Uncertainty-Aware Leave-One-Out Collaborative Learning for Scribble-Supervised Medical Image Segmentation

Zhaoyang Chen*, Tuchao Li
College of Artificial Intelligence and Big Data, Guangzhou Vocational University of Science and Technology, Guangzhou, Guangdong, China
**Corresponding Author*

Abstract: Dense pixel-level annotation is expensive in medical image segmentation, while scribble annotation provides a practical weak-supervision alternative. This study addresses pseudo-label self-confirmation in tri-branch scribble-supervised medical image segmentation. We propose an uncertainty-aware leave-one-out (LOO) collaborative learning framework built on heterogeneous CNN, Swin-UNet, and Mamba-UNet branches. Each branch is anchored by partial cross entropy on scribble pixels and receives dense pseudo supervision generated only from the other two branches; entropy-based confidence further weights pseudo labels during training and branch fusion during inference. Five-fold experiments on the ACDC cardiac MRI dataset show that the method preserves the strong baseline overlap performance, with Dice of 0.8794 compared with 0.8803 and IoU of 0.7921 compared with 0.7934. Precision changes slightly from 0.8662 to 0.8666, although HD95 and ASD do not consistently improve. These results demonstrate that the proposed strategy reduces direct self-confirmation and offers a lightweight pseudo-label governance mechanism without redesigning the backbone. The main contribution is a simple collaborative supervision formulation that separates sparse reliable scribble anchors from dense noisy pseudo-label expansion, providing a reproducible basis for future uncertainty- and boundary-aware weakly supervised medical segmentation.

Keywords: Medical Image Segmentation; Weak Supervision; Pseudo Label; Collaborative Learning; Vision Mamba;

1. Introduction

In cardiac MRI, dense training masks require clinicians to trace several anatomical structures over successive slices. The resulting contours are valuable for quantitative analysis, but producing them case by case is costly. Scribble annotation changes this workflow: a few class-specific strokes identify representative target regions without requiring a complete contour. The saved annotation effort comes with a clear consequence. However, most pixels presented during training have no human label.

The resulting learning problem is not simply one of missing labels. It is a question of how to expand a small set of trusted pixels without turning model errors into new training targets. Scribble pixels are sparse but human-provided, whereas pseudo labels are dense but model-generated. Training only on scribbles leaves much of the image unused; accepting every pseudo label can amplify false regions and misplaced boundaries. This work therefore treats the control of dense pseudo supervision, rather than backbone capacity alone, as the central design problem.

Multi-branch learning provides a natural way to reduce pseudo-label noise. Different architectures encode different inductive biases. CNN-based models such as U-Net [1] and nnU-Net [2] are effective at local texture and boundary modeling. Transformer-based medical segmentation models such as Swin-UNet [3] capture longer-range contextual interactions. Visual state-space models such as VMamba [4] and Mamba-based medical segmentation networks provide another form of efficient long-range modeling. Weak-Mamba-UNet [5] further shows that CNN, ViT, and Mamba branches can be combined for scribble-supervised medical image segmentation.

Despite this progress, the pseudo-label mechanism in a tri-branch framework deserves closer analysis. If a branch contributes to the pseudo label used to train itself, the pseudo target is not purely cross-supervised. The branch may reinforce its own mistakes, which is a form

of self-confirmation. This issue is particularly important under weak supervision because the sparse scribbles cannot fully correct dense pseudo-label errors.

This paper proposes an uncertainty-aware leave-one-out tri-branch collaborative learning strategy. The method keeps the heterogeneous CNN, Swin, and Mamba branches, but changes the supervision mechanism. For each target branch, the dense pseudo label is generated only by the other two branches. The target branch is excluded from its own teacher set. In addition, entropy-based confidence weighting is used to suppress uncertain pseudo-label regions. The resulting method is simple, easy to implement, and compatible with existing tri-branch scribble-supervised frameworks.

2. Related Work

2.1 Medical Image Segmentation Backbones

U-Net [1] established the encoder-decoder and skip-connection pattern that remains common in medical segmentation. Its direct transfer of high-resolution encoder features helps recover spatial detail during decoding. nnU-Net [2] showed that strong results also depend on adapting preprocessing and training configurations to the dataset, rather than on architecture alone. These observations motivate the CNN branch used in this study: it supplies a locally biased prediction source with explicit multi-scale feature reuse.

The other two methods adopt context differently. Swin-UNet [3] adapts the Swin Transformer into a U-shaped medical segmentation architecture and uses shifted-window self-attention for local-global feature learning. VMamba [4] adapts Mamba-style state-space modeling to vision through visual state-space blocks and 2D selective scanning. Weak-Mamba-UNet [5] combines CNN, ViT, and Mamba predictors for scribble-supervised segmentation. For our purposes, the relevance of these three models goes beyond simply offering off-the-shelf backbone choices. They produce three different views of the same sparsely annotated image, which makes it possible to ask a more specific question: how should one branch be supervised by the other two without using its own prediction as part of the target.

2.2 Scribble-Supervised Medical Image Segmentation

Weak supervision reduces annotation cost by training segmentation models from sparse labels such as points, boxes, or scribbles. Scribbles are attractive because they can mark object interiors with less effort than full masks. However, scribbles only supervise a small subset of pixels. Recent methods therefore rely on auxiliary constraints, consistency learning, or pseudo-label expansion. DMSPS [6] uses dynamically mixed soft pseudo labels for scribble-supervised medical image segmentation. ScribbleVS [7] proposes dynamic competitive pseudo-label selection and regional pseudo-label diffusion. These methods reflect a shared challenge: pseudo labels are necessary to expand supervision, but they must be controlled because noisy pseudo labels can harm training.

2.3 Collaborative and Multi-Teacher Learning

Collaborative learning trains multiple models or branches jointly so that they can provide supervision to each other. In scribble-supervised segmentation, this is useful because different branches may make different errors. However, collaboration is not automatically reliable. If the target branch participates in producing its own pseudo label, the supervision becomes partially self-generated. This paper addresses this issue through leave-one-out pseudo-label construction. Instead of using a single mixed pseudo label for all branches, each branch receives a branch-specific pseudo label generated by the remaining two branches.

3. Method

3.1 Problem Definition

Let x be a 2D medical image and y^s be its scribble label map. Only a subset of pixels is annotated. The unlabeled pixels are ignored when computing the sparse supervised loss. The task is to train a segmentation model that predicts a dense multi-class mask y from sparse scribble supervision.

The proposed framework contains three segmentation branches:

$$f_S(x), \quad f_M(x), \quad f_C(x) \tag{1}$$

where f_S denotes the Swin-UNet branch, f_M denotes the Mamba-UNet branch, and f_C denotes the CNN branch. Their softmax predictions are denoted as:

$$p_S, \quad p_M, \quad p_C \in [0,1]^{C\times H\times W} \tag{2}$$

3.2 Partial Cross Entropy on Scribble Pixels

Scribble labels are reliable but sparse. Therefore, each branch is first trained using partial cross entropy on annotated pixels only. Let Ω_s be the set of scribble-labeled pixels. For branch b, the scribble loss is:

$$\mathcal{L}_{pCE}^{b} = -\frac{1}{|\Omega_s|}\sum_{i\in\Omega_s} \log p_b(i, y_i^s) \quad (3)$$

This term anchors the training process to trustworthy human-provided labels and prevents pseudo labels from being the only supervision source.

3.3 Leave-One-Out Pseudo Label Generation

The baseline tri-branch framework fuses the predictions of all branches to obtain a pseudo label. Although this produces dense supervision, the target branch contributes to the pseudo label used to train itself. This may create self-confirmation.

To reduce this effect, the proposed method uses leave-one-out pseudo label generation. For each target branch, the teacher set excludes the target branch:

$$T_S = \{M, C\}, \quad T_M = \{S, C\}, \quad T_C = \{S, M\} \quad (4)$$

For target branch b, the teacher probability is obtained by confidence-weighted fusion of the predictions from T_b:

$$\bar{p}_b(i, k) = \frac{\sum_{t\in T_b} w_t(i) p_t(i,k)}{\sum_{t\in T_b} w_t(i) + \epsilon} \quad (5)$$

where $w_t(i)$ is the confidence weight of teacher branch t at pixel i. The branch-specific pseudo label is then:

$$\hat{y}_b(i) = \arg\max_k \bar{p}_b(i, k) \quad (6)$$

Thus, Swin is supervised by Mamba and CNN, Mamba is supervised by Swin and CNN, and CNN is supervised by Swin and Mamba. The pseudo label is dense, but it is no longer generated by the target branch itself.

3.4 Entropy-Based Confidence Weighting

Predictions with high entropy are less reliable. We compute a normalized entropy confidence map for each teacher branch:

$$c_t(i) = 1 - \frac{-\sum_{k=1}^{C} p_t(i,k) \log (p_t(i,k)+\epsilon))}{\log C} \quad (7)$$

The confidence can be sharpened using a power factor q:

$$w_t(i) = c_t(i)^q \quad (8)$$

In the main experiment, a mild confidence power is used to avoid over-suppressing uncertain pixels. The fused pseudo label is trained with a confidence-weighted Dice-style pseudo loss:

$$\mathcal{L}_{pseudo}^{b} = Dice(p_b, \hat{y}_b; w_b) \quad (9)$$

The pseudo loss is multiplied by a ramp-up factor $\lambda(t)$, so that early training relies more on scribble pixels and later training uses dense pseudo supervision more strongly:

$$\mathcal{L} = \sum_{b\in\{S,M,C\}} \left(\mathcal{L}_{pCE}^{b} + \lambda(t)\mathcal{L}_{pseudo}^{b}\right) \quad (10)$$

3.5 Uncertainty-Aware Ensemble Inference

At inference, the three branches can be combined by simple averaging or by uncertainty-aware weighting. The uncertainty-aware ensemble uses each branch confidence to weight its probability map:

$$p_{ens}(i, k) = \frac{\sum_b c_b(i) p_b(i,k)}{\sum_b c_b(i) + \epsilon} \quad (11)$$

The final label is obtained by:

$$y_{ens}(i) = \arg\max_k p_{ens}(i, k) \quad (12)$$

This inference strategy gives more influence to confident branches at each pixel and reduces the effect of uncertain predictions.

4. Experiments

4.1 Dataset

We evaluate the method on the ACDC cardiac cine-MRI dataset [8]. Each case contains annotated cardiac structures across a sequence of two-dimensional slices. During training, the optimizer receives the sparse scribble map. Pixels outside the scribbles are ignored by the partial cross-entropy term. Dense masks are retained for validation and testing. The same fold assignments and annotation protocol are used for the baseline and the proposed method, so the comparison isolates the change in collaborative supervision. Results are aggregated over five folds.

4.2 Implementation Details

The experiment contains three branches: Swin-UNet, Mamba-UNet, and a CNN-based segmentation branch. The Swin and Mamba branches are initialized with available pretrained weights. Input slices are extracted as 224 × 224 patches. This model uses stochastic gradient descent with a base learning rate of 0.01, a batch size of 24, and polynomial learning-rate decay, training each main run for 30,000 iterations.

In the proposed configuration, the target branch is excluded from pseudo-label construction. The loss uses a ramp-up setting of 5 for pseudo-label and an entropy-confidence power of 0.5. The

disagreement coefficient is set to 1.0, meaning teacher disagreement does not contribute an additional attenuation term in the LOO experiment. Stronger reliability weighting and self-mixing proved less stable in early tests, so they were excluded from the main comparison.

Training length, data folds, preprocessing, and evaluation code are kept identical across the baseline and LOO runs, with only the pseudo-label and fusion rules modified. All tables report the mean and standard deviation computed over five folds. The main evaluation follows this fixed protocol, whereas exploratory checkpoint comparisons serve a purely diagnostic purpose and are not presented as separate methods. This constrained design keeps the analysis centered on the supervision strategy rather than on checkpoint selection.

4.3 Evaluation Metrics

We report Dice similarity coefficient, intersection over union (IoU), 95th percentile Hausdorff distance (HD95), average surface distance (ASD), recall, precision, and pixel accuracy. Dice and IoU measure region overlap, whereas HD95 and ASD reflect boundary and surface consistency. In weakly supervised segmentation, distance metrics are especially important because pseudo-label errors may appear as boundary irregularities even when Dice remains similar.

5. Results

5.1 Overall Five-Fold Performance

Table 1. Boundary Distance Metrics on ACDC

Method	HD95	ASD
ensemble	1.8949 ± 0.5852	0.5469 ± 0.2025
LOO	2.0937 ± 0.5779	0.6134 ± 0.1822

Table 2. Overlap and Classification on ACDC

Method	Dice	IoU	Recall	Precision	Pixel Acc
ensemble	0.8803 ± 0.0172	0.7934 ± 0.0255	0.9102 ± 0.0190	0.8662 ± 0.0177	0.9938 ± 0.0013
LOO	0.8794 ± 0.0180	0.7921 ± 0.0264	0.9080 ± 0.0194	0.8666 ± 0.0173	0.9938 ± 0.0014

Table 1 reports the five-fold boundary distance metrics on ACDC. The five folds are used here as a stability statistic rather than as the main object of analysis. Lower values are better for HD95 and ASD, while higher values are better for the remaining metrics.

Table 2 reports the five-fold Overlap and classification on ACDC. The proposed LOO strategy maintains nearly the same region-overlap performance as the baseline. Dice decreases by 0.0009 and IoU decreases by 0.0013, while precision increases slightly by 0.0003. These differences are small relative to the five-fold standard deviations.

HD95 and ASD are worse for LOO, because suppressing self-confirmation at the pseudo-label level does not mean improve the boundary quality. The result supports a restrained interpretation: LOO changes the supervision mechanism and keeps the baseline in an accuracy level, but it does not mean provide a consistent overall performance gain.

5.2 Diagnostic Observations and Required Ablations

Figure 1 shows the dispersion for five-fold and is not intended to rank individual folds. The Dice and IoU changes are only -0.0009 and -0.0013, which places the LOO model close to the baseline in region overlap. HD95 and ASD increase by 0.1988 and 0.0665. In short, forgoing direct self-mixing preserves the overall foreground quality, but it does not address the boundary errors that drive surface-distance metrics.

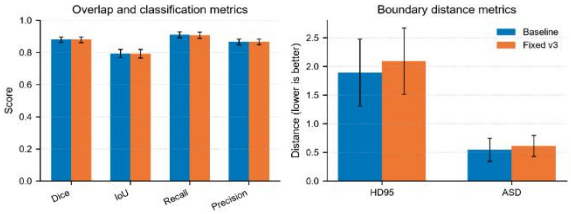

Figure 1. Overall Five-Fold Mean and Standard Deviation.

In earlier trials, we tested stronger reliability weighting and a few boundary-guidance strategies, but none of these variants led to stable improvements in the recorded results. For now, these findings should only be taken as preliminary diagnostics, not as conclusive ablation evidence.

5.3. Discussion

The five-fold comparison is not meant to show that LOO improves accuracy over the baseline. In fact, its mean scores are not higher. What LOO does is remove the direct self-supervision path, while keeping the overlap accuracy nearly unchanged. The contribution is therefore located in the training relation among the three branches, not in a new encoder or decoder. This distinction matters because it defines what the current evidence can, and cannot, establish.

With a single pseudo label fused from all three

predictions, the output of a target branch appears on both sides of its pseudo-supervised update. It helps create the target and is then optimized against that target. Leave-one-out fusion removes this direct path. For example, the CNN branch is trained against a dense target formed from the Swin and Mamba predictions only. The two teachers are not fully independent because all branches share the same images, scribbles, and training schedule, but the target no longer contains the student's current output.

Entropy is useful here, but its meaning is limited. When the two teachers produce a diffuse class distribution, their contribution to the pseudo loss is reduced. This rule is computationally light and requires no extra network. It should not, however, be interpreted as an estimate of correctness. A model may assign low entropy to a wrong class, particularly at a small structure or an ambiguous contour.

Distance metrics reveal this limitation more clearly than Dice. HD95 is sensitive to a small number of remote boundary errors, and ASD captures systematic contour shifts. simply down-weighting high-entropy pixels does not reliably remove either kind of error. A more robust confidence model would therefore need to go beyond predictive entropy alone—for example, by incorporating inter-branch disagreement, class-wise calibration, or a boundary-specific reliability term. This view is in line with scribble-supervised approaches that use explicit label propagation and consistency constraints to handle unlabeled regions [9,10].

In the current setup, branch collaboration ends at the output maps. The branches exchange probability maps only during pseudo-label generation and are fused again at inference; their intermediate features never interact. While this keeps the modification minimal and makes its effect easier to isolate, it also leaves potentially useful complementary information unused. Extensions could try distilling CNN boundary responses into the Mamba decoder, or aligning Transformer semantic features with Mamba features, provided that any such feature-level interaction is evaluated independently from the LOO supervision.

Checkpoint selection can also hide boundary variation. A checkpoint chosen solely by Dice can produce a mask with good overlap but poor surface geometry. In future experiments, the selection rule should be decided before testing and computed on the validation set using a score that combines Dice with HD95 (or ASD). Reporting that rule alongside the final checkpoint would improve the reproducibility of boundary comparisons and avoid letting test-set observations influence model selection.

6. Conclusion

We started from a concrete problem in tri-branch pseudo-supervision: a branch can inadvertently shape the very target used to supervise itself. Our solution replaces the shared target with three leave-one-out targets, which cuts that direct feedback loop without altering the underlying CNN, Transformer, or Mamba backbones. This makes the two supervision sources explicit—scribbles provide sparse anchors, while the other branches generate dense guidance for the remaining pixels.

The results also highlight what this setup does not yet solve. Dice and IoU remain nearly unchanged, showing that teacher exclusion preserves global overlap. Yet the higher HD95 and ASD reveal that entropy-based weighting alone cannot protect fine anatomical boundaries. Therefore, the path forward is not to add yet another branch, but to make collaboration more intelligent: better confidence calibration, explicit modeling of inter-branch disagreement, boundary-aware regularization, and a checkpoint selection criterion that jointly considers overlap and surface distance. In a broader sense, the key challenge for weakly supervised segmentation is no longer how to obtain sparse human labels, but how to reliably govern the dense machine-generated supervision that follows.

Acknowledgments

This work was supported by the 2026 University Research Project of Guangzhou Vocational University of Science and Technology under Grant No. 2026LG29 (Research on Improved Multimodal Semantic Fusion Doodle-Based Medical Image Segmentation).

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015, LNCS 9351, pp. 234-241, 2015.

[2] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-

based biomedical image segmentation," Nature Methods, vol. 18, no. 2, pp. 203-211, 2021.

[3] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation," arXiv:2105.05537, 2021.

[4] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu, "VMamba: Visual State Space Model," arXiv:2401.10166, 2024.

[5] Z. Wang and C. Ma, "Weak-Mamba-UNet: Visual Mamba Makes CNN and ViT Work Better for Scribble-based Medical Image Segmentation," arXiv:2402.10887, 2024.

[6] M. Han, X. Luo, and X. Xie, "DMSPS: Dynamically mixed soft pseudo-label supervision for scribble-supervised medical image segmentation," Medical Image Analysis, vol. 97, Art. no. 103274, 2024.

[7] T. Wang, X. Zhang, Y. Chen, Y. Zhou, L. Zhao, T. Tan, and T. Tong, "ScribbleVS: Scribble-Supervised Medical Image Segmentation via Dynamic Competitive Pseudo Label Selection," arXiv:2411.10237, 2024.

[8] O. Bernard, A. Lalande, C. Zotti et al., "Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?" IEEE Transactions on Medical Imaging, vol. 37, no. 11, pp. 2514-2525, 2018.

[9] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 3159-3167.

[10] K. Zhang and X. Zhuang, "CycleMix: A Holistic Strategy for Medical Image Segmentation from Scribble Supervision," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), New Orleans, LA, USA, 2022, pp. 11656-11665

