SELF-SUPERVISED PRE-TRAINING BASED ON CONTRASTIVE COMPLEMENTARY MASKING FOR SEMI-SUPERVISED CARDIAC IMAGE SEGMENTATION

Yubo Zhou[1], *Ran Gu*[1], *Shaoting Zhang*[1,2,3], *Guotai Wang*[1,2]

[1] School of Mechanical and Electrical Engineering,
University of Electronic Science and Technology of China, Chengdu, China
[2] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[3] SenseTime Research, Shanghai, China

ABSTRACT

Cardiac structure segmentation is important for heart disease diagnosis, and deep learning with a large number of annotations has obtained remarkable performance on this task. Semi-Supervised Learning (SSL) has the potential to reduce annotation costs. However, most SSL methods only leverage unlabeled images for consistency regularization or pseudo labels, ignoring their potential for feature learning with pre-training. In this work, we propose a novel framework that utilizes self-supervised pre-training for better semi-supervised segmentation. Our framework consists of two modules: 1) Self-supervised pre-training based on Contrastive Complementary Masking (CCM), where a contrastive loss is used for two networks that encode complimentary masked versions of the same input, in addition to a reconstruction loss to enhance global and local feature learning; 2) Semi-supervised segmentation with Cross Pseudo Supervision (CPS) between the two pre-trained networks, where each network is supervised by pseudo labels from the other to deal with unlabeled images. Experiments on the ACDC dataset showed that our method improved performance by 6.73 percentage points over baseline with a 5% annotation ratio, and outperformed three state-of-the-art semi-supervised methods.

***Index Terms*—** Self-supervised learning, Semi-supervised segmentation, Contrastive learning

1. INTRODUCTION

Cardiac Magnetic Resonance (CMR) imaging captures different image contrasts that are sensitive to cardiac physiologies and pathologies. Automatic segmentation of CMR image plays an important role in heart disease diagnosis [1]. Recently, deep learning models have achieved remarkable performance on CMR image segmentation when trained with a large number of annotations [2]. However, providing pixel-level annotations for CMR images is time-consuming and demands expertise, as the boundary between different tissues is complex. Semi-supervised training and self-supervised pre-training are two efficient techniques to improve model performances by leveraging unlabeled data, which can largely reduce the annotation cost [3].

Semi-supervised learning trains a model with a small set of annotated data and a large set of unannotated data to reduce the annotation cost. Existing semi-supervised methods can be roughly divided into two categories: consistency-based [4–6] and pseudo-label-based [7] methods. Despite that these methods can leverage unannotated images for performance improvement, they typically trained the models from scratch with random initialization, thereby ignoring the potential benefits of pre-training. Some other works [8, 9] have explored the application of transfer learning to semi-supervised training, but they either relied on supervised pre-training that cannot effectively leverage the unlabeled data [9], or only trained an encoder [8], which might not be well-suited for segmentation tasks.

On the other hand, self-supervised learning can effectively leverage a set of unlabeled images for feature learning via a pretext task that does not require human annotation. Recently, contrastive learning [10–12] and image reconstruction-based pretext tasks [13–15] have been shown effective for feature learning that is transferable to downstream tasks such as image classification, detection [8] and segmentation [16]. However, most of them focused on fully supervised downstream tasks that still require a large amount of annotations. Although Zhang *et al.* [8] applied self-supervised pre-training to a semi-supervised downstream task, they only pre-trained an encoder for object detection, while self-supervised pre-training strategies for semi-supervised medical image segmentation have rarely been investigated that has a potential to further improve the model's performance given a low annotation cost.

To better utilize unlabeled images for improving the segmentation performance, we propose a novel framework that utilizes self-supervised pre-training for semi-supervised cardiac image segmentation. A novel self-supervised pre-training method based on Contrastive Complementary Mask-

Corresponding author: Guotai Wang (guotai.wang@uestc.edu.cn).

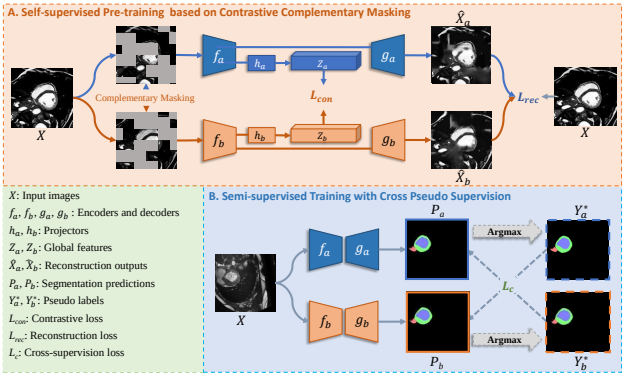

Fig. 1. Overview of our proposed framework that uses Contrastive Complementary Masking (CCM)-based pre-training for semi-supervised segmentation.

ing (CCM) is proposed, where two encoder-decoder networks are pre-trained for deriving visual representations of the dataset. The two networks accept two masked versions of the same input with complementary masks for contrastive feature learning and image reconstruction, and their features are encouraged to be consistent. In the downstream semi-supervised learning stage, the two pre-trained networks are trained by cross-supervision to leverage unlabeled images. Experiments on the Automated Cardiac Diagnosis Challenge (ACDC) [17] dataset demonstrated that our approach outperformed three existing semi-supervised and two self-supervised pre-training methods. When only 50% of the training data are labeled, the performance of our model is comparable to that of fully supervised learning.

2. METHODS

Our proposed framework is illustrated in Fig. 1. Firstly, we propose Contrastive Complementary Masking (CCM) to integrate masked image modeling and contrastive learning to pre-train two separate networks with the same architecture. The encoder generates global features for contrastive loss calculation, and the decoder outputs the reconstructed images for reconstruction loss calculation. Subsequently, the two pre-trained networks are leveraged for semi-supervised segmentation, where pseudo labels of unlabeled images are generated for cross-supervision.

2.1. Self-supervised Pre-training based on Contrastive Complementary Masking

Contrastive learning is an effective method for feature learning from unlabeled images [11]. However, most existing contrastive learning methods only learn a global feature representation in an encoder, which is insufficient for segmentation. To deal with this problem, we propose CCM that is more suitable for segmentation tasks. For an input image X_i, we first obtain a binary mask M^a by randomly setting half of all the blocks in the image as 0 and the other half as 1. Meanwhile, a complementary mask M^b is obtained by $M^b = 1 - M^a$. The masked images are denoted as $X_i^a = M^a \cdot X_i$ and $X_i^b = M^b \cdot X_i$, respectively. Note that X_i^a and X_i^b are complementary to each other, and they are sent into two different networks for contrastive feature alignment and image reconstruction. Let f_a and f_b denote the encoders of the two networks and g_a and g_b denote their decoders, respectively, the reconstructed images are denoted as $\hat{X}_i^a = g_a(f_a(X_i^a))$ and $\hat{X}_i^b = g_b(f_b(X_i^b))$. Given a batch size of N, the Mean Square Error (MSE) loss is used to compute the reconstruction loss between X_i and $\hat{X}_i^a$, $\hat{X}_i^b$:

$$L_{rec} = \frac{1}{2N}\sum_{i=1}^{N}(||X_i - \hat{X}_i^a||_2^2 + ||X_i - \hat{X}_i^b||_2^2) \quad (1)$$

In addition, we introduce projectors h_a and h_b after f_a and f_b, respectively. They obtain two global feature representations $Z_i^a = h_a(f_a(X_i^a))$ and $Z_i^b = h_b(f_b(X_i^b))$ from the two encoders, respectively. Given a batch size of N images, we obtain $2N$ feature representations. Z_i^a and Z_i^b from the same input X_i form a positive pair, and the other $2(N-1)$ features within the minibatch are negative counterparts for Z_i^a. The infoNCE [18] is used for contrastive loss:

$$L_{con} = -\log \frac{exp(sim(Z_i^a, Z_i^b)/\tau)}{\sum_{z\in\mathcal{N}\&z\neq Z_i^a} exp(sim(Z_i^a, z)/\tau)} \quad (2)$$

where $sim(\cdot,\cdot)$ is the cosine similarity between two features, $\mathcal{N}$ is the set of $2N$ features in the mini-batch, and τ denotes a temperature parameter. The whole training loss of self-supervised pre-training is:

$$L_{self} = L_{rec} + \lambda L_{con} \quad (3)$$

where λ is the trade-off weight. L_{rec} facilitates the learning of local information within individual images through image reconstruction, and L_{con} helps the network extract global information through instance discrimination.

2.2. Semi-supervised Segmentation with Cross Pseudo Supervision

To effectively leverage both of the two pre-trained networks, we employ a Cross Pseudo Supervision (CPS) strategy [7] for the downstream semi-supervised learning for two reasons: 1) CPS can be seamlessly integrated with our CCM module that provides two different networks initialized with the pre-trained weights; 2) The two networks have different decision boundaries and cross supervision between them is more robust to noisy pseudo labels, which overcomes the bias of a single model learning from its own pseudo labels.

Table 1. Dice (%) of different semi-supervised methods and self-supervised pre-training for cardiac segmentation.

Method	Annotation ratio: 5%				Annotation ratio: 10%			
	RV	Myo	LV	Avg	RV	Myo	LV	Avg
Full supervision	89.06±7.81	88.70±3.95	93.40±6.09	90.39±4.86	89.06±7.81	88.70±3.95	93.40±6.09	90.39±4.86
Baseline	72.74±22.14	79.11±9.62	87.93±10.36	79.93±11.83	85.18±8.05	83.44±5.23	89.51±8.43	86.04±5.00
CPS [7]	83.68±14.42	81.78±11.20	89.84±10.36	85.10±11.06	86.99±7.87	83.56±6.00	89.50±9.51	86.68±5.64
URPC [4]	76.72±18.69	78.79±10.87	86.20±12.13	80.57±11.87	87.87±8.11	82.98±7.22	91.18±6.75	87.34±4.80
UAMT [6]	74.19±21.24	81.30±8.36	89.14±8.23	81.54±11.14	86.24±8.02	82.61±6.80	89.74±9.19	86.20±5.59
CPS + SimCLR [11]	84.16±9.05	**82.82±7.14**	89.38±8.82	85.46±6.31	88.21±6.95	85.46±4.70	91.14±8.86	88.27±4.87
CPS + MG [13]	78.84±19.24	81.05±9.19	88.45±10.74	82.78±11.14	87.26±10.07	85.42±4.60	**92.35±4.58**	88.34±5.04
Ours	**87.21±7.21**	82.71±8.24	**90.05±9.57**	**86.66±6.09**	**89.16±6.83**	**85.78±4.55**	91.54±6.81	**88.83±4.43**

For semi-supervised learning, we use $\mathcal{D}^l$ and $\mathcal{D}^u$ to denote the labeled and unlabeled subsets, respectively. For an image X, the two pre-trained networks obtain two probability maps P_a and P_b, respectively. If X is from $\mathcal{D}^l$, we use Y to denote the ground truth label, and the supervised loss L_s is:

$$L_s = \frac{1}{2|\mathcal{D}^l|} \sum_{(X,Y)\in D^l} \big(\ell_{ce}(P_a, Y) + \ell_{ce}(P_b, Y)\big) \quad (4)$$

where ℓ_{ce} is the standard pixel-wise cross-entropy loss.

For an image X from $\mathcal{D}^u$, we apply argmax to P_a and P_b to obtain the one-hot pseudo labels Y_a^* and Y_b^*, respectively. Y_a^* is used to supervise the second network, and Y_b^* is used to supervise the first network. The cross pseudo supervision loss L_c is formulated as:

$$L_c = \frac{1}{2|D^u|} \sum_{X\in\mathcal{D}^u} \big(\ell_{ce}(P_a, Y_b^*) + \ell_{ce}(P_b, Y_a^*)\big) \quad (5)$$

The final loss for semi-supervised training is:

$$L_{semi} = L_s + \alpha L_c \quad (6)$$

where α is the trade-off weight for cross-supervision loss.

3. EXPERIMENTS AND RESULTS

3.1. Data and Implementation

We validated the effectiveness of our method on the ACDC dataset [17] which contains 200 annotated short-axis cardiac cine-MRI scans from 100 subjects. Each volume has 6-21 slices with an inter-slice spacing ranging from 5.0-10.0 mm. The matrix size ranges from 154×154 to 428×512. We randomly divided the scans at the patient level into 70%, 10% and 20% for training, validation and testing, respectively. Each slice was resized to 256×256, and normalized by z-score. We compared the performance of different methods trained with an annotation ratio ranging from 5% to 50%. The segmentation performances of the Right Ventricle (RV), Left Ventricle (LV) and Myocardium (Myo) were quantitatively evaluated by the Dice Similarity Coefficient (DSC) and the Average Symmetric Surface Distance (ASSD).

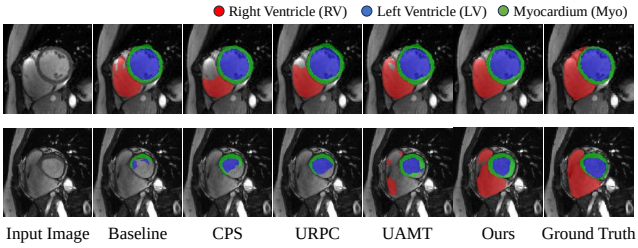

Fig. 2. Visual comparison of different semi-supervised methods for cardiac image segmentation.

Due to the relatively small number of subjects in the ACDC dataset, the segmentation networks were implemented by U-Net with a ResNet50 backbone [19], and we added a projector consisting of two linear layers after each encoder to obtain global features with a dimension of 1000. For self-supervised pre-training, we used the Adam optimizer with an initial learning rate of 0.001 and applied the cosine annealing learning rate policy. The epoch number was 400, and the batch size was 40, λ was set to 1, and the temperature coefficient τ was 0.07 according to SimCLR [11]. The block size for CCM was 16×16. For semi-supervised segmentation, we set α to 0.1, with an epoch number of 300 and a batch size of 24. The learning rate was halved when the loss did not decrease for 40 epochs. The other hyperparameters were set as the same values as those in the self-supervised pre-training.

There was no additional data augmentation applied in self-supervised pre-training except for complementary masking. For semi-supervised learning, to alleviate overfitting, we employed random flipping, rotation, crop, and gamma correction, gaussian noise to augment the images. No post-processing was used for the segmentation results. All the compared methods were implemented with PyTorch and PyMIC [20]. All the experiments were conducted on a computer with one NVIDIA GeForce RTX 3090 GPU.

3.2. Comparison with Existing Methods

We compared our method with three existing semi-supervised methods: CPS [7], Uncertainty Rectified Pyramid Consistency (URPC) [4] and Uncertainty-Aware Mean Teacher

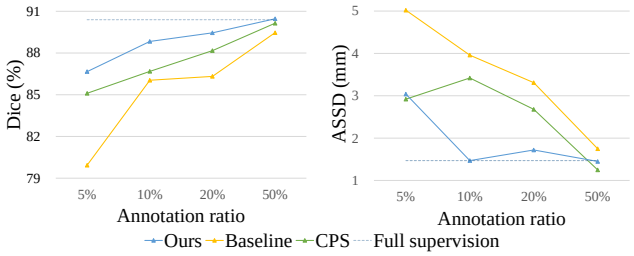

Fig. 3. Dice (%) and ASSD (mm) of methods with different annotation ratios.

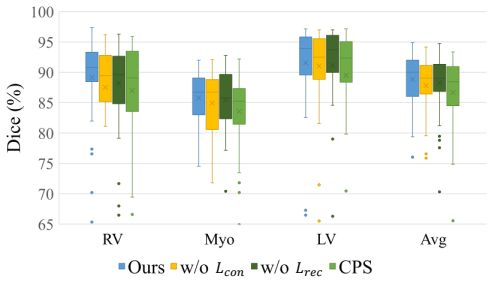

Fig. 4. Ablation study for proposed methods with an annotation ratio of 10%.

(UAMT) [6] as well as two existing self-supervised pre-training methods: SimCLR [11] and MG (Models Genesis) [13]. They were also compared with the baseline of learning only from the annotated images, and supervised learning with a 100% annotation ratio (i.e., full supervision).

The quantitative comparison under two different annotation ratios (5% and 10%) is shown in Table 1. Among the existing methods, URPC [4] obtained the best performance when the annotation ratio was 10%, but it performed the worst when the annotation ratio was 5%. This is mainly because consistency-based methods with a single network were easier to overfit to the small amount of labeled data. In contrast, CPS [7] and our method performed more stable by using two networks to generate pseudo labels for cross-supervision. Besides, compared with random initialization of the two networks in CPS [7], our method with self-supervised pre-training achieved the best performance at both annotation ratios. A visual comparison between these methods with a 10% annotation ratio is shown in Fig. 2. In the first case, CPS has an under segmentation of the RV, and our method segmented the RV successfully. In the second case, all the existing methods missed some part of the Myo and the RV, and our result is closer to the ground truth.

We also conducted performance comparison across various annotation ratios from 5% to 50%, and the results are shown in Fig. 3. It can be observed that our method largely outperformed both the baseline and CPS [7] under different annotation ratios. It's worth noting that when using 50% labeled data, the average Dice of our method reaches that of fully supervised learning (90.46% vs 90.14%).

3.3. Ablation Study

For ablation study, we evaluated the impact of pre-training without contrastive learning or without reconstruction with an annotation ratio of 10%. The experimental results in Fig. 4 demonstrate that relying solely on one of the two modules only leads to a slight improvement from CPS [7] for the semi-supervised cardiac image segmentation. In contrast, by applying the proposed CCM strategy, the model can learn both global and local features of the data, which significantly surpassed CPS [7] without pre-training.

For each class of the segmentation task, our method achieved the best segmentation performance. The CPS [7] without pre-training obtained an average Dice of 86.68%, and our method improved it to 88.83%, which was higher than pre-training without the contrastive loss L_{con} (87.83%) and pre-training without the reconstruction loss L_{rec} (88.27%), thus proving the effectiveness of CCM strategy.

4. CONCLUSION

In conclusion, we have introduced a novel self-supervised pre-training-based framework for semi-supervised cardiac image segmentation. It comprises two critical modules. First, a Contrastive Complementary Masking (CCM) strategy is proposed to pre-train two networks from the entire training data without using human annotations, and it learns local features through image reconstruction tasks and global features through instance discrimination tasks. Subsequently, these two pre-trained networks are employed to generate pseudo-labels for cross-supervision in the downstream semi-supervised segmentation, which effectively leverages labeled and unlabeled data to enhance the segmentation performance. Experimental results on the ACDC [17] dataset demonstrate that by employing the self-supervised pre-training, the unannotated images are more effectively leveraged for improving performance of semi-supervised learning, and our method outperformed several state-of-the-art semi-supervised segmentation methods. Future research could be extending CCM-based pre-training strategy to other semi-supervised methods to assess its effectiveness. Moreover, it's worthwhile to assess the generalizability of our framework by applying it to additional datasets, broadening its scope and applicability.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access. Ethical approval was not required as confirmed by the license attached with the open-access data.

6. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant 62271115.

7. REFERENCES

[1] Caroline Petitjean and Jean-Nicolas Dacher, "A review of segmentation methods in short axis cardiac MR images," *Medical image analysis*, vol. 15, no. 2, pp. 169–184, 2011.

[2] Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert, "Deep learning for cardiac image segmentation: a review," *Frontiers in Cardiovascular Medicine*, vol. 7, pp. 25, 2020.

[3] Shanshan Wang, Cheng Li, Rongpin Wang, Zaiyi Liu, Meiyun Wang, Hongna Tan, Yaping Wu, Xinfeng Liu, Hui Sun, Rui Yang, et al., "Annotation-efficient deep learning for automatic medical image segmentation," *Nature communications*, vol. 12, no. 1, pp. 5915, 2021.

[4] Xiangde Luo, Guotai Wang, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Dimitris N Metaxas, and Shaoting Zhang, "Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency," *Medical Image Analysis*, vol. 80, pp. 102517, 2022.

[5] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *NeurIPS*, vol. 30, 2017.

[6] 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 *MICCAI*, 2019, pp. 605–613.

[7] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *CVPR*, 2021, pp. 2613–2622.

[8] Xuanye Zhang, Kaige Yin, Siqi Liu, Zhijie Feng, Xiaoguang Han, Guanbin Li, and Xiang Wan, "Self-and semi-supervised learning for gastroscopic lesion detection," in *MICCAI*, 2023, pp. 83–93.

[9] Abulikemu Abuduweili, Xingjian Li, Humphrey Shi, Cheng-Zhong Xu, and Dejing Dou, "Adaptive consistency regularization for semi-supervised transfer learning," in *CVPR*, 2021, pp. 6923–6932.

[10] Xinlei Chen and Kaiming He, "Exploring simple siamese representation learning," in *CVPR*, 2021, pp. 15750–15758.

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020, pp. 1597–1607.

[12] Yuanfan Guo, Canqian Yang, Tiancheng Lin, Chunxiao Li, Rui Zhang, Rong Wu, and Yi Xu, "Self supervised lesion recognition for breast ultrasound diagnosis," in *ISBI*. IEEE, 2022, pp. 1–4.

[13] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang, "Models genesis: Generic autodidactic models for 3D medical image analysis," in *MICCAI*, 2019, pp. 384–393.

[14] Guotai Wang, Jianghao Wu, Xiangde Luo, Xinglong Liu, Kang Li, and Shaoting Zhang, "MIS-FM: 3D medical image segmentation using foundation models pretrained on a large-scale unannotated dataset," *arXiv preprint arXiv:2306.16925*, 2023.

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022, pp. 16000–16009.

[16] Chenxin Tao, Xizhou Zhu, Weijie Su, Gao Huang, Bin Li, Jie Zhou, Yu Qiao, Xiaogang Wang, and Jifeng Dai, "Siamese image modeling for self-supervised vision representation learning," in *CVPR*, 2023, pp. 2132–2141.

[17] 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 TMI*, vol. 37, no. 11, pp. 2514–2525, 2018.

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

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[20] Guotai Wang, Xiangde Luo, Ran Gu, Shuojue Yang, Yijie Qu, Shuwei Zhai, Qianfei Zhao, Kang Li, and Shaoting Zhang, "Pymic: A deep learning toolkit for annotation-efficient medical image segmentation," *Computer Methods and Programs in Biomedicine*, vol. 231, pp. 107398, 2023.