SWS-NET: An Image Segmentation Framework For Chronic Wounds Based On Self-Supervised Learning

Jiyun Li*, Sheng Yang† and Chen Qian‡
School of Computer Science and Technology, Donghua University
Shanghai, China
Email: *jyli@dhu.edu.cn, †jensenyoungc@outlook.com, ‡chen.qian@dhu.edu.cn

***Abstract*—Automatic monitoring and evaluation of chronic wounds usually requires massive labeled data sets for segmentation training. Because of the high cost of time and labor, these data are usually difficult to obtain. In order to improve the segmentation effect of the wound image with a small number of labeled samples for training, this paper proposes an image segmentation framework based on self-supervision, summarize a relatively optimal pre-training task for chronic wound image segmentation, minimizes the redundancy between the symmetric network projection output by learning the feature information of two views generated by the same chronic wound image under different distortion transformation, and finally learns valuable knowledge that is conducive to the downstream wound image segmentation task. In addition, this framework also optimize the network structure and loss of the segmentation model. The experimental results show that after self-supervised learning pre-training with a full amount of unlabeled data, the segmentation framework can achieve significant improvement in precision(up to 8%), recall(up to 5%), and MIoU(up to 9%) by fine-tuning with only a small amount of labeled data, This can provide a clear optimization direction for the application of self-supervised learning to specific image segmentation.**

Keywords-Chronic Wound; Image Segmentation; Redundancy; Self-Supervised; Pre-Training

I. INTRODUCTION

Chronic wound is a kind of wound that cannot be healed in a short time through regular treatment. In wound healing, clinicians need to continuously measure and evaluate the wound to monitor the healing process and treatment effect. The development of computer technology has brought convenience to the whole work process. Among them, chronic wound image segmentation is an essential step in computer wound measurement. The current mainstream depth convolution neural network (DCNN) performs well in medical image segmentation [1-2].

However, this model based on supervised learning needs a lot of labeled data to train a model with good results. In chronic wounds, obtaining data sets with large-scale pixel labeling is time-consuming, labor-consuming, expensive, and requires a professional clinical experience. In order to reduce the burden of annotation, many methods other than supervised learning have been proposed to improve the labeling efficiency of medical imaging, including semi-supervised learning [3-4], self-supervised learning [5-8], transfer learning [9-10], in which self-supervised learning has obvious advantages over other methods. There is no need for any labeling data during the pre-training phase. It can learn the useful representation of unlabeled data through pre-training tasks to better solve the problem of scarcity of labeled data. Many articles have proved its effectiveness [11-14] on well-known public data sets (such as ImageNet [15]). However, the related research on specific image segmentation, especially chronic wound images, is still relatively rare. So we plan to introduce a self-supervised learning method, targeted optimization for chronic wounds, to improve the segmentation effect when the data on chronic wounds are scarce.

The main contribution of this research is to propose a network framework for chronic wound image segmentation based on self-supervised learning, which we named SWS-NET. In this framework, many unlabeled data are used in the pre-training phase to get useful knowledge for downstream segmentation tasks, which is learned by comparing the similarity of the same image under different distortions from the unlabeled data. Compared with the mainstream segmentation models (such as Unet), it improves the segmentation precision of chronic wound images in the case of sparsely labeled data. In this process, we focus on the effect of the pre-training model trained by different distortion methods on the image segmentation task of chronic wounds. Finally, we summarize a relatively optimal pre-training task for chronic wound image segmentation.

We subsequently developed a small program to realize the segmentation application of chronic wound images on mobile devices such as mobile phones. Therefore, we also made some lightweight modifications to the convolution module of the segmentation network.

II. RELATED WORK

A. Image Segmentation

Currently, the encoder-decoder structure is one of the most popular end-to-end image segmentation frameworks. The full convolution network (FCN) [1] based on this structure has achieved relatively successful image segmentation results, but it reduces the interpretability of the model, resulting in poor segmentation results for specific types of images (such as medical images). With further research, Ronneberger et al [2]. proposed an encoder-decoder structure with high applicability, namely U-Net, which provides a more reliable backbone network for wound image segmentation framework. Deeplab [16] is similar to it, but the network structure will be much more complex, which is not convenient for subsequent improvement.

DOI reference number: 10.18293/SEKE2023-134

Since Unet was proposed, it has been concerned and applied by many scholars, such as Attention Unet [17] proposed by O. Oktay et al., and the initial convolution layer introduced by Narinder Singh Punn et al. [18], which further strengthens feature extraction. However, this leads to time-consuming training process, and when annotation data of chronic wound images are scarce, it is easy to over-fit. After that, Francois Chollet et al. proposed an RCA-IUnet network structure [19], which discusses and uses depthwise separable convolutions, integrates the advantages of attention filter, mixing pool, and initial convolution layer, reduces the complexity of convolution neural network structure, and provides a good solution for the segmentation model of this article.

B. *Self-supervised Learning*

Given the lack of annotation samples, more and more scholars have focused on the research of self-supervised learning. The agent tasks of early self-supervised learning are usually geometric transformation prediction, flip, rotation angle prediction [20], and jigsaw puzzle [21]. These methods in the initial phase of self-supervision need a well-defined task to be effective for a specific image, and reasonable constraints[22] are needed to prevent the model from obtaining trivial constant (i.e., collapsed) embeddings.

At present, the well-known contrastive learning methods are MoCo [23], SimCLR [13], BYOL [12], SimSiam [24], and Barlow Twins [25]. These methods adopt specific methods to avoid model collapse, which increase the complexity of the model, and rely heavily on the comparison of negative samples. Barlow Twins proposed a new optimization direction to avoid these problems. Inspired by biology [26], remove redundancy between networks, which brings a lot of inspiration to our framework.

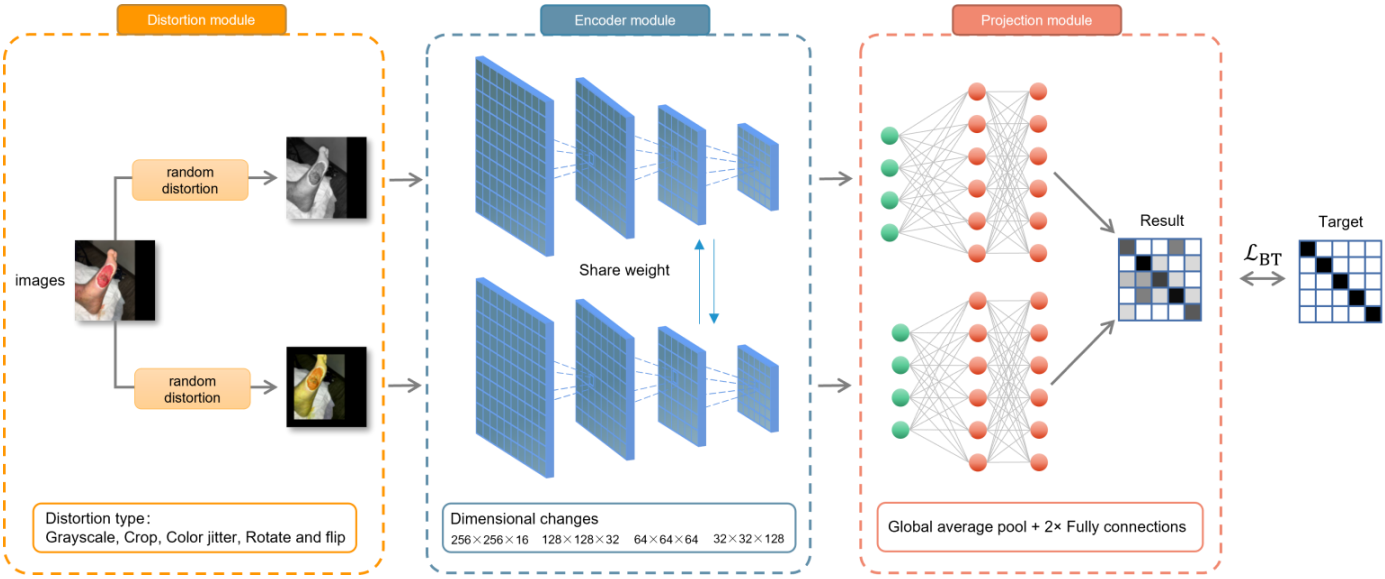

Figure 1. The overall architecture of the pre-training phase

III. METHOD

As shown in Figure 1, we have constructed a self-supervised learning framework to explore and learn the multi-dimensional features of unlabeled data sets by pre-training the encoder of the segmentation model and finally fine tune the pre-trained encoder with a small amount of labeled data downstream. This framework is mainly composed of several parts: First, the distortion module generates two distortion views for all images in the same batch sampled from the data set; the second is the encoder module, which will be used for pre-training and applied to the downstream segmentation task; then there is the projection module, which is responsible for projecting the three-dimensional features output by the encoder into the one-dimensional space, and narrowing the distance between positive examples in this space. Finally, the decoder module is used to fine-tune the downstream segmentation tasks and restore the prediction mask image through upsampling.

A. *Self-Supervision Phase*

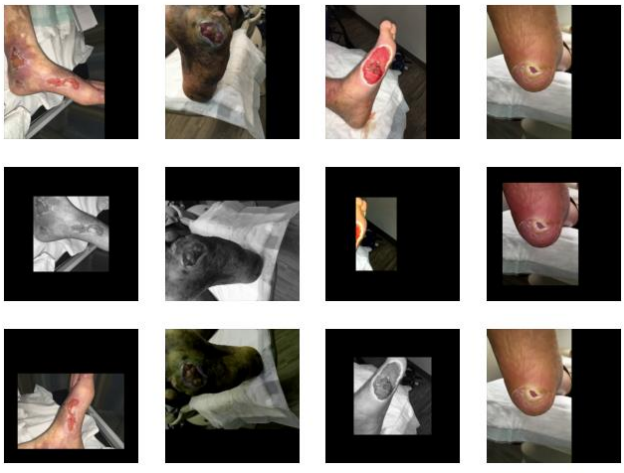

Figure 2. Example of pre-task in pre-training phase (the top is the original image)

At this phase, we have set up four image distortion methods in the distortion module: rotation&flip, grayscale, crop, and color jitter. By assigning different weights to the probabilities of these distortion modes, we made targeted optimization on the chronic wound image and finally obtained an optimal distortion strategy. The processed image effect is shown in the Figure 2.

The decoder is modified from the original Unet framework (see Figure 3). After the coder trains two distorted pictures with shared parameters, the training data is projected into a one-dimensional space. The projection network consists of two layers, each containing full connection, ReLU activation, and batch normalization processing. Finally, we use the loss to make the correlation matrix of the features from different perspectives close to the identity. The definition of loss as expressed in Eq. 1 and 2.

$$\mathcal{L}_{BT} = \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2 \tag{1}$$

$$C_{ij} \triangleq \frac{\sum_b z^A_{b,i} z^B_{b,j}}{\sqrt{\sum_b (z^A_{b,i})^2} \sqrt{\sum_b (z^B_{b,j})^2}} \tag{2}$$

where λ is a constant, weighing the importance of the first and second terms of loss, and C is a cross-correlation matrix calculated along the batch dimension between the outputs of two identical networks. b represents the batch index of the input sample, and i, j represents the network output's vector dimension. Ultimately, the value of C will range from -1 to 1, -1 means completely irrelevant, and 1 means complete correlation.

After we use many unlabeled chronic wound data sets for the pre-training of contrastive learning, the improvement effect is still relatively limited compared to classification tasks. From the related paper [27-29], we can know that this is because the comparison of different distortions of the same image is more focused on extracting the global representation. However, it is constrained to improving segmentation effect, such as pixel-by-pixel prediction tasks. To solve this problem, Xiangyun Zhao et al. [30] proposed a pre-training feature extractor using pixel-by-pixel, label-based contrast loss. Through experiments, we found that the loss of Barlow Twins is also suitable for this pre-training strategy, After application, the segmentation effect has been further improved.

B. Segmentation Training Phase

Our segmentation framework is shown in Figure 3, which can be divided into two parts: the encoder and the decoder. The encoder is divided into four downsamples, and the decoder is the corresponding four upsampling. Then, by splicing the fragments in the upsampling process and the downsampling process, The stitching retains more dimension and location information, allowing the following neural network layer to freely choose between shallow and deep features, which is more advantageous for semantic segmentation tasks.

In the convolution module part, each module is composed of two depth separation convolutions [31-32] and one ordinary convolution. We use deep separation convolution to replace the ordinary convolution in the original Unet framework, which makes the network more lightweight and convenient for our fine-tuning and application on mobile devices.

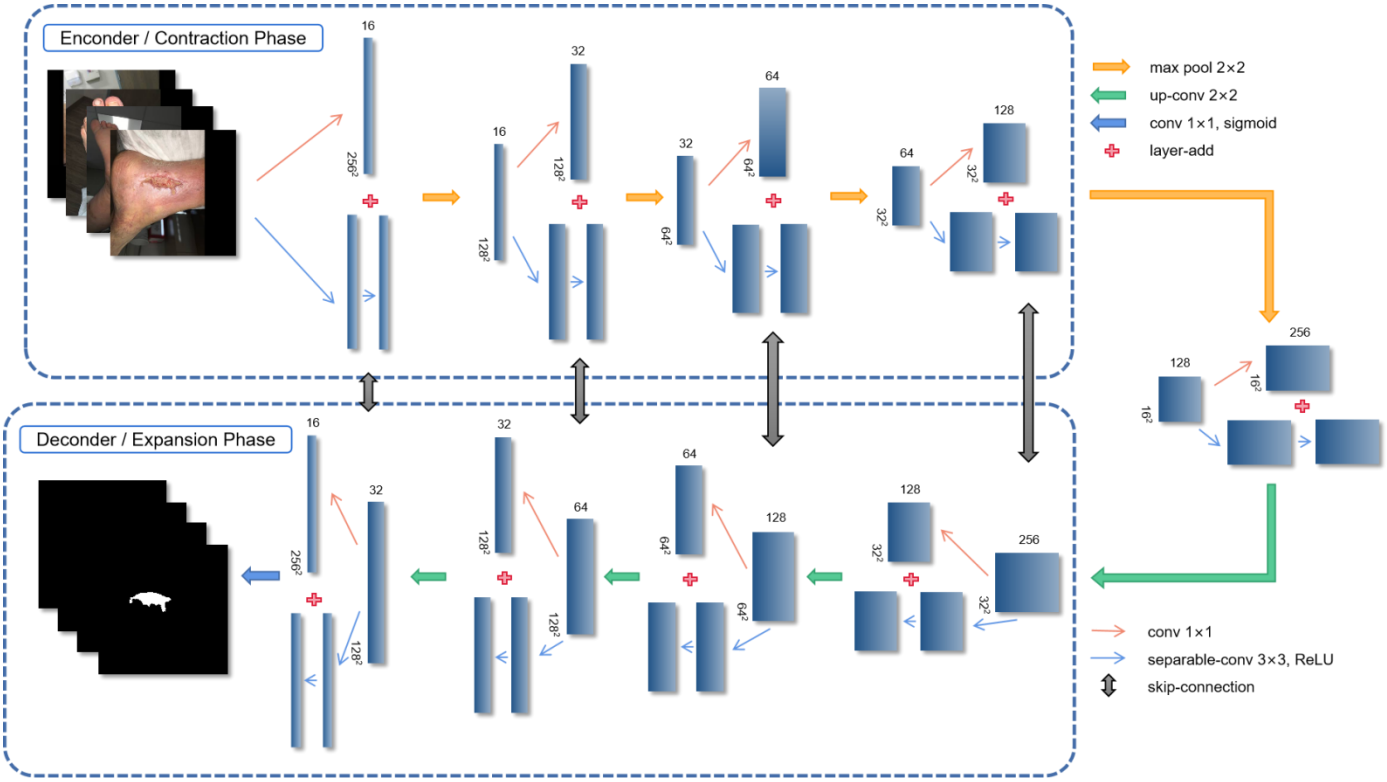

Figure 3. The architecture of the segmentation model

In the loss function part, we use the loss, which combines the binary cross entropy and the dice coefficient, as shown in Eqs 3, 4, and 5. However, the cross entropy loss will be dominated by the class with more pixels for the wound image that only accounts for a small part of the background area. For smaller objects, it is difficult to learn their characteristics, thus reducing the effectiveness of the network. To alleviate this problem, we introduce the Dice coefficient loss. The Dice coefficient calculates the intersection ratio between the segmented prediction result area and the ground truth area, neglects a large number of background pixels, and solves the problem of imbalance between positive and negative samples.

$$\mathcal{L}_{\mathrm{BCE}} = -(1-\mathrm{y})\log(1-\mathrm{x}) - \mathrm{y}\log(\mathrm{x}) \tag{3}$$

$$\mathcal{L}_{\mathrm{Dice}} = 1 - \frac{2|\mathrm{X}\cap\mathrm{Y}|}{|\mathrm{X}|+|\mathrm{Y}|} \tag{4}$$

$$\mathcal{L}_{Final} = \lambda\mathcal{L}_{BCE} + \mu\mathcal{L}_{Dice} \tag{5}$$

In the above formula, x represents the actual pixel value, and y represents the predicted pixel value. X represents the number of pixels of the real wound segmented, and Y represents the number of pixels of the predicted wound because the intersection ratio will increase with the improvement of the effect. In order to conform to the optimization direction of the loss, a subtraction process is made, and the value is kept between 0 and 1. λ and μ represent the weight parameter of the loss function.

IV. EXPERIMENT

In this article, our experiment can be roughly divided into three stages: dataset processing, segmentation experiment, and ablation experiment: The processing of data sets is an important part of self-supervised learning, including the division of training sets and test sets and the setting of pre-tasks; In the part of segmentation experiment, we tested the segmentation effect under different amounts of data, and prove that our proposed framework can improve the accuracy obviously when the labeled data is scarce. Finally, in the ablation experiment part, we verified the influence of the models trained by different distortion methods on the experimental results.

A. Dataset processing

Our dataset is from the chronic wound dataset publicly available on Kaggle, with a total of 1010 chronic wound images and their corresponding segmented mask images. In order to use additional chronic wound data for self-supervised pre-training, we used the multi-category dataset of chronic wound images also publicly available on Kaggle, with a total of 2023 images, classified into diabetes foot ulcer, burns, normal, pressure ulcer, skin tear, surgical wound, trauma, and venous wound. It can balance different types of wound samples in the dataset and avoid the abnormal error of segmentation of a particular wound type caused by the scarcity of specific types of samples. We need to divide the data volume of pre-training data to verify its effect under different data volumes. The specific division is shown in the Table 1.

TABLE I. DIVISION OF DATASET IN THE SEGMENTATION TASK PHASE

	10%	20%	30%	40%	50%
Test Dataset	110	110	110	110	110
Training Dataset	74	144	216	289	360
Validation Dataset	18	36	54	72	90

TABLE II. EXPERIMENTAL RESULTS UNDER DIFFERENT DATA VOLUMES

Data volume	10%		20%		30%		40%		50%	
pre-training	has	no	has	no	has	no	has	no	has	no
Precision	0.622	0.581	0.784	0.704	0.867	0.817	0.914	0.897	0.934	0.941
Recall	0.403	0.311	0.582	0.543	0.709	0.654	0.771	0.737	0.831	0.776
MIoU	0.547	0.455	0.693	0.606	0.783	0.732	0.829	0.805	0.862	0.844

B. Segmentation experiment

The framework is based on a Python 3.8 environment and developed with the TensorFlow V2.6 library. Before the training starts, all the pictures are adjusted to 256×256, the batch size of each training step is set to 8, use the SGD optimizer with a learning rate of 1e-3 and momentum parameter of 0.9. The weight decay parameter of the full connection layer is set to 4e-3. Finally, we use 10%, 20%, 30%, 40%, and 50% data volumes to verify the effect of contrastive learning, and the results are shown in Table II.

We have selected precision, recall, and MIoU as our evaluation indicators, and their calculation formula is shown in Eqs 6, 7, 8:

$$\mathrm{Precesion} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{6}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{7}$$

$$\mathrm{MIoU} = \frac{1}{\mathrm{k}}\sum_{\mathrm{i}=1}^{\mathrm{k}} \frac{\mathrm{TP}}{\mathrm{FN+FP+TP}} \tag{8}$$

From the experimental results, we can see that compared with the model without self-supervised pre-training, the self-supervised framework we proposed can achieve a certain improvement under different fine-tuning data amounts such as 10%, 20%, 30%, 40%, 50%, and can achieve a more significant improvement when the data volume is low, for example, in the case of 20% data volume, it can achieve more than eight percentage points of improvement.

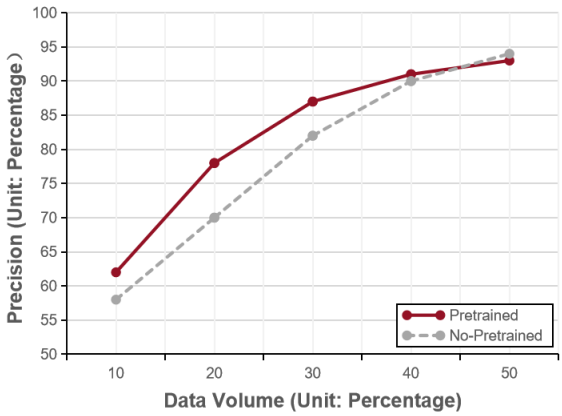

Figure 4. Precision comparison chart

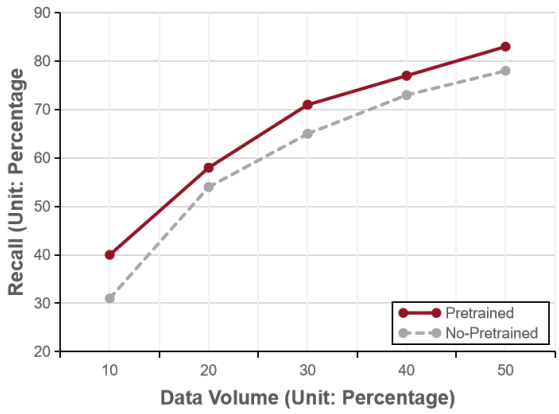

Figure 5. Recall comparison chart

By observing the Figures 4 and Figure 5, we can find that, on the whole, both the precision and the recall rate show more obvious advantages than the model without pre-training under small data volume, but the trend of its curve is still slightly different. With the increase in data volume, the improved range of precision shows a trend of increasing and decreasing. When the data volume is large enough, the precision may even be lower than the model directly trained by supervised learning. With the increase in the amount of data, the increase in recall shows a trend of decreasing at first and then increasing. There is a negative correlation between recall and accuracy. Since the two evaluation indicators, precision and recall rate, have their respective focus, in order to more fairly evaluate the improvement effect under different data volumes, we draw the improvement effect diagram of MIoU (see Figure 6). Through observation, our model has significantly improved when the data volume is small. Although the gap between the precision will be gradually narrowed after the data volume is increased, this more balanced indicator can prove the effectiveness of the self-supervised pre-training framework.

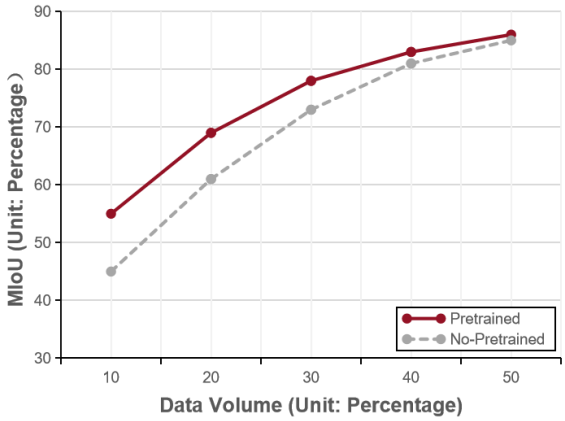

Figure 6. MIoU comparison chart

C. *Ablation experiment for pre-task*

Our current pre-task applies four mainstream pre-tasks: rotation&flipping, grayscale image, clipping, and color jitter. At the same time, a comparison between the original image and the mask image is added to improve further the learning effect of the pre-task on the segmented task. In this part of experiment, we gradually remove the superimposed distortion method to evaluate the accuracy impact. Through previous experiments, we can achieve more remarkable improvement in 20% of the labeled data sets, making the ablation experiment intuitive. So the ablation experiment is also carried out with 20% data volume. The specific results are shown in Figure 7.

Through vertical comparison, our improved Unet framework can consistently achieve higher segmentation precision than the original Unet framework, which shows that the primary performance of our improved segmentation model exceeds the original Unet model. At the same time, the gap between the original model and our model is gradually reducing with the gradual reduction of distortion. It can also be proved that some of the fine-tuning we have done are targeted optimization for these distortion tasks.

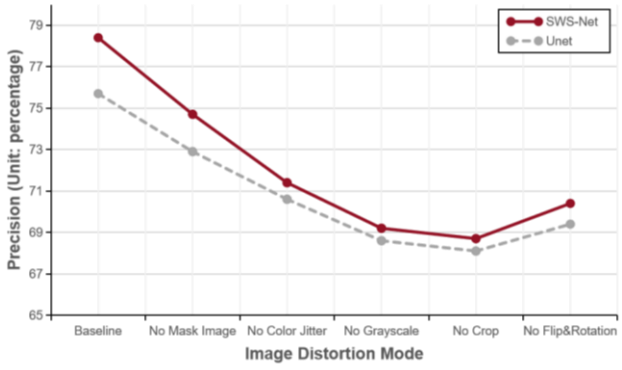

Figure 7. The influence of different image distortion on the result

Through horizontal comparison, we can see that the most important factor affecting the segmentation precision is the contrastive learning between the original and mask images. Moreover, in the segmentation task, the importance of grayscale is no less than the color jitter, which differs from the result of the previous different distortion methods on the classification task. In the classification task, the impact of color jitter accounts for a large proportion, leading to the model not being robust enough to remove some distortion methods. Therefore, a good pre-task is the key to the effectiveness of the self-supervised learning framework. The single-image distortion method may even have the opposite effect.

V. Conclusion And Prospect

We propose a network framework for chronic wound image segmentation based on self-supervised learning, alleviate the problem of the scarcity of chronic wound labeling data and provide a better segmentation accuracy for the segmentation of chronic wound images than the mainstream benchmark. In this process, we also discussed the effect of the pre-training model trained by different distortion methods on the downstream segmentation task. Through experiments, the effectiveness of our experiment has been proved, which provides reference methods and ideas for further research and application.

However, there are also some areas that need to be improved during the experiment. According to the experimental data, the recall rate has been at a relatively low level. What causes the low recall rate and what optimization methods can improve the recall rate. At present, there are two reasons to guess: one is the image quality problem in the data set, and the other is the loss function in the fine-tuning stage, which needs to be further explored by later generations.

References

[1] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” 2015.

[2] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” Cham: Springer International Publishing, 2015, pp. 234–241.

[3] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng, “Transformation-Consistent Self-Ensembling Model for Semisupervised Medical Image Segmentation,” *IEEE Trans. Neural Netw. Learning Syst.*, vol. 32, no. 2, pp. 523–534, 2021,

[4] S. Li, Z. Zhao, K. Xu, Z. Zeng, and C. Guan, “Hierarchical Consistency Regularized Mean Teacher for Semi-supervised 3D Left Atrium Segmentation,” 2021.

[5] Z. Zeng, Y. Xulei, Y. Qiyun, Y. Meng, and Z. Le, “SeSe-Net: Self-Supervised deep learning for segmentation,” *Pattern Recognition Letters*, vol. 128, pp. 23–29, 2019,

[6] Z. Zhou *et al.*, “Models Genesis: Generic Autodidactic Models for 3D Medical Image Analysis,” Cham: Springer International Publishing, 2019, pp. 384–393.

[7] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, “A Survey on Contrastive Self-Supervised Learning,” *Technologies*, vol. 9, no. 1, p. 2, 2020,

[8] J. Xu and National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, 03056, Ukraine, “A Review of Self-supervised Learning Methods in the Field of Medical Image Analysis,” *IJIGSP*, vol. 13, no. 4, pp. 33–46, 2021,

[9] Z. Zhao, K. Xu, S. Li, Z. Zeng, and C. Guan, “MT-UDA: Towards Unsupervised Cross-modality Medical Image Segmentation with Limited Source Labels,” Cham: Springer International Publishing, 2021, pp. 293–303.

[10] F. Zhuang *et al.*, “A Comprehensive Survey on Transfer Learning,” *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, 2021,

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” 2009.

[12] J.-B. Grill *et al.*, “Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning,” *Neural Information Processing Systems*, 2020.

[13] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” *International Conference on Machine Learning*, 2020.

[14] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, “Unsupervised Learning of Visual Features by Contrasting Cluster Assignments,” *Neural Information Processing Systems*, 2020.

[15] I. Misra and L. van der Maaten, “Self-Supervised Learning of Pretext-Invariant Representations,” 2020.

[16] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking Atrous Convolution for Semantic Image Segmentation,” *ArXiv*, 2017.

[17] O. Oktay *et al.*, “Attention U-Net: Learning Where to Look for the Pancreas,” *ArXiv*, 2018.

[18] N. S. Punn and S. Agarwal, “Inception U-Net Architecture for Semantic Segmentation to Identify Nuclei in Microscopy Cell Images,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 16, no. 1, pp. 1–15, 2020,

[19] S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised Representation Learning by Predicting Image Rotations,” *ICLR*, 2018.

[20] M. Noroozi and P. Favaro, “Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles,” Cham: Springer International Publishing, 2016, pp. 69–84.

[21] C. Doersch, A. Gupta, and A. A. Efros, “Unsupervised Visual Representation Learning by Context Prediction,” 2015.

[22] I. Goodfellow *et al.*, “Generative adversarial networks,” *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020,

[23] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum Contrast for Unsupervised Visual Representation Learning,” 2020.

[24] X. Chen and K. He, “Exploring Simple Siamese Representation Learning,” 2021.

[25] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, “Barlow Twins: Self-Supervised Learning via Redundancy Reduction,” *ICML*, 2021.

[26] H. B. Barlow, “Possible Principles Underlying the Transformations of Sensory Messages,” The MIT Press, 2012, pp. 216–234.

[27] K. Chaitanya, Ertunç Erdil, N. Karani, and E. Konukoglu, “Contrastive learning of global and local features for medical image segmentation with limited annotations,” *NeurIPS*, 2020,

[28] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. V. Gool, “Exploring Cross-Image Pixel Contrast for Semantic Segmentation,” 2021.

[29] Y. Xie, J. Zhang, Z. Liao, Y. Xia, and C. Shen, “PGL: Prior-Guided Local Self-supervised Learning for 3D Medical Image Segmentation,” *ArXiv*, 2020.

[30] X. Zhao *et al.*, “Contrastive Learning for Label Efficient Semantic Segmentation,” 2021.

[31] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” 2017.

[32] A. Howard *et al.*, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” *ArXiv*, 2017.