Uncertainty-Guided Pixel Contrastive Learning for Semi-Supervised Medical Image Segmentation

Tao Wang[1] , **Jianglin Lu**[1] , **Zhihui Lai**[1,2*] , **Jiajun Wen**[1] , **Heng Kong**[3]
[1]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[2]Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, China
[3]Shenzhen University General Hospital, Shenzhen, China
taovvang@gmail.com, JianglinLu@outlook.com, {lai_zhi_hui, enjoy_world}@163.com, generaldoc@126.com

Abstract

Recently, contrastive learning has shown great potential in medical image segmentation. Due to the lack of expert annotations, however, it is challenging to apply contrastive learning in semi-supervised scenes. To solve this problem, we propose a novel uncertainty-guided pixel contrastive learning method for semi-supervised medical image segmentation. Specifically, we construct an uncertainty map for each unlabeled image and then remove the uncertainty region in the uncertainty map to reduce the possibility of noise sampling. The uncertainty map is determined by a well-designed consistency learning mechanism, which generates comprehensive predictions for unlabeled data by encouraging consistent network outputs from two different decoders. In addition, we suggest that the effective global representations learned by an image encoder should be equivariant to different geometric transformations. To this end, we construct an equivariant contrastive loss to strengthen global representation learning ability of the encoder. Extensive experiments conducted on popular medical image benchmarks demonstrate that the proposed method achieves better segmentation performance than the state-of-the-art methods.

1 Introduction

Medical image segmentation plays an important role in computer-aided diagnosis system. Supervised learning methods based on deep learning have achieved great performance [Ronneberger *et al.*, 2015; Cao *et al.*, 2021; Li *et al.*, 2021] relying on a large number of labeled data. However, it is difficult to obtain large-scale medical image annotations due to the requirements of professional clinical knowledge and time consumption on data collection and labeling. Semi-supervised learning can leverage both labeled data and unlabeled data, which greatly reduces the dependence on annotations. Semi-supervised learning aims to explore the internal information of unlabeled data to improve the performance of the model. Some popular semi-supervised learning strategies include self-training using pseudo-labels [Qiao *et al.*, 2018; Chen *et al.*, 2021a], self-ensembling [Tarvainen and Valpola, 2017; Yu *et al.*, 2019], entropy minimization [Vu *et al.*, 2019], and consistency regularization [Laine and Aila, 2016; Ouali *et al.*, 2020]. The above-mentioned methods utilize unlabeled data during training stage by constructing trusted labels or forcing the prediction consistency when the input data encounters disturbances. However, these methods make the classification of each pixel independent, which ignores the internal correlation between pixels (or features) of images.

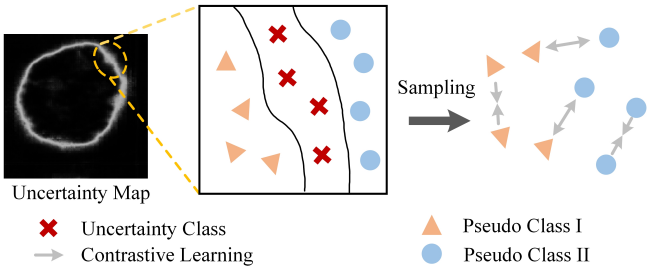

Figure 1: Workflow of uncertainty guidance. We aim to reduce the possibility of noise sampling by removing uncertainty region in the uncertainty map so as to ensure the effectiveness of contrastive learning.

In order to strengthen the connection between pixels, some recent works have applied contrastive learning to segmentation tasks [Hu *et al.*, 2021]. The methods of contrastive learning have achieved superior performance in the self-supervised representation learning of natural images. The core idea of contrastive learning is that the representations of similar samples should be alike, and the representations of different kinds of samples should be different. How to define similar samples is the key in contrastive learning. Image-level contrastive learning defines similar samples as different transformations of the same image, and those from different images are defined as dissimilarity. However, similar pixels are densely distributed in the segmentation task. So the definition of dissimilar samples is not suitable for pixel-level contrastive learning. To solve this problem, [Wang *et al.*, 2021] uses segmentation labels to construct contrast samples for supervised segmentation tasks. For unlabeled data, [Chen *et al.*, 2021b] uses

*Corresponding author

the predicted pseudo labels to determine the sample category. [Zhong *et al.*, 2021] uses the spatial consistency of weakly enhanced images to construct similar samples, and construct dissimilar samples by a simple cross-image and pseudo-label weighting heuristic. In fact, using pseudo-labels to construct samples is likely to be inconsistent with the actual semantic categories, which may lead to a noise sampling on contrastive learning. In addition, pixel contrastive learning only establishes the association of local pixels, ignoring the learning of global representation information.

In this paper, our objectives aim to 1) solve the noise sampling problem of contrastive learning using pseudo-labels and 2) strengthen the global representation learning ability of the encoder. To achieve these goals, we propose a contrastive learning method based on uncertainty. Fig. 1 shows the core idea of our method. For unlabeled data, we use the uncertainty map to guide the region of pseudo-labels sampling and reduce the number of wrong samples. Then the sample contrastive loss is calculated to optimize the network and reduce the uncertainty area of predictions. To obtain a better uncertainty map, we design a consistency learning strategy with CNN decoder and Transformer decoder, which can obtain accurate predictions from different views using the structural differences between two decoders. In addition, segmentation models should have the ability to identify geometric transformations. Base on this, we define an equivariant contrastive loss to force the network to learn the identification information of geometric transformations by adding a transformation category prediction in the representation learning stage.

In summary, our contributions mainly include:

- We propose a novel uncertainty guided contrastive learning method, which can effectively alleviate noise sampling from pseudo-labels of unlabeled data.
- A consistency learning strategy for heterogeneous decoders based on CNN and transformer is designed, which can obtain reliable prediction results and uncertainty map by consistency training on unlabeled data.
- We define an equivariant contrastive loss for global representation learning, which equips the model with discrimination ability to distinguish different geometric transformations of images.

2 Related Work

2.1 Semi-supervised Medical Image Segmentation

Without the requirement of large-scale labeled data, semi-supervised learning has attracted much attention in medical image segmentation. Existing semi-supervised medical image segmentation methods mainly involves in entropy minimization, pseudo label self-training, collaborative training and consistency learning. Entropy minimization [Vu *et al.*, 2019] suggests that high-quality prediction results should have a low entropy, and hence it conducts model learning by minimizing the information entropy of the prediction probability distribution. Pseudo label self-training [Chen *et al.*, 2021a] performs class supervised learning by predicting pseudo labels for the unlabeled data. Co-training [Qiao *et al.*, 2018] assumes that there are multiple decision views containing complementary information, and designs different classifiers to learn different views to promote segmentation performance. Consistency learning [Verma *et al.*, 2019; Laine and Aila, 2016; Tarvainen and Valpola, 2017; Ouali *et al.*, 2020] makes an assumption that even if an image sample encounters some disturbances, say input disturbance or model disturbance, the prediction results from the sample should not change. Motivated by such intuition, these methods conduct model training by encouraging consistent prediction of unlabeled disturbed samples. Inspired by collaborative training and consistency learning, we propose to characterize the complementary information of data from different views, using the structural differences between CNN and transformer, and apply consistency constraints to train the model.

2.2 Contrastive Learning

In image-level representation learning, contrastive learning can make full use of unlabeled data to learn effective visual representation, in which the core idea is to strengthen the discrimination of the learned visual representation by narrowing similar pairs (positive) and separating dissimilar pairs (negative) based on some similarity constraints. The key point of image-level contrastive learning is how to construct contrastive samples. A feasible solution is proposed in [He *et al.*, 2020], which increases the number of contrastive samples by introducing memory bank and momentum contrast.

Recently, some works [Chaitanya *et al.*, 2020; Wang *et al.*, 2021; Zhong *et al.*, 2021; Hu *et al.*, 2021] have been proposed to extend contrastive learning from image-level to pixel-level for image segmentation. The main idea of pixel-level contrastive learning is to construct pixel sample pairs with the help of segmentation labels. For unlabeled data, sample pairs are constructed by using pseudo labels or spatial structure. Nevertheless, these methods may encounter the problem of noise sampling during the process of constructing sample pairs. To alleviate this problem, we suggest using prediction uncertainty to guide sample sampling and reduce the number of noise samples. In addition, pixel-level contrastive learning lacks the capture ability of global representation, which urges us to impose a constraint of prior knowledge in representation learning for segmentation task.

2.3 Uncertainty Estimation

In semi-supervised learning, uncertainty can be used to evaluate the quality of model predictions for better use of unlabeled data. The measure methods of estimating uncertainty mainly include 1) using information entropy of the prediction probability distribution, 2)using the deviation of multiple predition results of the same input under different disturbances [Yu *et al.*, 2019], and 3) calculating the variance of different prediction results of the same input [Zheng and Yang, 2021]. However, these methods are time-consuming and lack of reliability. In our method, we estimate the uncertainty by calculating the entropy of the average probability distribution obtained by different predictors to overcome these problems.

3 Methodology

Given a label dataset $D_L = \{(x_i, y_i), i = 1, \ldots, N\}$ and an unlabeled dataset $D_U = \{x_j, j = 1, \ldots, M\}$, where

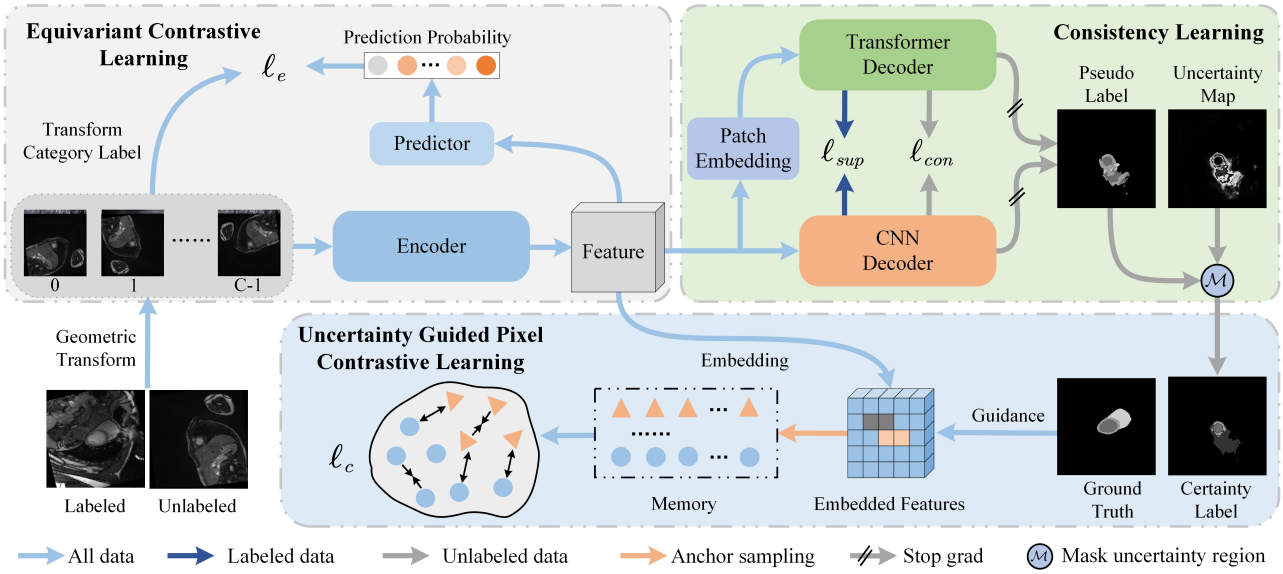

Figure 2: Overview of our method. Arrows of different colors represent the processing flows of different data. In particular, for unlabeled data, the certainty labels are obtained by mask option using pseudo labels and uncertainty maps. For pixel contrastive learning, contrastive anchors are selected by the certainty labels of unlabeled data and the ground truth of labeled data.

$M \gg N$, the images in D_L and D_U firstly go through a geometric transformation and then feed into an encoder network to extract multi-scale features. After that, these features will be sent to the following three branches, including *consistency learning* branch, *uncertainty-guided contrastive learning* branch, and *equivariant contrastive learning* branch. For the consistency learning branch, we propose a heterogeneous consistency network to predict the segmentation results, which is driven by a supervision loss ℓ_{sup} and a consistency loss ℓ_{con}. ℓ_{sup} is calculated by ground truth of D_L, and ℓ_{con} is calculated by prediction consistency of D_U. For the uncertainty-guided constrastive learning branch, we build and maintain a memory queue to preserve enough samples for constrastive learning. The selection of samples in the memory queue depends on the labels of D_L and certainty labels of D_U. For the selected samples, we impose a pixel-level contrastive loss ℓ_c to make the pixels of the same class close to each other and the pixels of different classes far away from each other. For the equivariant contrastive learning branch, we perform geometric transformation category prediction on all labeled and unlabeled data, and design an equivariant contrastive loss ℓ_e to force the encoder to be robust to geometric transformation. For ease of understanding, Fig.2 gives an illustration of the overall architecture and the training process of our proposed method. In summary, the total objective of our method is:

$$\ell = \ell_{sup} + \lambda_t \ell_{con} + \lambda_1 \ell_c + \lambda_2 \ell_e \tag{1}$$

In this paper, we set $\lambda_1 = \lambda_2 = 0.1$ and λ_t is a temperature parameter that increases from 0 to 0.01. The following shows the above-mentioned three branches in details.

3.1 Consistency Learning Between Decoders

In the consistency learning branch, we design a simple yet effective network structure to achieve the following two goals: 1) using unlabeled data to promote the learning of segmentation network, and 2) obtaining reliable uncertainty estimation from network outputs. It is demonstrated that using co-training strategy can obtain better segmentation performance, of which the core idea is to make different classification predictions from different views and then regard the differences of predictions as the measurement criterion of uncertainty estimation. Instead of using the same architecture that requires adding some disturbances for co-training, inspired by [Luo *et al.*, 2021], we adopt a simple yet effective scheme that takes advantage of the congenital differences between transformer decoder and CNN decoder. Specifically, we construct a heterogeneous predictor to constrain the two decoders to generate consistent predictions. And then the entropy of the mean prediction is used to estimate the uncertainty map.

Patch Embedding and Position Encoding. We choose two different decoders $f_\theta^t(\cdot)$ and $f_\theta^c(\cdot)$ from Swin-UNet and UNet. Through the encoder, we can obtain a set of features$\{f_i, i = 0, \ldots, 3\}$. Before inputing into $f_\theta^t(\cdot)$, we need to reshape the feature $f_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ into a sequence of flattened patches $x_p^i \in \mathbb{R}^{P_i^2 \times C_i}$ and embed them into a D-dimensional space using a learnable linear projection E. To preserve the spatial information, we add the absolute location encoding to the embedded patches:

$$PE(x_i) = [x_p^{i1}E; x_p^{i2}E; \cdots; x_p^{iP_i^2}E] + E_{pos} \tag{2}$$

where $E \in \mathbb{R}^{(P_i^2 \times C_i) \times D}$ is the patch embedding projection, and $E_{pos} \in R^{N \times D}$ denotes the absolute location encoding.

Consistency Learning. Given an input image, we can get two predicted probability distributions p_t and p_c from two decoders. For labeled data, we use ground thuth to calculate the supervision segmentation loss:

$$\ell_{sup} = \mathcal{L}_{seg}(p_c, y) + \alpha \mathcal{L}_{seg}(p_t, y) \tag{3}$$

$$\mathcal{L}_{seg} = \frac{1}{2}(\mathcal{L}_{CE} + \mathcal{L}_{Dice}) \tag{4}$$

where $\mathcal{L}_{CE}$, $\mathcal{L}_{Dice}$ are the cross-entropy loss and dice loss, and y is ground truth of labeled data. We use $f_\theta^c(\cdot)$ as the main predictor, so α is set to 0.4. In the inference stage, prediction from CNN branch is the final result. For unlabeled data, consistency loss is calculated as follows:

$$\ell_{con} = \mathcal{L}_{dis}(p_c, p_t) \tag{5}$$

where $\mathcal{L}_{dis}$ is a distance measure between two output probability distributions. In this work, we choose to use mean squared error (MSE) as a distance measure.

3.2 Uncertainty-Guided Contrastive Learning

Image segmentation task generally involves with cross entropy loss for conducting pixel-level classification. However, such loss makes the classification of each pixel independent and hence ignores the relationship between the pixels. To solve this problem, we design a pixel-level contrastive learning mechanism that classifies the pixels of the same category (semantic label) as positive samples, and the pixels of different categories as negative samples. The relationship between pixels is established by reducing the distance between positive samples and enlarging the distance between negative samples in an embedding space. In order to effectively use unlabeled data for pixel-level contrastive learning, we estimate the uncertainty of unlabeled pixels and the pixels with a higher certainty are selected as anchor points for contrastive learning.

Mask Uncertainty Region. We choose the predictive entropy as the metric to approximate the uncertainty. Specifically, we first calculate the average probability distribution of the prediction results $\hat{p} = (p_c + p_t)/2$, and then calculate the entropy for the probability distribution of each pixel in the channel dimension. It can be summarized as:

$$u = -\sum_c \hat{p}_c \log(\hat{p}_c + \epsilon) \tag{6}$$

where ϵ is a very small constant to avoid singularity. We believe that the prediction with large entropy is uncertain in category. When calculating the pseudo labels, those uncertainty predictions are removed as a non-sampling region, and then the determined pseudo labels are obtained:

$$y_p = Argmax(\hat{p})_{|u<H} \tag{7}$$

where H is a threshold to mask the uncertain labels, and y_p is the final certainty pseudo label.

Anchor Sampling. We use the labels of labeled images and the certainty pseudo labels of unlabeled images as the basis for the use of contrastive samples. Because the original image resolution is too large, the cost of contrastive learning in the original image size is expensive and the prototype vectors of pixels contain less semantic information. Therefore, we use contrastive learning in the feature space with low resolution. Firstly, the features extracted from the encoder will be embedded into the D-dimensional space, where each D-dimensional feature vector represents the prototype vector of the pixel. Then, the labels are downsampled to the same resolution, the category is specified for each prototype vector, and the vector in the uncertainty region is not sampled. We adopt the strategy of random sampling with a fixed number of samples for each category. If the number of samples from same category is small, we will sample anchors from other categories. The number of contrastive negative samples greatly affects the performance of contrastive learning, but a large number of negative samples will produce a lot of overhead. A better solution is to use a fixed size external storage to store the sampled samples and update the storage content with the training. In our method, we set up a memory queue to store the collected samples. In each iteration, the randomly selected samples are used as anchors to calculate the contrastive loss, and then they are updated to the memory queue.

Pixel Contrastive Loss. The prototype vectors and their category of pixels are saved in the sample queue. We use the popular InfoNCE [van den Oord *et al.*, 2018] loss function to calculate the contrastive loss. In each iteration, we randomly sample M anchors and calculate the contrastive loss for each anchor. Then average the loss of all anchors as the overall contrastive loss. The specific calculation is as follows:

$$\ell_c^i = -\frac{1}{|P_i|}\sum_{v_i^+ \in P_i} log\frac{e^{cos(v_i, v_i^+)/\tau}}{e^{cos(v_i, v_i^+)/\tau} + \sum_{v_i^- \in N_i} e^{cos(v_i, v_i^-)/\tau}} \tag{8}$$

$$\ell_c = \frac{1}{M}\sum_{i=1}^{M} \ell_c^i \tag{9}$$

where P_i and N_i denote prototype vector's collections of the positive and negative samples for pixel i. v_i is the prototype vector of pixel i, v_i^+ is a positive prototype vector, v_i^- is an negative vector and τ is a temperature hyper-parameter.

3.3 Equivariant Contrastive Loss

For conducting contrastive learning, some previous work constructs positive samples by different transformations of the same image. However, some transformations do not accord with the prior knowledge of segmentation tasks[Dangovski *et al.*, 2021], such as geometric transformation. In this paper, we suggest that the effective feature representations required by segmentation task should be equivariant (or discriminative) to different geometric transformations.

Based on the above, we consider adding the equivariant contrastive loss to the representation learning of the segmentation model to learn the global information. Specifically, we define the segmentation model as encoder-decoder form: $f(x_i) = f_\gamma(f_\theta(x_i))$. For an image x_i, when it passes through some geometric transformation $G(\cdot)$, the corresponding segmentation result will also change, that is:

$$f(G(x_i)) = G(f(x_i)) \tag{10}$$

Then, we can deduce:

$$f_\theta(G(x_i)) \neq f_\theta(x_i) \tag{11}$$

Therefore, we can explicitly strengthen the learning of this geometric transformation information in $f_\theta(\cdot)$. We add a classification predictor $p_\phi(\cdot)$ to predict the discrimination results

Method	$ACDC_{\|X_L\|=68}$		$ACDC_{\|X_L\|=136}$		$ISIC_{\|X_L\|=91}$		$ISIC_{\|X_L\|=181}$	
	Dice[%]	Jaccard[%]	Dice[%]	Jaccard[%]	Dice[%]	Jaccard[%]	Dice[%]	Jaccard[%]
-	65.32	51.12	82.09	70.79	67.02	53.49	68.91	56.69
MT	74.20	60.53	85.75	75.69	69.87	57.19	70.64	58.63
UA-MT	74.24	60.68	83.56	72.94	69.33	56.20	75.67	63.42
EM	76.70	63.40	83.39	72.23	66.11	51.89	71.10	58.54
DCT	73.28	60.16	82.76	71.40	70.09	57.13	75.98	63.68
CCT	74.39	61.28	83.94	72.68	69.53	56.94	73.42	62.13
CPS	74.76	61.40	85.06	74.67	71.87	57.35	78.09	65.74
Ours	**80.05**	**67.66**	**88.11**	**79.15**	**72.67**	**57.85**	**79.48**	**67.25**

Table 1: The comparison of different mothods on ACDC dataset and ISIC dataset on different semi-supervised settings. The first row represents the baseline results of supervised training only using labeled data. $|X_L|$ represents the number of labeled images.

ℓ_{con}	ℓ_e	ℓ_c^{sup}	ℓ_c^{pseudo}	ℓ_c	Dice[%]	Jaccard[%]
					82.09	70.79
✓					84.34	73.56
	✓				85.06	74.76
✓	✓				86.05	76.21
✓	✓	✓			86.50	76.73
✓	✓	✓	✓		87.09	77.61
✓	✓	✓		✓	**88.11**	**79.15**

Table 2: Quantitative results of ablation study on ACDC dataset.

Feature	Resolution	Metrics	
		Dice[%]	Jaccard[%]
$Conv_1$	56×56	87.15	77.64
$Conv_2$	28×28	**88.11**	**78.95**
$Conv_3$	14×14	87.63	78.46
$Conv_4$	7×7	86.67	76.32

Table 3: The comparison results of pixel contrastive learning on the features of different resolutions using the proposed method.

of geometric transformation. Our equivariant contrastive loss function is as follows:

$$\ell_e = \frac{1}{C}\sum_{i=0}^{C-1}\mathcal{L}_{CE}(p_\phi(f_\theta(G^i(x))), i) \tag{12}$$

where the geometric transformation $G^i(\cdot)$ represents the four-fold rotation in this paper, so $C = 4$.

4 Experiments

4.1 Experimental Setup

Datasets and Metircs. We validate the proposed method on two public datasets:

- **ACDC dataset** [Bernard *et al.*, 2018] contains 200 annotated short-axis cardiac MR-cine images from 100 patients. We divide the dataset in a ratio of 7:3 to obtain the training set and verification set. According to different semi-supervised experiment settings, 136 images from 7 patients and 68 images from 3 patients in the training set are labeled respectively. See SSL4MIS[1] for details.
- **ISIC dataset** [Codella *et al.*, 2018] includes 2594 dermoscopy images, and we use 1815 images for training and 779 images for validation. In training set, 5% (91) and 10% (181) images are labeled for different semi-supervised experiment settings.

All images in both datasets are resized to 224 × 224 to meet the input requirements of the proposed method. We use standard data augmentation to enlarge training set, including random cropping, random rotating, random flipping and color jittering. In our method, the category of transformation will be recorded to calculate the ℓ_e. In order to evaluate the performance of our method, we select Dice Coefficient (denoted as Dice) and Jaccard Index (denoted as Jaccard) as evaluation metrics.

Implementation Details. For fair comparisons, all the methods used in the experiments choose UNet as the benchmark architecture for image segmentation. We use ResNet-50 to replace the encoder part of UNet and initialize its parameters with the weights pre-trained on ImageNet. We adopt SGD as an optimizer with a weight decay of 0.0005 and a momentum of 0.9. The initial learning rate is set to 0.01, which will reduce to 0.001 by polynomial scheduler strategy during training. We implement the methods using PyTorch library and train them on a NVIDIA RTX 2080Ti GPU. The batch size is set to 16, where 8 images are labeled. All methods perform 6000 iterations during training.

4.2 Quantitative Comparison

Compared Methods. We compare our method with some recent semi-supervised segmentation methods including: Meat-Teacher (MT) [Tarvainen and Valpola, 2017], Entropy Minimization (EM) [Vu *et al.*, 2019], Uncertainty-Aware Mean Teacher (UA-MT) [Yu *et al.*, 2019], Deep Co-training (DCT) [Qiao *et al.*, 2018], Cross-Consistency Training (CCT) [Ouali *et al.*, 2020] and Cross Pseudo Supervision (CPS) [Chen *et al.*, 2021a]. For all comparison methods, we adopt the official hyper-parameter settings.

Main Results. Table 1 shows our quantitative comparative experimental results on ACDC and ISIC datasets. The first row represents the performance of the baseline model trained with labeled data only. Compared with the baseline model,

[1]https://github.com/HiLab-git/SSL4MIS

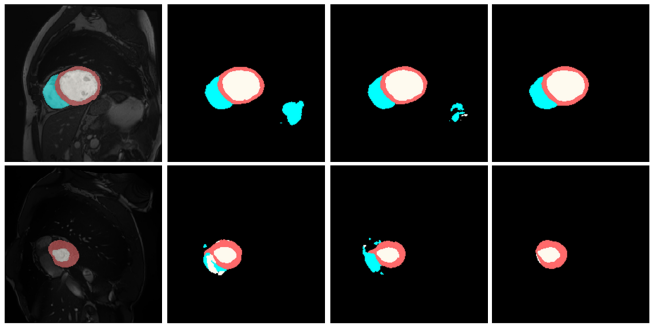

Figure 3: Visual comparison of the segmentation results produced by different methods on ACDC dataset. The proposed method predicts the best results.

our method can effectively use unlabeled data to achieve great performance. In different datasets and different semi-supervised settings, our proposed method obviously outperforms the comparison methods. Especially, when only 68 labeled images are used in ACDC dataset, our method improves Dice by more than 3% compared with the other methods.

Visual Comparisons. Fig.3 shows some visual comparisons between different methods when using 136 labeled images on ACDC dataset. We chose two methods MT and CPS which performed better in the experiment for comparison. Compared with MT and CPS, our method has better prediction results and less false predictions.

4.3 Ablation Study

Table 2 shows the results of ablation experiments of our method on ACDC dataset with 136 labeled images. We choose the UNet model that only uses labeled data for supervised training as the baseline (first row), and gradually increase the proposed components to prove their effectiveness. Besides, we further add two additional comparison settings, including 1) using only labeled data for contrastive learning (ℓ_c^{sup}) and 2) using pseudo labels for contrastive learning (ℓ_c^{pseudo}) to demonstrate the effectiveness of our proposed uncertainty-guided contrastive learning method. The experimental results show that each part of our proposed method has a positive impact. The introduction of contrastive learning effectively establishes the relationship between pixels and improves the performance of the model. Compared with the pseudo label scheme, the proposed method makes full use of unlabeled data and hence brings significant performance improvements (Dice increases by about 1%).

Contrast on Different Feature Scales. The resolution of different features has an important impact on the selection of contrastive learning samples. To find an appropriate feature scale, we explore the effect of different scales for contrastive learning on the ACDC dataset. Table 3 shows the results of contrastive learning under four different feature scales. As we can see, the performance of contrastive learning under low resolution ($Conv_4$) is poor, which might be caused by the semantic inconsistency from label downsampling. The higher resolution feature ($Conv_1$) also brings performance degradation. The potential reason is that the high-resolution pixel vectors contain less semantic information. Therefore, we believe that using middle-level features for contrastive learning can bring better segmentation performance.

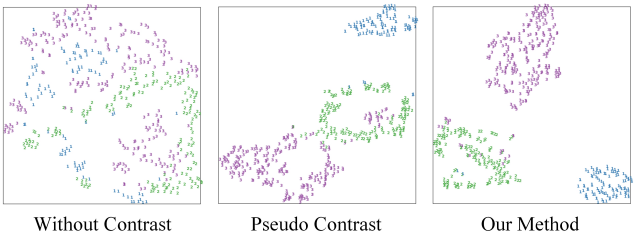

Figure 4: Visualization results of the pixel features obtained by different methods on ACDC dataset. The dimension of pixel features is reduced by t-SNE algorithm. Colors represent pixel categories.

Visualization of Features. In Fig.4, we use t-SNE algorithm to reduce the dimension of pixel features for visualization. From left to right, they are the results of training without contrastive learning, contrastive learning using pseudo labels and the proposed method. Compared with the first one, the proposed method can equip the pixel representations with better intra-class compactness and inter-class separability, which indicates the effectiveness of contrastive learning for segmentation task. Compared with the second one, our method has better aggregation results, the potential reason of which is that our method can reduce the possibility of noise sampling.

5 Conclusion

We propose an uncertainty-guided pixel contrastive learning method for semi-supervised medical image segmentation, which uses uncertainty to solve the noise sampling problem of unlabeled data in pixel contrastive learning. To estimate uncertainty, a heterogeneous consistency learning strategy is elaborately designed based on the decoders of CNN and Transformer. In addition, we construct an equivariant contrastive loss to strengthen the global representation learning ability of our model. Extensive experiments demonstrate that our method can achieve the state-of-the-art performance.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China under Grant 61976145 and Grant 61802267, in part by the Shenzhen Municipal Science and Technology Innovation Council under Grants JCYJ20210324094413037 and JCYJ20190813100801664, and in part by the Guangdong Basic and Applied Basic Research Foundation 2021A1515011318.

References

[Bernard *et al.*, 2018] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning

techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

[Cao *et al.*, 2021] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv:2105.05537*, 2021.

[Chaitanya *et al.*, 2020] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *arXiv:2006.10511*, 2020.

[Chen *et al.*, 2021a] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021.

[Chen *et al.*, 2021b] Zejian Chen, Wei Zhuo, Tianfu Wang, Wufeng Xue, and Dong Ni. Bootstrap representation learning for segmentation on medical volumes and sequences. *arXiv:2106.12153*, 2021.

[Codella *et al.*, 2018] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. 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 (ISBI 2018)*, pages 168–172, 2018.

[Dangovski *et al.*, 2021] Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljačić. Equivariant contrastive learning. *arXiv:2111.00899*, 2021.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[Hu *et al.*, 2021] Xinrong Hu, Dewen Zeng, Xiaowei Xu, and Yiyu Shi. Semi-supervised contrastive learning for label-efficient medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 481–490, 2021.

[Laine and Aila, 2016] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv:1610.02242*, 2016.

[Li *et al.*, 2021] Shaohua Li, Xiuchao Sui, Xiangde Luo, Xinxing Xu, Yong Liu, and Rick Goh. Medical image segmentation using squeeze-and-expansion transformers. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 807–815, 8 2021.

[Luo *et al.*, 2021] Xiangde Luo, Minhao Hu, Tao Song, Guotai Wang, and Shaoting Zhang. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. *arXiv:2112.04894*, 2021.

[Ouali *et al.*, 2020] Yassine Ouali, Céline Hudelot, and Myriam 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.

[Qiao *et al.*, 2018] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision*, pages 135–152, 2018.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015.

[Tarvainen and Valpola, 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv:1703.01780*, 2017.

[van den Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018.

[Verma *et al.*, 2019] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 3635–3641, 7 2019.

[Vu *et al.*, 2019] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

[Wang *et al.*, 2021] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. *arXiv:2101.11939*, 2021.

[Yu *et al.*, 2019] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–613, 2019.

[Zheng and Yang, 2021] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, 129(4):1106–1120, 2021.

[Zhong *et al.*, 2021] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7273–7282, 2021.