7}$$

3.3. Collaborative intra-inter Contrastive Learning

To improve the quantity and quality of positive and negative samples, we design the CCL strategy, which leverages the complementarity of inter-image and intra-image feature representations. This strategy mainly includes the SBCL and CGCL modules. The SBCL and CGCL modules complement each other through feature learning at different levels. The inter-image features provided by SBCL enhance the diversity of pseudo-labels within the image, while the high-quality pseudo-labels provided by CGCL improve reliability for inter-image feature learning. The two modules support each other in the feature embedding space, achieving high cohesion within categories and strong separation between categories through a more comprehensive contrastive learning strategy.

3.3.1. Symmetric Bidirectional Contrastive Learning

The effectiveness of contrastive learning is influenced by the quantity and quality of negative samples. Therefore, leveraging the characteristics of the dual-branch network structure, we propose a symmetric bidirectional sampling mechanism to improve both the quantity and quality of positive and negative samples. Building upon this sampling mechanism, we design the SBCL module, as illustrated in Fig. 2(a).

The symmetric bidirectional sampling mechanism relies on the dual-branch network for cross-sampling between the labeled and unlabeled data branches, and applies a confidence-based sampling for negative samples. Firstly, we choose pixels from an augmented view as anchors. Then, the pixels at the same position with anchors from another augmented view are regarded as positive samples. Finally, for each branch, the original image of the other branch is used as the source of negative samples. We exclude pixels in the same category as the anchor, and then sample negative samples according to prediction confidence by selecting the top-K most confident samples. This mechanism can improve the quantity and quality of positive and negative samples.

Furthermore, we propose the SBCL module. Initially, the features extracted from encoder are input into the projection head, obtaining the corresponding feature maps Z_L, Z_L^1, Z_L^2 and Z_U, Z_U^1, Z_U^2. After that, we obtain positive and negative samples for the labeled and unlabeled data branches.

For the labeled data branch, the pixel sampled from the augmented view $\psi_L^1 \in Z_L^1$ serves as the anchor, and the positive samples $\psi_L^2 \in Z_L^2$ are selected from another augmented view. The negative samples $\psi_n \in Z_U (n = 1, 2, ..., N)$ are obtained from the feature map of the original image of unlabeled data. The contrastive loss for this anchor is calculated as:

$$l_c(\psi_L^1, \psi_L^2) = -log \frac{\exp(\cos(\psi_L^1, \psi_L^2)/\tau)}{\exp(\cos(\psi_L^1, \psi_L^2)/\tau) + \sum_{\psi_n \in Z_U, n=1}^{N} \exp(\cos(\psi_L^1, \psi_L^2)/\tau)}, \quad (8)$$

where τ is the temperature coefficient. Similarly, the loss $l_c(\psi_L^2, \psi_L^1)$ is calculated while the pixel of another augmented view serves as the anchor.

Finally, the aforementioned symmetric bidirectional contrastive loss is calculated for each pixel in the feature maps. We introduce this symmetric design in the calculation of contrastive loss. The core idea is to ensure consistency in the comparison between the two augmented views. Therefore, the total loss for the labeled data branch is calculated as:

$$l_{sbcl}(Z_L^1, Z_L^2, Z_U) = \frac{1}{N_L^1} \sum_{\psi_L^1 \in Z_L^1} l_c(\psi_L^1, \psi_L^2) + \frac{1}{N_L^2} \sum_{\psi_L^2 \in Z_L^2} l_c(\psi_L^2, \psi_L^1), \quad (9)$$

where N_L^1 and N_L^2 represent the number of pixels in the feature maps Z_L^1, Z_L^2 of the two augmented views in the labeled data branch, respectively. Similarly, the loss $l_{sbcl}(Z_U^1, Z_U^2, Z_L)$ is calculated for the unlabeled data branch.

Therefore, the total loss for SBCL is defined as:

$$l_{sbcl} = \frac{1}{2}(l_{sbcl}(Z_L^1, Z_L^2, Z_U) + l_{sbcl}(Z_U^1, Z_U^2, Z_L)). \tag{10}$$

3.3.2. Certainty-Guided Contrastive Learning

To fully leverage the feature representations within each view and obtain more generalized and robust features, we maximize agreement within positive pairs and minimize agreement within negative pairs in the learned embedding space to achieve better intra-class cohesion and inter-class separation. The proposed CGCL module is illustrated in Fig. 2(b).

Firstly, the features extracted from the encoder are input into the embedding head. The encoder features are embedded into a D-dimensional space to obtain prototype feature vectors for each pixel. These prototype feature vectors are respectively saved into the vector sets X_L, X_L^1, X_L^2 and X_U, X_U^1, X_U^2. These sets are used for subsequent positive and negative sample sampling.

Secondly, to improve the quality of negative samples and reduce the false negative rate, we select pixels with high-confidence as anchors through utilizing the certainty pseudo-labels. Specifically, the certainty pseudo-labels are downsampled to the size of the prototype vector sets, assigning a category to each pixel in the prototype vector. In the vector sets, the most top-K confident samples are selected, and negative pixels are sampled based on the prediction confidence.

Then, in contrastive learning, a large number of samples can introduce considerable costs. To reduce costs in the CGCL, a memory queue is employed to store the prototype vector sets and their corresponding pixel categories. The memory queue is continuously updated during the model training process. This mechanism helps to reduce computational costs and enables the model to learn from a large number of samples while maintaining focus on high-confidence samples. The popular InfoNCE [38] is used to calculate the contrastive loss. In each iteration, we choose M anchors according to the prediction confidence and calculate the contrastive loss for each anchor. In the labeled data branch, the loss for each anchor is formulated as:

$$l^i = -\frac{1}{|P_L^i|} \sum_{e_i^+ \in N_L^i} log \frac{\exp(\cos(e_i, e_i^+)/\tau)}{\exp(\cos(e_i, e_i^+)/\tau) + \sum_{e_i^- \in N_L^i} \exp(\cos(e_i, e_i^-)/\tau)}, \tag{11}$$

where e_i represents the anchor, P_L^i and N_L^i denote the prototype vector sets for positive and negative samples of the anchor, respectively. e_i^+ is a positive

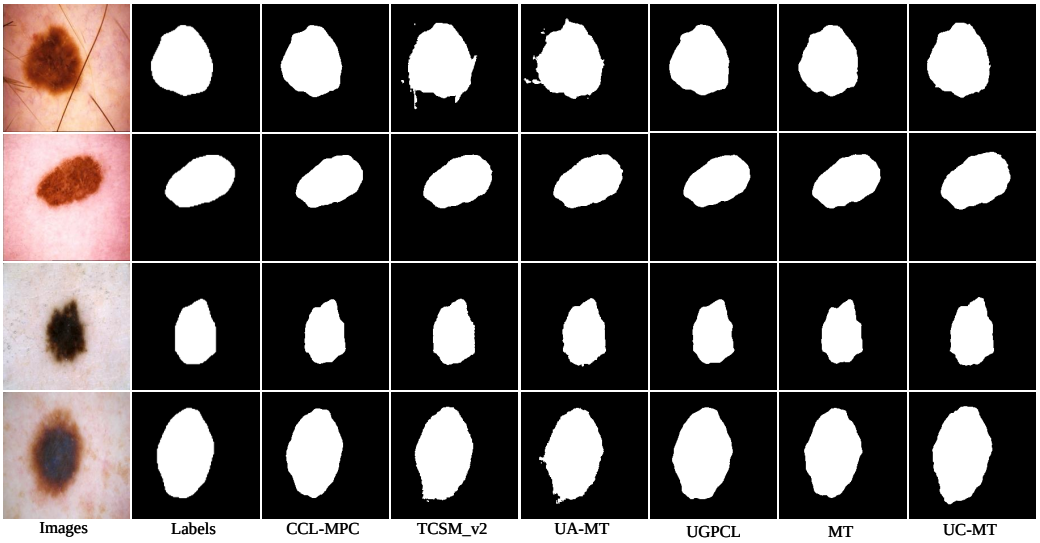

Figure 3: Visualization results generated by different methods on the ISIC2018 dataset. CCL-MPC yields the most favorable predictions.

sample, e_i^- is a negative sample. Finally, we average the loss of all anchors to obtain the total contrastive loss for the labeled data,

$$l_{cgcl}^{L} = \frac{1}{M} \sum_{i=1}^{M} l^i. \tag{12}$$

Similarly, the loss l_{cgcl}^{U} for the unlabeled data is used to calculated. Finally, the overall loss for CGCL is formulated as:

$$l_{cgcl} = \frac{1}{2}(l_{cgcl}^{L} + l_{cgcl}^{U}). \tag{13}$$

3.4. Training Objective

The total loss of CCL-MPC is the summation of the supervised loss, heterogeneous decoding consistency loss, certainty-augmented consistency loss, symmetric bidirectional contrastive learning loss, and certainty-guided contrastive learning loss. The total loss of CCL-MPC is given:

$$l_{total} = l_{sup} + \lambda_1 l_{hdc} + \lambda_2 l_{cac} + \lambda_3 l_{sbcl} + \lambda_4 l_{cgcl}. \tag{14}$$

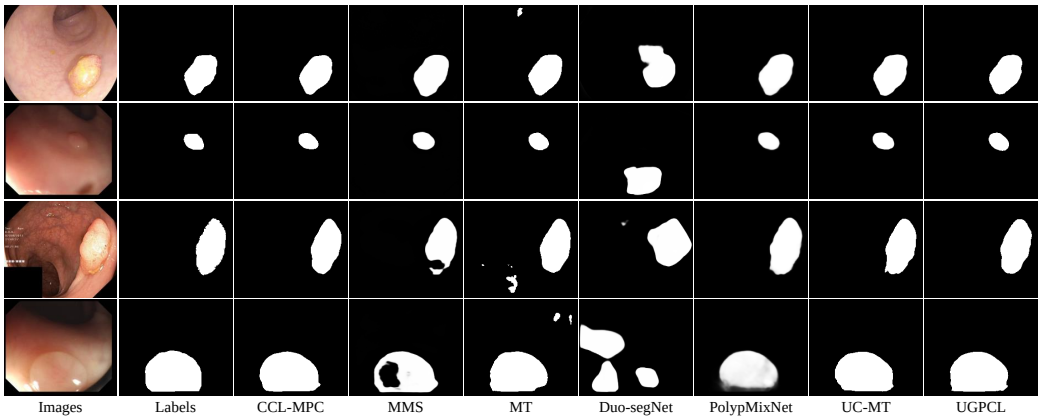

Figure 4: Visualization results generated by different methods on the CVC-300 dataset. CCL-MPC yields the most favorable predictions.

$\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 0.1$ is set to balance the weights of the losses. Additionally, the supervised segmentation loss l_{sup} is calculated using the Ground Truth for the labeled data:

$$l_{sup} = l_{sup}(p_c, y) + \alpha l_{sup}(p_t, y), \tag{15}$$

where p_c and p_t are the prediction results of CNN-based and Transformer-based decoders, respectively. α is a weighting coefficient and is set to 0.4. y is the Ground Truth. Additionally, l_{sup} is a combination of cross-entropy and Dice loss. In the inference stage, the prediction of the CNN-based decoder is the final segmentation result.

4. Experiments

4.1. Datasets

To evaluate the effectiveness of the proposed CCL-MPC, we conduct comparative experiments and ablation studies with different labeled data ratios on different medical image segmentation tasks, including skin lesion segmentation from dermoscopy images and polyp segmentation from colonoscopy images.

4.1.1. Skin lesion dataset

The skin lesion segmentation dataset is from the 2018 International Skin Imaging Collaboration (ISIC2018) [33]. The ISIC2018 dataset consists of

3694 images. The image size ranges from 500×500 to 6661×4422. Specifically, 2594 images are used as the training dataset, 100 images are used as the validation dataset, and 1000 images are used as the testing dataset.

4.1.2. Polyp datasets

The polyp segmentation datasets include: Kvasir-SEG [34], CVC-ClinicDB [35], and CVC-300 [36]. Similar to previous methods [37], the training dataset includes 1450 images, 900 images from Kvasir-SEG and 550 images from CVC-ClinicDB. Some details of these datasets are provided as follows.

(1) *Kvasir-SEG*. This dataset is part of the Medical Multimedia Challenge and contains 1000 images of varying sizes, ranging from 332×487 to 1920×1072. We use 900 images for training and 100 images for testing.

(2) *CVC-ClinicDB*. This dataset comes from 31 colonoscopy sequences and contains 612 images with a size of 384×288. We use 550 images for training and 62 for testing.

(3) *CVC-300*. This dataset comes from 44 colonoscopy sequences and contains 912 images with a size of 574 × 500. We use 60 images for testing.

4.2. Evaluation metrics

To comprehensively evaluate the effectiveness of CCL-MPC, we use the following three widely used evaluation metrics: Mean Dice ($mDice$), Mean Intersection over Union ($mIoU$), and Mean Absolute Error (MAE). *Dice* coefficient is a metric for evaluating segmentation quality, which measures the similarity between segmentation results and the Ground Truth. $mDice$ is the average value of *Dice* coefficient for all samples. Specifically, $mDice$ is calculated as:

$$mDice = \frac{1}{n}\sum_{i=1}^{n}\frac{2 \times |X_i \cap Y_i|}{|X_i| + |Y_i|}, \tag{16}$$

where n is the total number of samples, X_i is the predicted area of the ith sample, Y_i is the Ground Truth of the ith sample, $|X_i \cap Y_i|$ is the number of pixels in the intersection of predicted area and the Ground Truth, $|X_i|$ and $|Y_i|$ represent the number of pixels in the predicted area and the Ground Truth, respectively. $mDice$ is between 0 and 1. The closer the value is to 1, the higher the segmentation accuracy.

IoU calculates the ratio of intersection to union between segmentation results and true labels. $mIoU$ is the average value of IoU for all samples.

Specifically, $mIoU$ is calculated as:

$$mIoU = \frac{1}{n}\sum_{i=1}^{n}\frac{|X_i \cap Y_i|}{|X_i \cup Y_i|}, \tag{17}$$

where $|X_i \cup Y_i|$ represents the number of pixels in the union of predicted area and the true area. The closer the value is to 1, the higher the segmentation accuracy.

MAE is widely used to measure the average distance between the segmentation results and the Ground Truth. Specifically, MAE is calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|X_i - Y_i|. \tag{18}$$

The smaller the MAE value, the better the segmentation accuracy.

4.3. Implementation Details

The experiments are conducted using PyTorch on a single Nvidia GeForce RTX 3090 GPU. For fairness, all methods choose U-Net as the benchmark architecture for image segmentation. ResNet-50 is employed to substitute the encoder of U-Net, initializing its parameters with pre-trained weights on ImageNet. The Stochastic Gradient Descent (SGD) optimizer is used for training, with a weight decay of 0.0005 and a momentum of 0.9. The initial learning rate is set to 0.01, which reduces to 0.001 through a polynomial scheduler strategy during training. The image size is set to 224×224. The batch size is set to 8. CCL-MPC is evaluated on the validation dataset every 500 iterations, and the ultimate model is chosen for testing.

4.4. Comparative Experiment

CCL-MPC is compared with the existing state-of-the-art semi-supervised segmentation methods, including MT [10], UA-MT [29], Transformation-Consistent Self-ensembling Model (TCSM_v2) [26], CPS [28], UGPCL [8], Uncertainty-guided Collaborative Mean-Teacher (UCMT) [32], Min-Max Similarity (MMS) [20], PolypMixNet [39] and Duo-SegNet [40]. In addition, we regard the U-Net that only uses labeled data for supervised training as SupOnly.

Table 1: Quantitative results generated by different methods on the ISIC2018 dataset with the 5%, 12.5% and 25% labeled data ratios, respectively. The best values are in bold.

Methods	Labeled data ratio = 5%			Labeled data ratio = 12.5%			Labeled data ratio = 25%		
	mDice[%]↑	mIoU[%]↑	MAE↓	mDice[%]↑	mIoU[%]↑	MAE↓	mDice[%]↑	mIoU[%]↑	MAE↓
SupOnly	79.19	69.83	0.1061	81.59	72.86	0.0912	83.44	74.80	0.0795
MT [10]	79.38	69.78	0.1075	82.68	73.95	0.0898	82.71	74.00	0.0904
UAMT [29]	75.06	66.53	0.1362	78.10	71.17	0.1110	80.97	73.07	0.0911
UCMT [32]	83.64	76.07	0.0779	85.56	78.74	0.0676	**88.19**	**79.21**	**0.0651**
TCSM_V2 [26]	75.46	67.01	0.1301	78.85	70.13	0.1135	81.54	73.77	0.0866
CPS [28]	79.87	72.58	0.1112	81.72	73.71	0.0957	82.26	74.85	0.0874
UGPCL [8]	83.50	75.93	0.0823	83.68	74.95	0.0874	84.75	75.93	0.0823
MMS [20]	78.29	69.23	0.1132	78.96	69.33	0.1176	79.34	70.49	0.1062
CCL-MPC	**84.34**	**77.07**	**0.0763**	**86.28**	**78.55**	**0.0634**	87.13	77.96	0.0707

Table 2: Quantitative results generated by different methods on the Kvasir-SEG, CVC-ClinicDB, and CVC-300 with the 5% labeled data ratio. The best values are in bold.

Methods	Labeled data ratio = 5%					
	Kvasir-SEG		CVC-ClinicDB		CVC-300	
	mDice[%]↑	mIoU[%]↑	mDice[%]↑	mIoU[%]↑	mDice[%]↑	mIoU[%]↑
SupOnly	75.12	64.96	69.13	59.55	80.97	72.14
MT [10]	73.74	62.08	60.16	49.25	63.81	51.11
UCMT [32]	81.53	73.73	73.45	65.64	83.56	76.18
UGPCL [8]	82.14	76.52	72.20	65.41	84.51	76.36
MMS [20]	81.40	73.35	68.83	58.94	74.18	64.05
PolypMixNet [39]	82.92	74.27	71.98	65.16	84.31	77.61
Duo-SegNet [40]	68.16	61.91	58.49	51.65	63.92	60.52
CCL-HPC	**83.23**	**76.87**	**74.89**	**66.11**	**86.53**	**77.84**

4.4.1. Results on ISIC2018

Table I shows the quantitative evaluation results of these methods on the ISIC2018 dataset with 5%, 12.5%, and 25% labeled data ratios. The first row shows the performance of the baseline model trained only with labeled data. At 5% and 12.5% labeled data ratios, the $mDice$ scores of CCL-MPC are 84.34% and 86.28%, the $mIoU$ scores are 77.07% and 78.55%, and the MAE scores are 0.0763 and 0.0634, respectively, all of which achieve the best segmentation performance. At 25% labeled data ratio, the $mDice$, $mIoU$, and MAE of CCL-MPC are 87.13%, 77.96%, and 0.0707, respectively, which is second only to UCMT, but still better than the segmentation results of other methods. UGPCL achieves suboptimal performance on the ISIC2018. However, UGPCL cannot fully leverage the available data, resulting in a limitation of insufficient positive and negative samples. In contrast, CCL-MPC benefits from the SBCL and CGCL modules, improving the segmentation performance. In addition, CCL-MPC achieves greater performance improve-

Table 3: Quantitative results generated by different methods on the Kvasir-SEG, CVC-ClinicDB, and CVC-300 with the 12.5% labeled data ratio. The best values are in bold.

Labeled data ratio = 12.5%						
Methods	Kvasir-SEG		CVC-ClinicDB		CVC-300	
	mDice[%]↑	mIoU[%]↑	mDice[%]↑	mIoU[%]↑	mDice[%]↑	mIoU[%]↑
SupOnly	76.28	65.88	72.79	63.21	83.19	74.25
MT [10]	78.15	68.41	67.53	56.78	69.39	57.38
UCMT [32]	85.23	77.65	77.56	71.98	84.34	77.25
UGPCL [8]	84.74	77.19	74.65	67.71	86.51	78.71
MMS [20]	81.87	73.92	70.17	60.83	78.28	67.92
PolypMixNet [39]	82.20	77.02	73.61	66.63	85.45	76.66
Duo-SegNet [40]	69.91	62.57	59.51	52.39	64.45	60.02
CCL-HPC	**85.78**	**77.81**	**78.17**	**72.12**	**87.18**	**79.12**

Table 4: Quantitative results generated by different methods on the Kvasir-SEG, CVC-ClinicDB, and CVC-300 with the 25% labeled data ratio. The best values are in bold.

Labeled data ratio = 25%						
Methods	Kvasir-SEG		CVC-ClinicDB		CVC-300	
	mDice[%]↑	mIoU[%]↑	mDice[%]↑	mIoU[%]↑	mDice[%]↑	mIoU[%]↑
SupOnly	78.33	69.56	73.97	66.12	83.71	74.89
MT [10]	82.42	73.64	74.75	66.33	81.70	72.34
UCMT [32]	84.64	76.13	78.69	72.61	86.25	78.06
UGPCL [8]	82.98	75.28	76.00	68.71	84.74	76.89
MMS [20]	81.46	75.53	76.54	71.05	80.88	71.12
PolypMixNet [39]	84.02	77.60	78.41	72.24	87.35	79.31
Duo-SegNet [40]	70.56	63.67	61.09	54.38	67.67	61.19
CCL-HPC	**86.37**	**78.29**	**79.79**	**72.80**	**87.93**	**80.49**

ment at the 5% labeled data ratio, which indicates that CCL-MPC can use unlabeled data more effectively and obtain better segmentation performance than other methods. In summary, the proposed CCL-MPC demonstrates strong performance across testing datasets with different labeled data ratios.

Fig. 3 illustrates the visualization results of different methods on the ISIC2018 with a 5% labeled data ratio. In contrast to others, CCL-MPC yields superior segmentation results, particularly in the boundaries of skin lesions. CCL-MPC captures sufficiently more positive and negative samples, which can improve both intra-class cohesion and inter-class separation. In comparison to other methods, CCL-MPC can capture finer structural details and obtain the smoother boundaries. Therefore, CCL-MPC has better effects on boundary extraction than other popular semi-supervised methods.

4.4.2. Results on polyp datasets

Tables II, III, and IV show the quantitative evaluation results of these methods on CVC-300, Kvasir-SEG, and CVC-ClinicDB with labeled data ratios of 5%, 12.5%, and 25%. In Tables II, III, and IV, in terms of $mDice$ and $mIoU$, CCL-MPC outperforms other methods and achieves the best segmentation performance. First, in comparison to SupOnly, which only employs labeled data, CCL-MPC demonstrates a significant improvement in segmentation performance. In particular, there is the 9.5% increase in $mDice$ (76.28% to 85.78%) compared to SupOnly on the Kvasir-SEG with a 12.5% labeled data ratio. Similarly, leveraging unlabeled data, CCL-MPC also achieves substantial improvements on the segmentation performance across different scales and datasets. Secondly, as the proportion of labeled data increases, the segmentation performance of CCL-MPC continues to improve. Notably, when the labeled ratio increases from 5% to 12.5%, $mDice$ improved by 3.28% on the CVC-ClinicDB dataset.

Fig. 4 further shows the visualization results of different methods on the CVC-300 at a 5% labeled data ratio. The polyp segmentation task is relatively challenging due to the diversity of polyp shapes and sizes, as well as the blurry boundaries caused by the adhesion between polyps and surrounding tissues. UCMT produces the second-best results by training with high-confidence pseudo-labels, lacking the capability to mine richer feature representations. The proposed CCL-MPC can leverage intra-image and inter-image feature representations to avoid small isolated polyps during the segmentation process, thereby reducing the possibility of mis-segmenting small polyps. Observably, CCL-MPC achieves superior segmentation results compared to other popular semi-supervised segmentation methods.

4.5. Ablation Study

The U-Net that only uses labeled data for supervised training is chosen as the baseline (the 1st row in Tables V and VI), and the proposed components are gradually introduced to demonstrate their effectiveness. Ablation studies are conducted on the ISIC2018 and CVC-300 with the 5%, 12.5%, and 25% labeled data ratio.

4.5.1. The Effects of HDC and CAC

To better demonstrate the effectiveness of MPC, we select three labeled data ratios and evaluated MPC on two different medical image segmentation tasks. Simultaneously, to validate the effectiveness of the HDC and

Table 5: Quantitative results of ablation study on the CVC-300 with the 5%, 12.5%, and 25% labeled data ratio. The best values are in bold.

	MPC		CCL		Labeled data ratio = 5%		Labeled data ratio = 12.5%		Labeled data ratio = 25%	
l_{sup}	l_{hdc}	l_{cac}	l_{sbcl}	l_{cgcl}	mDice[%]↑	mIoU[%]↑	mDice[%]↑	mIoU[%]↑	mDice[%]↑	mIoU[%]↑
✓					80.97	72.14	83.19	74.25	83.71	74.89
✓	✓				82.78	73.13	83.97	75.23	84.89	75.82
✓		✓			82.93	72.99	83.60	74.92	85.22	76.57
✓	✓	✓			84.10	74.09	84.12	75.60	85.95	78.51
✓	✓	✓		✓	84.98	74.89	85.31	77.79	86.06	77.79
✓	✓	✓	✓		85.84	77.27	86.73	78.45	86.35	78.78
✓	✓	✓	✓	✓	**86.53**	**77.84**	**87.18**	**79.12**	**87.93**	**80.49**

Table 6: Quantitative results of ablation study on the ISIC2018 with the 5%, 12.5%, and 25% labeled data ratio. The best values are in bold.

	MPC		CCL		Labeled data ratio = 5%		Labeled data ratio = 12.5%		Labeled data ratio = 25%	
l_{sup}	l_{hdc}	l_{cac}	l_{sbcl}	l_{cgcl}	mDice[%]↑	mIoU[%]↑	mDice[%]↑	mIoU[%]↑	mDice[%]↑	mIoU[%]↑
✓					79.19	69.83	81.59	72.86	83.44	74.80
✓	✓				80.43	71.17	83.07	74.91	84.97	76.37
✓		✓			81.53	73.12	84.20	75.49	83.69	75.00
✓	✓	✓			81.85	73.96	83.87	75.61	85.19	76.62
✓	✓	✓		✓	83.07	75.11	85.12	77.70	85.44	76.80
✓	✓	✓	✓		82.14	75.28	84.96	76.36	86.16	77.70
✓	✓	✓	✓	✓	**84.34**	**77.07**	**86.28**	**78.55**	**87.13**	**77.96**

CAC within MPC, we separately introduce HDC and CAC based on the baseline, as shown in the 2nd and 3rd rows of Tables V and VI. Initially, compared to the baseline, incorporating HDC alone improves the $mIoU$ and $mDice$ across various labeled data ratios and datasets. HDC obtains feature representations from different perspectives through heterogeneous network architecture, effectively utilizing the advantages of CNN and Transformer. Therefore, HDC can utilize the complementarity of two architectures to extract more comprehensive and robust feature representations. Subsequently, compared to the baseline, integrating CAC alone also improves the $mIoU$ and $mDice$ across various labeled data ratios and datasets. This improvement is attributed to CAC to utilize certainty pseudo-labels for auxiliary supervision, improving the reliability of pseudo-labels and allowing consistency learning to focus on high certainty regions, resulting in more reliable segmentation results. Meanwhile, CAC performs consistency regularization on two different augmented views, encouraging the prediction results of the two different augmented views to remain consistent. This approach improves consistency within the feature space by minimizing the cosine distance between different views.

In addition, compared to the baseline, the use of CAC and HDC achieve more significant improvements in segmentation performance. Especially at a 5% labeled data ratio, $mDice$ is improved by 3.13% on the CVC-300. The main reason may be that there is good complementarity between the two types of consistency learning. On the contrary, consistency regularization can be performed in a more complementary way, ensuring that the model focuses on more robust feature representations. On the ISIC2018, the combination of HDC and CAC significantly improve $mDice$ and $mIoU$ scores, demonstrating that the MPC can effectively improve the segmentation performance of the model for complex skin lesions. We also observe similar performance improvements on the CVC-300, demonstrating that the MPC has wide applicability in different types of medical images.

4.5.2. The Effects of SBCL and CGCL

To better demonstrate the effectiveness of CCL, we selected three labeled data ratios and validated CCL on two different public datasets. Additionally, to verify the effectiveness of the SBCL and CGCL modules within CCL, we introduce SBCL and CGCL, respectively based on the third row (utilizing HDC and CAC simultaneously), as shown in the 4th and 5th rows of Tables V and VI. Firstly, with the introduction of CGCL, both $mDice$ and $mIoU$ have been improved. This improvement is attributed to that CGCL uses certainty pseudo-labels to select pixels within the same image, selecting positive and negative samples with higher certainty and reducing the false negative sample rate. Meanwhile, by introducing a memory queue, CGCL can learn more effectively from a large number of samples, which helps improve the performance of the model. Secondly, with the introduction of SBCL, $mIoU$ and $mDice$ were both improved on these two datasets with different labeled data ratios. The proposed symmetric bidirectional sampling mechanism expands the scope of contrastive learning by capturing potential relationships between different images, significantly enriching the number of positive and negative samples.

In addition, the segmentation performance is greatly improved when both SBCL and CGCL modules are used simultaneously. These two modules complement each other through feature learning at different levels, effectively utilizing the potential relationships between labeled and unlabeled data to capture feature representations with better generalization performance. The inter-image features provided by SBCL enhance the diversity of pseudo-labels within the image, while the high-quality pseudo-labels provided by CGCL

achieve reliability for inter-image feature learning. The two modules support each other in the feature embedding space, achieving high cohesion within categories and strong separation between categories through a more comprehensive contrastive learning strategy. It is worth noting that SBCL can achieve better performance compared to CGCL. This can be attributed to the symmetric bidirectional sampling mechanism, which provides more diverse samples and enhances intra-class cohesion and inter-class separation. Additionally, it is worth noting that the combination of SBCL and CGCL significantly improved the segmentation accuracy of the model on the ISIC2018, especially in terms of boundary extraction. On the CVC-300, this combination also demonstrated superior performance, especially when dealing with polyps with various shapes and sizes.

5. Conclusion

In this work, we have proposed a semi-supervised method CCL-MPC for medical image segmentation. First, we devise the CCL to collaboratively integrate the SBCL and CGCL modules, enriching feature representations and enhancing intra-class cohesion and inter-class separation. The proposed SBCL effectively leverages the potential for feature learning between images. Importantly, we introduce a novel symmetric bidirectional sampling mechanism to capture the intrinsic features across different images. Simultaneously, the proposed CGCL effectively utilizes the feature learning capability within images. It ensures high-confidence samples and reduces potential noise during the training process using certainty pseudo-labels and the top-K method. Additionally, we propose MPC, which leverages a dual-branch network architecture and two augmented views to improve the quality and diversity of training data, thereby improving the capacity to capture robust features. Extensive experiments demonstrate that CCL-MPC achieves the state-of-the-art performance for medical image segmentation.

6. Acknowledgements

This work is partly supported by National Natural Science Foundation of China (Nos. 61861024, 62271296, and 62201334), Scientific Research Program Funded by Shaanxi Provincial Education Department (Nos. 23JP022, and 23JP014), and General Project of Key Research and Development Programs in Shaanxi Province, China, Social Development Area (No. 2022SF-105).

References

[1] O. Ronneberger, P. Fischer, T. Brox, "U-Net: convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention*, 2017, 40(4): 834–848.

[2] C. L. Chen, G. Papandreou, I. Kokkinos, et al., "Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 234–241.

[3] R. Wang, T. Lei, R. Cui, et al., "Medical image segmentation using deep learning: a survey," *IET Image Processing*, 2022, 16(5): 1243–1267.

[4] D. Zhang, G. Huang, Q. Zhang, et al., "Cross-modality deep feature learning for brain tumor segmentation," *Pattern Recognition*, 2021, 110: 107562.

[5] D. Zhang, G. Huang, Q. Zhang, et al., "Exploring task structure for brain tumor segmentation from multi-modality MR images," *IEEE Transactions on Image Processing*, 2020, 29: 9032–9043.

[6] X. Zhao, C. Fang, D. Fan, et al., "Cross-level contrastive learning and consistency constraint for semi-supervised medical image segmentation," *IEEE International Symposium on Biomedical Imaging*, 2022, 1–5.

[7] C. Chen, J. Han, K. Debattista., "Virtual category learning: a semi-supervised learning method for dense prediction with extremely limited labels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(8): 5595–5611.

[8] T. Wang, J. Lu, Z. Lai, et al., "Uncertainty-guided pixel contrastive learning for semi-supervised medical image segmentation," *Proceedings of the International Joint Conference on Artificial Intelligence*, 2022, 1444–1450.

[9] T. Lei, D. Zhang, X. Du, et al., "Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network," *IEEE Transactions on Medical Imaging*, 2022, 1265–1277.

[10] A. Tarvainen and H. Valpola, "Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in Neural Information Processing Systems*, 2017, 30.

[11] Z. Zhao, H. Wang, T. Lei, et al., "Balanced feature fusion collaborative training for semi-supervised medical image segmentation," *Pattern Recognition*, 2025, 157: 110856.

[12] X. Luo, J. Chen, T. Song, and G. Wang, "Semi-supervised medical image segmentation through dual-task consistency," *Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, 2021, 35(10): 8801–8809.

[13] T. Lei, H. Liu, Y. Wan, et al., "Shape-guided dual consistency semi-supervised learning framework for 3D medical image segmentation," *IEEE Transactions on Radiation and Plasma Medical Sciences*, 2023, 719–731.

[14] H. Wu, F. Xiao, and C. Liang, "Dual contrastive learning with anatomical auxiliary supervision for few-shot medical image segmentation," *Proceedings of the European Conference on Computer Vision*, 2022, 417–434.

[15] H. Wu, W. Xie, J. Lin, and X. Guo, "ACL-Net: semi-supervised polyp segmentation via affinity contrastive learning," *Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, 2023, 37(3): 2812–2820.

[16] H. Basak and Z. Yin, "Pseudo-Label guided contrastive learning for semi-supervised medical image segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, 19786–19797.

[17] C. Tang, X. Zeng, L. Zhou, et al., "Semi-supervised medical image segmentation via hard positives oriented contrastive learning," *Pattern Recognition*, 2024, 146: 110020.

[18] Q. Yu, N. Xi, J. Yuan, et al., "Source-free domain adaptation for medical image segmentation via prototype-anchored feature alignment and

contrastive learning," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2023, 3–12.

[19] H. Basak and Z. Yin, "Semi-supervised domain adaptive medical image segmentation through consistency regularized disentangled contrastive learning," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2023, 260–270.

[20] A. Lou, K. Tawfik, X. Yao, et al., "Min-Max Similarity: a contrastive semi-supervised deep learning network for surgical tools segmentation," *IEEE Transactions on Medical Imaging*, 2023, 42(10): 2832–2841.

[21] Y. Zhang, Z. Wang, Y. Wang, et al., "Boundary-aware contrastive learning for semi-supervised nuclei instance segmentation," *arXiv*, 2024, 2402.04756.

[22] Z. Wang and C. Ma, "Dual-contrastive dual-consistency dual-transformer: a semi-supervised approach to medical image segmentation," *Proceedings of the IEEE International Conference on Computer Vision*, 2023, 870–879.

[23] X. Zhao, Z. Qi, S. Wang, et al., "RCPS: rectified contrastive pseudo supervision for semi-supervised medical image segmentation," *arXiv*, 2023, 2301.05500.

[24] L. Yang, Z. Zhao, L. Qi, et al., "Shrinking class space for enhanced certainty in semi-supervised learning," *Proceedings of the IEEE International Conference on Computer Vision*, 2023, 16187–16196.

[25] B. Liu, N. Xu, J. Lv, et al., "Revisiting pseudo-label for single-positive multi-label learning," *International Conference on Machine Learning*, 2023, 22249–22265.

[26] X. Li, L. Yu, H. Chen, et al., "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, 2020, 32(2): 523–534.

[27] Y. Liu, Y. Tian, Y. Chen, et al., "Perturbed and strict mean teachers for semi-supervised semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, 4258–4267.

[28] X. Chen, Y. Yuan, G. Zeng and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, 2613–2622.

[29] L. Yu, S. Wang, X. Li, et al., "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," *Medical Image Computing and Computer Assisted Intervention*, 2019, 605–613.

[30] Y. Wu, M. Xu, Z. Ge, et al., "Semi-supervised left atrium segmentation with mutual consistency training," *Medical Image Computing and Computer Assisted Intervention*, 2021, 297–306.

[31] J. Chen, J. Zhang, K. Debattista, et al., "Semi-supervised unpaired medical image segmentation through task-affinity consistency," *IEEE Transactions on Medical Imaging*, 2022, 42(3): 594–605.

[32] Z. Shen, P. Cao, H. Yang, et al., "Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation," *Proceedings of the International Joint Conference on Artificial Intelligence*, 2023, 4199–4207.

[33] N. C. F. Codella, D. Gutman, E. M. Celebi, et al., "Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging," *IEEE International Symposium on Biomedical Imaging*, 2018, 168–172.

[34] D. Jha, P. H Smedsrud, M. A Riegler, et al., "Kvasir-SEG: a segmented polyp dataset," *MultiMedia Modeling: 26th International Conference*, 2020, 451–462.

[35] J. Bernal, F J. Sánchez, G. Fernández-Esparrach, et al., "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, 2015, 99–111.

[36] D. Vázquez, J. Bernal, J. Sánchez, et al., "A benchmark for endoluminal scene segmentation of colonoscopy images," *Journal of Healthcare Engineering*, 2017, 4037190.

[37] D. Fan, G. Ji, T. Zhou, et al., "Pranet: parallel reverse attention network for polyp segmentation," *International Conference on Medical Image Computing and Computer-assisted Intervention*, 2020, 263–273.

[38] O. Aäron, L. Yazhe and V. Oriol, "Representation learning with contrastive predictive coding," *arXiv*, 2018, abs/1807.03748.

[39] X. Jia, Y. Shen, J. Yang, et al., "PolypMixNet: enhancing semi-supervised polyp segmentation with polyp-aware augmentation," *Computers in Biology and Medicine*, 2024, 170: 108006.

[40] H. Peiris, Z. Chen, G. Egan, et al., "Duo-SegNet: adversarial dual-views for semi-supervised medical image segmentation," *Medical Image Computing and Computer Assisted Intervention*, 2021, 428–438.