# appendix of SAM-SP: SAM-guided Semantic Alignment and Pseudo-label Fine-tuning for Scribble-Supervised Medical Image Segmentation

**Anonymous Submissions**

## Dataset and Evaluation Metric

**ACDC**: The ACDC dataset (Bernard et al. 2018) contains cardiac magnetic resonance imaging (CMR) data from 100 patients, with expert annotations for the left ventricle (LV), right ventricle (RV), and myocardium (Myo). Each patient's data includes images from both the end-diastolic (ED) and end-systolic (ES) phases of the cardiac cycle.

**MSCMR**: The MSCMR dataset (Zhuang 2018, 2019) consists of late gadolinium-enhanced (LGE) MRI scan data from 45 patients with cardiomyopathy, including expert annotations for the left ventricle (LV), right ventricle (RV), and myocardium (Myo). Compared to non-contrast MRI, LGE MRI poses more significant challenges for segmentation tasks (Zhang and Zhuang 2022). Since (Zhang and Zhuang 2022) provides manual scribble annotations only for the training set of 25 patients, we follow their 25:5:15 data split for training, validation, and testing.

**CHAOS T1 and T2**: The CHAOS dataset (Kavur et al. 2021) consists of multi-organ abdominal segmentation data from 20 patients, with two MRI sequences: T1-DUAL (in-phase and out-of-phase) and T2-SPIR, collected using a 1.5T Phillips MRI scanner. The dataset includes two MRI sequences: T1-DUAL (CHAOS T1) and T2-SPIR (CHAOS T2). CHAOS T1 includes in-phase and out-of-phase images, which are well-aligned and share a common ground truth, while CHAOS T2 has an independent ground truth. The annotations cover four target organs: liver (LIV), left kidney (LK), right kidney (RK), and spleen (SPL). (Yang et al. 2024) provides manual scribble annotations for both T1 and T2.

**Metrics**: The Dice coefficient reflects the overall segmentation quality and is particularly useful for evaluating the accuracy of the segmentation, especially when there is a strong focus on the size and shape of the target region. HD95, on the other hand, emphasizes the precision of the boundaries and is ideal for assessing the accuracy of segmentation results in terms of detail, especially in tasks that require accurate boundary localization. In practical applications, combining these two metrics provides a more comprehensive and accurate evaluation of model performance. Furthermore, most research in the field of scribble-supervised medical semantic segmentation is based on one or both of these evaluation metrics.



## Implementation Details

We implemented our method and other comparative methods based on PyTorch and ran them on a single RTX 4090 GPU. During the network training process, we normalized the input images to the range [0, 1] and resized them to 256×256 to fit the network input.

Our Segmentor uses DMPLS (Luo et al. 2022) as the base segmentation network architecture, and following its approach, a dropout layer (ratio = 0.5) is inserted before each convolutional block in one branch to introduce perturbations. Moreover, independent CSF modules are added at all skip connections between its encoder and decoder, and an output convolutional layer is introduced after the two decoders to replace the summation of random weights. To minimize the joint loss function for optimizing the segmentation model, we employed the SGD optimizer with a weight decay of 1e-4, a momentum of 0.9, and an initial learning rate of 1e-2. Additionally, we adopted a polynomial learning rate schedule for dynamic adjustment and set the batch size to 16.

For SAM, unless otherwise specified, we use the ViT-H model to generate segmentation masks and adopt the setting of 10 prompt points per category. To minimize the joint loss function for fine-tuning SAM, we use the AdamW optimizer with an initial learning rate of 5e-6 and a weight decay of 1e-4. For SAM fine-tuning, we freeze the weights of the Image encoder and Prompt encoder, and only unlock the weights of the Mask Decoder for training. For each iteration of SAM fine-tuning, we perform 30 epochs of training.

## Experiment

### Comparisons with Scribble-Supervised Methods

Due to space constraints in the main text, we are unable to provide results for all categories. Therefore, we have included these details in the appendix. Tables **??** and 2 present the per-category results of our method compared with other methods on the ACDC and MSCMR datasets. However, per-category results for the CHAOS dataset are not shown. This is because, although we have complete per-category results

for our own method, other methods do not report their per-category results. To avoid misleading readers, we have chosen not to present the per-category results for this dataset.

### Comparisons with SAM-Based Methods

In this section, we compare our method with SAM-based approaches, including SAM(Kirillov et al. 2023), MedSAM(Ma et al. 2023), and SAM-Med(Cheng et al. 2023). These methods all combine class-specific masks generated by Random Sample point prompts, using them as pseudo-label generators to supervise the segmentation network.

SAM(Kirillov et al. 2023): We evaluate both the smallest (ViT-B) and largest (ViT-H) versions of the Segment Anything Model (SAM) trained on natural images. SAM generates DMPLS using a random point sampling strategy and employs a multi-class aggregation method.

SAM-Med(Cheng et al. 2023): SAM-Med is a SAM ViT-B model enhanced with adapter layers in the image encoder. We use randomly sampled points as input prompts. Following (Cheng et al. 2023), we evaluate SAM-Med2D in two configurations: with and without adapter layers.

MedSAM(Ma et al. 2023): MedSAM is a SAM ViT-B model fine-tuned with bounding box prompts on 1.5M biomedical image–segmentation pairs. We evaluate MedSAM using bounding boxes only, as its performance with point or mask prompts was found to be suboptimal. The bounding boxes are computed as the minimum enclosing boxes of the masks, with each dimension expanded by [3, 7] pixels.

Table 3 presents a comparison between our method and other SAM-based methods. The difference from the tables in the main text is that this table shows the quality of pseudo-labels generated by each method, which further demonstrates the effectiveness of our fine-tuning approach.

### SAM-SP* Details

SAM-SP* adopts ViT-b weights as the initialization for iterative fine-tuning. On the ACDC dataset, the model is fine-tuned for two rounds, and the resulting weights are then used to generate pseudo-labels for MSCMR, CHAOS T1, and CHAOS T2 datasets, which in turn supervise the segmentation networks. Similarly, the model is fine-tuned on MSCMR for two rounds, and the resulting weights are used to generate pseudo-labels for ACDC. In this way, the iteratively tuned SAM weights are never exposed to the target datasets during training, ensuring a fair comparison with methods such as MedSAM and SAMMed.

### Data Sources

Regarding the results on the ACDC dataset in Table 1 (comparison table with scribble methods on ACDC and MSCMR), the values of pCE, RLoss, MLoss, S2L, USTM, DMPLS, and Unet are taken from the original study (Luo et al. 2022). The results of ScribbleVC and ScribFormer are derived from our own experiments, as their original papers did not adopt five-fold cross-validation. All other results in the table are directly cited from their respective original literatures.

For the results on the MSCMR dataset in Table 1 (comparison table with scribble methods on ACDC and MSCMR), the results of DMPLS, ScribbleVC, and Unet are cited from (Li et al. 2023), while the results of pCE, RLoss, MLoss, S2L, and USTM are obtained from our experiments. The results of the remaining methods are all based on the data provided in their original papers.

All results in Table 2 (comparison table with scribble methods on CHAOS) are directly cited from the original study (Yang et al. 2024).

In Table 3 (comparison table with other SAM-based methods), the results of SparseMamba-PCL are taken from its original literature, and the remaining results are all derived from our own experimental evaluations.

The data sources corresponding to the tables in our supplementary materials are the same.

### Visualization

Fig. 1 displays the visual comparison of results between our method and other SAM-based methods on ACDC, MSCMR, ChaosT1, and ChaosT2 datasets. We observe that in the displayed images, our method closely aligns with the ground truth (GT) compared to the others.

Fig. 2 displays the visual comparison of pseudo labels between our method and other scribble-supervised methods on ACDC. Since we have replicated more comparison methods on this dataset, we only present the visual results for this dataset. We observe that in the displayed images, our method closely aligns with the ground truth (GT).

Fig. 3 shows a comparison of the results before and after fine-tuning iterations on ACDC and MSCMR datasets. It is evident that, compared to the results before the iterations, the method after fine-tuning is much closer to the ground truth (GT).

## Limited discussion

Although our method achieves state-of-the-art (SoTA) results, there are some limitations with the iterative fine-tuning process of SAM-SP. First, iterative fine-tuning requires significant computational resources, especially when conducting multiple rounds of training on various datasets. This leads to substantial increases in computation time and memory consumption, which can become a bottleneck in resource-constrained situations. Therefore, optimizing the training strategy or using more powerful hardware may be necessary for practical applications to improve efficiency.

Furthermore, it is important to note that while SAM-SP performs well on multiple datasets, its multi-round iterative training approach means that we are not using an end-to-end model. Each iteration depends on the previous pseudolabel generation and fine-tuning, and this stepwise optimization process can introduce additional complexity and instability, particularly when dealing with different datasets or new tasks, which may require further adjustments and validation.

## References

Bernard, O.; Lalande, A.; Zotti, C.; Cervenansky, F.; and Yang, X. 2018. Deep Learning Techniques for Automatic

Table 1: Comparison results on ACDC dataset. Best in bold, second-best underlined.

| Method | Pub. | DSC↑ | | | | HD95↓ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | LV | MYO | RV | Mean | LV | MYO | RV |
| **LOSS** | | | | | | | | | |
| pCE | CVPR | 0.686 | 0.766 | 0.668 | 0.625 | 173.3 | 167.7 | 165.1 | 187.2 |
| RLoss | ECCV | 0.856 | 0.896 | 0.817 | 0.856 | 6.9 | 7.0 | 6.9 | 7.9 |
| MLoss | TIP | 0.839 | 0.876 | 0.832 | 0.809 | 27.7 | 37.9 | 28.2 | 17.1 |
| **MIX** | | | | | | | | | |
| CycleMix | CVPR | 0.848 | 0.883 | 0.798 | 0.863 | – | – | – | – |
| TriMix | ACCV | 0.888 | 0.923 | 0.864 | <u>0.877</u> | 5.9 | 4.4 | 4.3 | 8.9 |
| **PL** | | | | | | | | | |
| S2L | MICCAI | 0.832 | 0.856 | 0.806 | 0.833 | 38.9 | 65.2 | 37.1 | 14.6 |
| USTM | PR | 0.786 | 0.785 | 0.756 | 0.815 | 102.2 | 139.6 | 112.2 | 54.7 |
| DMPLS | MICCAI | 0.872 | 0.913 | 0.842 | 0.861 | 9.9 | 12.1 | 9.7 | 7.9 |
| SRPA | BIBM | 0.884 | <u>0.926</u> | 0.851 | 0.874 | 4.1 | – | – | – |
| SC-Net | MICCAI | 0.872 | 0.915 | 0.839 | 0.862 | 6.5 | 8.1 | 6.7 | 4.6 |
| ScribbleVC | MM | 0.844 | 0.871 | 0.817 | 0.843 | 6.9 | 7.0 | 7.9 | 6.0 |
| ScribFormer | TMI | 0.861 | 0.904 | 0.851 | 0.830 | <u>1.8</u> | <u>1.8</u> | 2.4 | **1.3** |
| BayesianWSS | MICCAI | 0.875 | – | – | – | 7.0 | – | – | – |
| SPNet | ISBI | 0.886 | 0.923 | **0.888** | 0.846 | – | – | – | – |
| CIM | TIP | 0.879 | 0.923 | 0.855 | 0.860 | 7.3 | 6.6 | 6.6 | 8.9 |
| QMaxViT-Unet+ | CBM | <u>0.891</u> | 0.924 | 0.864 | 0.884 | **1.3** | **1.2** | **1.1** | **1.7** |
| SAM-SP (Ours) | – | **0.899** | **0.931** | <u>0.879</u> | **0.887** | 2.2 | 2.6 | <u>1.6</u> | 2.3 |
| **MASKS** | | | | | | | | | |
| U-Net | MICCAI | 0.898 | 0.933 | 0.883 | 0.882 | 7.0 | 8.1 | 5.9 | 6.9 |

Table 2: Comparison results on MSCMR dataset. Best in bold, second-best underlined.

| Method | Pub. | DSC↑ | | | | HD95↓ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | LV | MYO | RV | Mean | LV | MYO | RV |
| **LOSS** | | | | | | | | | |
| pCE | CVPR | 0.515 | 0.597 | 0.373 | 0.574 | 219.9 | 216.7 | 217.7 | 225.4 |
| RLoss | ECCV | 0.836 | 0.903 | 0.800 | 0.807 | 8.7 | 11.3 | 6.4 | 8.3 |
| MLoss | TIP | 0.812 | 0.833 | 0.789 | 0.814 | 54.3 | 45.3 | 65.8 | 51.8 |
| **MIX** | | | | | | | | | |
| CycleMix | CVPR | 0.800 | 0.870 | 0.739 | 0.791 | – | – | – | – |
| TriMix | ACCV | 0.874 | 0.922 | 0.836 | 0.865 | – | – | – | – |
| **PL** | | | | | | | | | |
| S2L | MICCAI | 0.766 | 0.724 | 0.731 | 0.842 | 188.1 | 194.7 | 161.6 | 208.1 |
| USTM | PR | 0.662 | 0.711 | 0.547 | 0.728 | 203.7 | 198.5 | 204.3 | 208.1 |
| DMPLS | MICCAI | 0.796 | 0.881 | 0.644 | 0.863 | – | – | – | – |
| SRPA | BIBM | 0.874 | 0.912 | 0.827 | 0.883 | 12.1 | – | – | – |
| ScribbleVC | MM | 0.868 | 0.921 | 0.830 | 0.852 | – | – | – | – |
| ScribFormer | TMI | 0.839 | 0.896 | 0.813 | 0.807 | – | – | – | – |
| SPNet | ISBI | 0.883 | **0.928** | **0.851** | 0.872 | – | – | – | – |
| CIM | TIP | 0.853 | 0.905 | 0.809 | 0.845 | 41.4 | 45.2 | 42.1 | 33.8 |
| QMaxViT-Unet+ | CBM | <u>0.884</u> | 0.923 | <u>0.842</u> | <u>0.888</u> | <u>2.2</u> | <u>2.3</u> | <u>2.2</u> | <u>2.1</u> |
| SAM-SP (Ours) | – | **0.890** | <u>0.927</u> | 0.837 | **0.905** | **1.9** | **2.0** | **1.8** | **1.8** |
| **MASKS** | | | | | | | | | |
| U-Net | MICCAI | 0.770 | 0.850 | 0.721 | 0.738 | – | – | – | – |

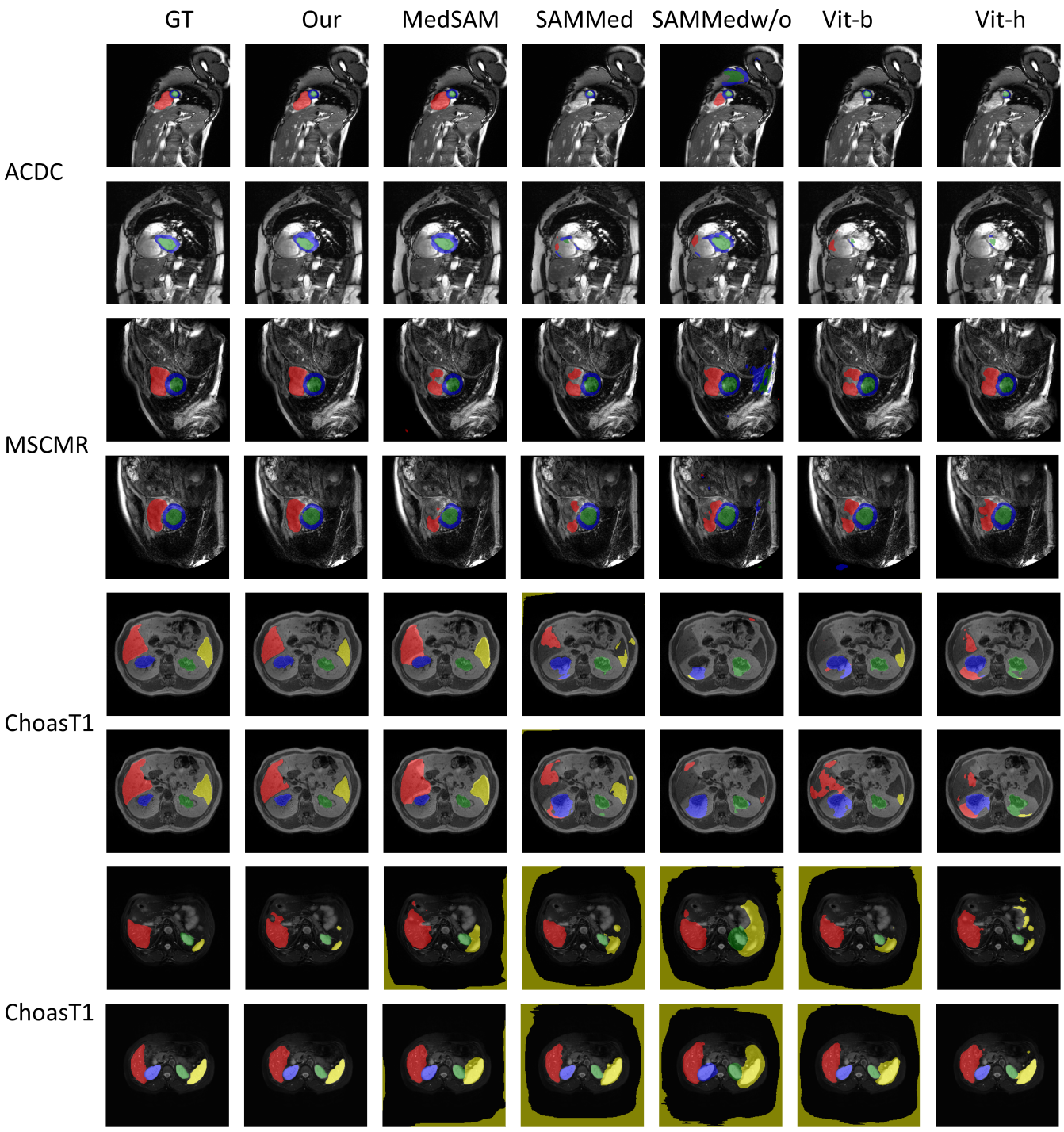


Figure 1: Our method and SAM-based methods yield the following results across each dataset.

Table 3: Comparison with SAM-based methods across all datasets.

| Method | Pub. | Model | ACDC | | | MSCMR | | | CHAOS-T1 | | | CHAOS-T2 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | MDSC | MHD95 | PLDSC | MDSC | MHD95 | PLDSC | MDSC | MHD95 | PLDSC | MDSC | MHD95 | PLDSC |
| **Fine-tuned SAM + DMPLS** | | | | | | | | | | | | | | |
| SAM-B | ICCV23 | ViT-b | 0.872 | 3.8 | 0.811 | 0.866 | 9.5 | 0.763 | 0.781 | 28.1 | 0.879 | 0.652 | 44.1 | 0.884 |
| SAM-H | ICCV23 | ViT-h | 0.873 | 4.3 | 0.832 | 0.870 | 2.1 | 0.765 | 0.805 | 23.7 | 0.906 | 0.655 | 39.4 | 0.910 |
| SAMMed w/o | ArXiv23 | ViT-b | 0.795 | 14.5 | 0.394 | 0.737 | 67.1 | 0.822 | 0.689 | 38.4 | 0.359 | 0.569 | 49.9 | 0.807 |
| SAMMed | ArXiv23 | ViT-b | 0.829 | 5.6 | 0.627 | 0.847 | 2.83 | 0.530 | 0.764 | 38.7 | 0.803 | 0.631 | 42.6 | 0.761 |
| MedSAM | Nature23 | ViT-b | 0.829 | 5.4 | 0.717 | 0.824 | 17.9 | 0.735 | 0.792 | 28.3 | 0.902 | 0.651 | 43.8 | 0.908 |
| **SAM-based** | | | | | | | | | | | | | | |
| SparseMamba-PCL | ArXiv25 | ViT-b | 0.891 | 5.5 | – | 0.821 | 41.8 | – | 0.738 | 16.8 | – | – | – | – |
| SAM-SP* | – | ViT-b | 0.895 | 2.9 | 0.890 | 0.884 | 2.0 | 0.847 | **0.844** | **8.7** | 0.919 | **0.793** | **2.7** | 0.913 |
| SAM-SP | – | ViT-h | **0.899** | **2.2** | **0.901** | **0.890** | **1.9** | **0.906** | 0.819 | 15.6 | 0.915 | 0.785 | 7.8 | 0.909 |

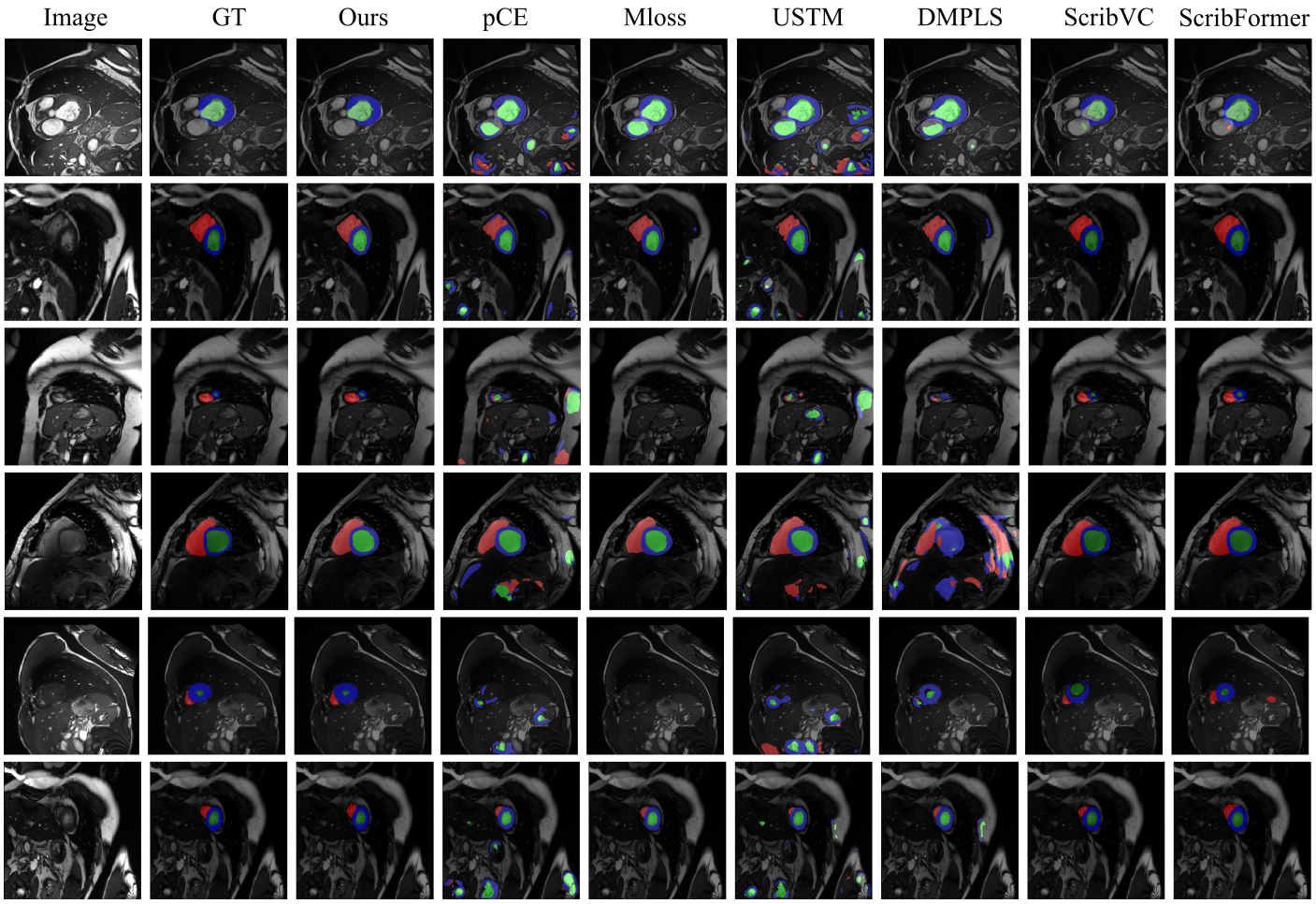


Figure 2: Our method and other scribble-supervised methods yield the following results.

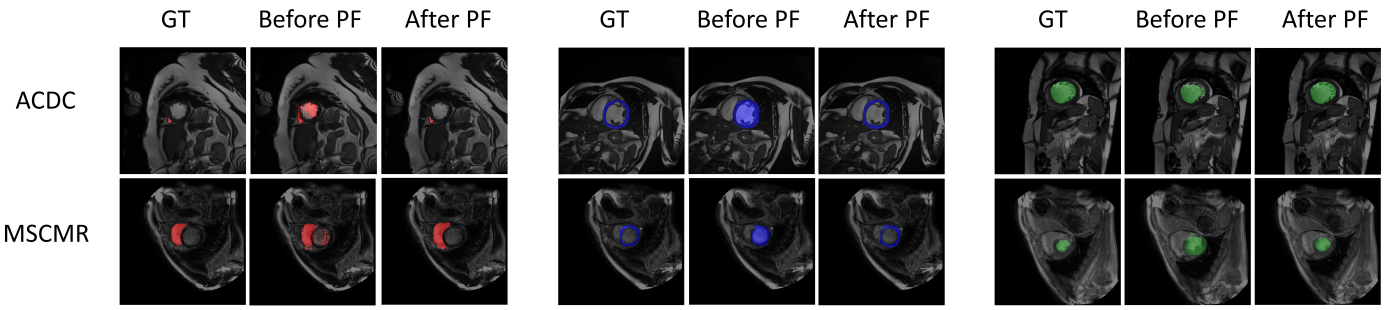


Figure 3: Comparison of pseudo labels before and after iteration

MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? *IEEE Transactions on Medical Imaging*, 37(11): 2514–2525.

Cheng, J.; Ye, J.; Deng, Z.; Chen, J.; Li, T.; Wang, H.; Su, Y.; Huang, Z.; Chen, J.; Sun, L. J. H.; He, J.; Zhang, S.; Zhu, M.; and Qiao, Y. 2023. SAM-Med2D. arXiv:2308.16184.

Kavur, A. E.; Gezer, N. S.; Barış, M.; Aslan, S.; Conze, P.-H.; Groza, V.; Pham, D. D.; Chatterjee, S.; Ernst, P.; Özkan, S.; Baydar, B.; Lachinov, D.; Han, S.; Pauli, J.; Isensee, F.; Perkonigg, M.; Sathish, R.; Rajan, R.; Sheet, D.; Dovletov, G.; Speck, O.; Nürnberger, A.; Maier-Hein, K. H.; Bozdağı Akar, G.; Ünal, G.; Dicle, O.; and Selver, M. A. 2021. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69: 101950.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. B. 2023. Segment Anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3992–4003.

Li, Z.; Zheng, Y.; Luo, X.; Shan, D.; and Hong, Q. 2023. ScribbleVC: Scribble-supervised Medical Image Segmentation with Vision-Class Embedding. *Proceedings of the 31st ACM International Conference on Multimedia*.

Luo, X.; Hu, M.; Liao, W.; Zhai, S.; Song, T.; Wang, G.; and Zhang, S. 2022. Scribble-Supervised Medical Image Segmentation via Dual-Branch Network and Dynamically Mixed Pseudo Labels Supervision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention(MICCAI)*.

Ma, J.; He, Y.; Li, F.; Han, L.-J.; You, C.; and Wang, B. 2023. Segment anything in medical images. *Nature Communications*, 15.

Yang, Z.; Lin, D.; Ni, D.; and Wang, Y. 2024. Non-iterative scribble-supervised learning with pacing pseudo-masks for medical image segmentation. *Expert Systems with Applications*, 238: 122024.

Zhang, K.; and Zhuang, X. 2022. CycleMix: A Holistic Strategy for Medical Image Segmentation from Scribble Supervision. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11646–11655.

Zhuang, X. 2018. Multivariate Mixture Model for Cardiac Segmentation from Multi-Sequence MRI. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

Zhuang, X. 2019. Multivariate Mixture Model for Myocardial Segmentation Combining Multi-Source Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12): 2933–2946.

## References

Bernard, O.; Lalande, A.; Zotti, C.; Cervenansky, F.; and Yang, X. 2018. 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.

Cheng, J.; Ye, J.; Deng, Z.; Chen, J.; Li, T.; Wang, H.; Su, Y.; Huang, Z.; Chen, J.; Sun, L. J. H.; He, J.; Zhang, S.; Zhu, M.; and Qiao, Y. 2023. SAM-Med2D. arXiv:2308.16184.

Kavur, A. E.; Gezer, N. S.; Barış, M.; Aslan, S.; Conze, P.-H.; Groza, V.; Pham, D. D.; Chatterjee, S.; Ernst, P.; Özkan, S.; Baydar, B.; Lachinov, D.; Han, S.; Pauli, J.; Isensee, F.; Perkonigg, M.; Sathish, R.; Rajan, R.; Sheet, D.; Dovletov, G.; Speck, O.; Nürnberger, A.; Maier-Hein, K. H.; Bozdağı Akar, G.; Ünal, G.; Dicle, O.; and Selver, M. A. 2021. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69: 101950.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. B. 2023. Segment Anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3992–4003.

Li, Z.; Zheng, Y.; Luo, X.; Shan, D.; and Hong, Q. 2023. ScribbleVC: Scribble-supervised Medical Image Segmentation with Vision-Class Embedding. *Proceedings of the 31st ACM International Conference on Multimedia*.

Luo, X.; Hu, M.; Liao, W.; Zhai, S.; Song, T.; Wang, G.; and Zhang, S. 2022. Scribble-Supervised Medical Image Segmentation via Dual-Branch Network and Dynamically Mixed Pseudo Labels Supervision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention(MICCAI)*.

Ma, J.; He, Y.; Li, F.; Han, L.-J.; You, C.; and Wang, B. 2023. Segment anything in medical images. *Nature Communications*, 15.

Yang, Z.; Lin, D.; Ni, D.; and Wang, Y. 2024. Non-iterative scribble-supervised learning with pacing pseudo-masks for medical image segmentation. *Expert Systems with Applications*, 238: 122024.

Zhang, K.; and Zhuang, X. 2022. CycleMix: A Holistic Strategy for Medical Image Segmentation from Scribble Supervision. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11646–11655.

Zhuang, X. 2018. Multivariate Mixture Model for Cardiac Segmentation from Multi-Sequence MRI. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

Zhuang, X. 2019. Multivariate Mixture Model for Myocardial Segmentation Combining Multi-Source Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12): 2933–2946.