SEMI-CERVIXSEG: A MULTI-STAGE TRAINING STRATEGY FOR SEMI-SUPERVISED CERVICAL SEGMENTATION

Juntao Jiang[1] *Yali Bi*[2] *Chunlin Zhou*[1*] *Yong Liu*[1,3*] *Jiangning Zhang*[1,4]

[1]College of Control Science and Engineering, Zhejiang University
[2]College of Computer and Information Science, Southwest University
[3]State Key Laboratory of Industrial Control Technology, Zhejiang University [4]Youtu Lab, Tencent

2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI) | 979-8-3315-2052-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/ISBI60581.2025.10981295

ABSTRACT

Image segmentation plays a critical role in computer-aided diagnosis and treatment planning for cervical cancer. Obtaining a large number of labeled images for supervised cervical segmentation is often labor-intensive and time-consuming. In this paper, we propose a multi-stage semi-supervised learning framework (*Semi-CervixSeg*) to address the cervical segmentation task in ultrasound images for *Fetal Ultrasound Grand Challenge: Semi-Supervised Cervical Segmentation* in ISBI 2025. Specifically, in the initial stage, we utilize unlabeled data through a multi-view random augmentation strategy, using consistency constraints and a contrastive learning method. Subsequently, a progressive multi-stage training strategy is adopted to generate and optimize pseudo-labels, further improving segmentation results. Experimental results demonstrate that the proposed method significantly improves segmentation performance compared with supervised methods. As a technical report for the challenge, this paper elaborates on our methodology and experimental findings in detail. The code can be accessed at `https://github.com/juntaoJianggavin/Semi-CervixSeg`.

***Index Terms*—** Semi-CervixSeg, cervical segmentation, semi-supervised learning, multi-stage training, contrastive learning, pseudo-label, FUGC ISBI 2025

1. INTRODUCTION

Cervical cancer is one of the most common health risks among women, and early diagnosis is essential to reduce mortality rates. Transvaginal ultrasound (TVUS) is the preferred method for observing the cervix as it provides detailed anatomical structures. Accurate segmentation of cervical muscle in ultrasound images is crucial for analyzing deep muscle structures, evaluating their functionality, and developing personalized treatment plans.

In recent years, the advent of deep-learning techniques has brought about revolutionary changes to computer-aided diagnosis. In the context of ultrasound image segmentation, several optimization methods have been proposed to further enhance its accuracy and efficiency, leveraging the power of deep-learning algorithms to achieve more reliable results. For example, many research efforts focus on fetal head segmentation in ultrasound images, exploring the use of artificial intelligence methods to achieve accurate positioning and thus provide better clinical support [1, 2, 3, 4, 5, 6, 7, 8, 6, 9, 10].

It is worth noting that existing supervised learning models typically rely on large amounts of high-quality annotated data. However, annotating medical images is time-consuming and expensive and requires domain expertise. To address this issue, semi-supervised learning (SSL) has gradually become an important research direction in the field of medical imaging. By combining a small amount of annotated data with a large amount of unannotated data, SSL can improve model performance while reducing annotation costs.

Semi-supervised segmentation methods focus on two prominent categories: *consistency regularization* and *pseudo-label generation*. Consistency-based methods require the model to produce consistent outputs for different augmented versions of the same input. For instance, contrastive learning-based methods [11, 12, 13] leverage unannotated data by imposing consistency constraints on predictions for differently augmented inputs, thereby improving model performance. Pseudo-labeling methods [14, 15, 16] employ pre-trained models to generate predicted labels for unannotated data and use them as training data for subsequent stages.

The "*Fetal Ultrasound Grand Challenge: Semi-Supervised Cervical Segmentation*" focuses on the semi-supervised segmentation of cervical muscles, aiming to propose more effective solutions and promote the development of automated cervical image segmentation systems. Building on the MICCAI PSFHS 2023 [3] and MICCAI IUGC 2024 challenges [4], this competition extends from fully supervised settings to semi-supervised settings, encouraging participating teams to explore how unlabeled data can be utilized to improve segmentation performance.

In response to this competition, we propose a multi-stage training strategy, *Semi-CervixSeg*, focusing on semi-supervised learning for cervical segmentation tasks. This

*Corresponding author

method introduces contrastive learning combined with consistency regularization in the initial stage. In the subsequent stages, pseudo-labels are generated and refined in multiple training sessions. The main contributions of this paper are as follows:

- In the initial training stage, we propose a method combining the supervised loss for labeled data and the consistency loss for unlabeled data.
- A multi-view random augmentation strategy is designed, applying various random augmentations on unannotated data for consistency constraints, improving the generalization ability of the model.
- A progressive optimization multi-stage training framework is designed. Through iterative generation and refinement of pseudo-labels in different stages, segmentation performance is gradually improved.

Experimental results show that the proposed method achieves impressive performance on the FUGC ISBI 2025 Challenge Dataset.

2. METHODS: SEMI-CERVIXSEG

2.1. Initial Stage: Contrastive Learning for the Unlabeled data

In the initial stage, we use the supervised loss for labeled data and the consistency loss for unlabeled data, and the final loss is the sum of these two losses. The contrastive learning method applies various random augmentations to unlabeled data to generate samples from different perspectives. The model then makes predictions on these augmented samples and calculates the consistency loss between different predictions to constrain the model's learning. The workflow for the initial stage can be seen in Figure 1.

2.1.1. Model Selection

In the first stage, RWKV-UNet [17] is used as the image segmentation model due to its strong capability to capture local and global information and its remarkable performance in datasets of different modalities, including breast lesions segmentation task in ultrasound images (Breast Ultrasound Images Dataset (BUSI)) [18].

2.1.2. Loss Function Design

The loss in the initial stage is a combination of the supervised loss and the consistency loss. It allows the model to learn from both labeled and unlabeled data simultaneously. The total loss function is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{supervised}} + \mathcal{L}_{\text{consistency}} \tag{1}$$

Supervised Loss: The supervised loss combines the Dice loss and the cross-entropy loss for training on labeled data. The supervised loss function can be expressed as:

$$\mathcal{L}_{\text{supervised}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{C E}} \tag{2}$$

Consistency Loss: The consistency loss is the core of this method and is used for training on unlabeled data. The core idea is that the model's predictions for different augmented versions of the same unlabeled data should be consistent. Specifically, multiple random augmentations are applied to the unlabeled data to obtain different augmented data versions, which are then input into the model to get the corresponding outputs. To ensure the comparability of different augmented outputs, inverse transformations are applied to the augmented outputs. The consistency loss is defined as the mean squared error between different augmented outputs, which can be expressed as:

$$\mathcal{L}_{\text{consistency}} = \frac{1}{M} \sum_{j=1}^{M} \left(I_1(o_{1j}) - I_2(o_{2j}) \right)^2 \tag{3}$$

where M is the number of elements in the output, and o_{1j} and o_{2j} are the j - th elements of the outputs after two different augmented data are input into the model, respectively. And I_1 and I_2 are the inverse transformations for each augmentation method.

2.2. Subsequent Stages: Pseudo-label Generation and Refinement

Pseudo-label generation uses a pre-trained model (trained in the initial stage) on labeled data to predict unlabeled data labels, creating pseudo-labels and expanding the training dataset with abundant unlabeled data. Multiple training sessions optimize pseudo-labels and improve model performance. Early pseudo-labels may be inaccurate, but with each training, the model regenerates them based on updated parameters, enhancing their quality.

2.2.1. Model Selection

In these stages, RWKV-UNet and PVT-EMCAD-B2 [19, 20] are used as image segmentation models. Due to the CUDA version requirements of RWKV-UNet, it cannot be the model submitted to the competition platform. EMCAD is an efficient multi-scale convolutional attention decoder for medical image segmentation, showing adaptability to different encoders. RWKV-UNet is used to complete the first two training sessions and PVT-EMCAD-B2 is used to complete the last three training sessions.

2.2.2. Training Process

Let the model be f, where θ represents the model parameters. For the unlabeled dataset $\mathcal{D}_u = \{x_j\}_{j=1,}^{N_u}$, the pseudo-labels $\tilde{y}_j^{(n)}$ generated during the n-th training can be expressed as:

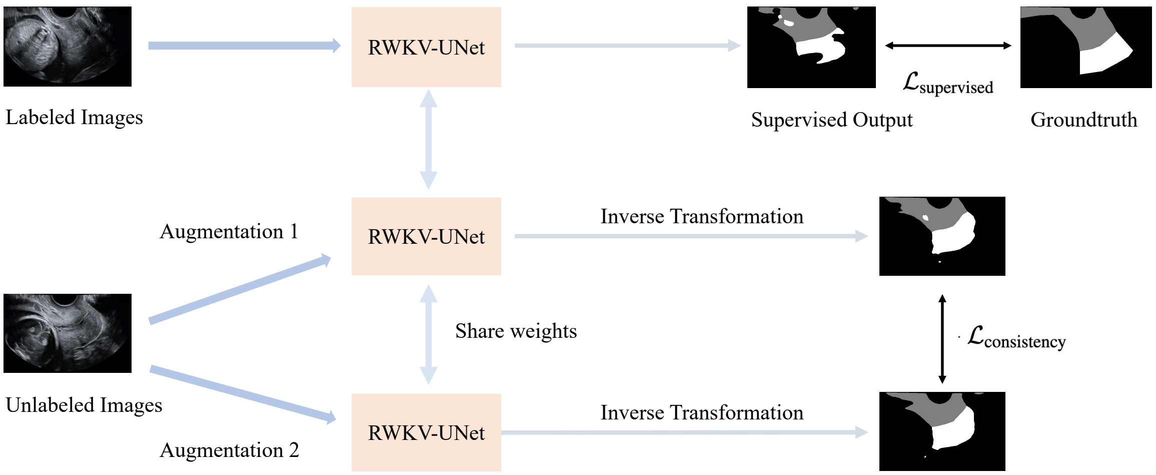

Fig. 1. A contrastive learning method is used in the initial stage. The final loss is the sum of the supervised loss for labeled data and the consistency loss for unlabeled data.

$$\tilde{y}_j^{(n)} = f_{\theta^{(n)}}\left(x_j\right), \quad j = 1, \cdots, N_u \tag{4}$$

Here, $\theta^{(n)}$ is the model parameter at the beginning of the n-th training.

Then merge the labeled dataset $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{N_l}$ and the unlabeled dataset with the pseudo-labels generated in the n-th training into a new dataset $\mathcal{D}^{(n)} = \left\{\left(x_k, \hat{y}_k^{(n)}\right)\right\}_{k=1}^{N_l+N_u}$, where:

$$\hat{y}_k^{(n)} = \begin{cases} y_k, & k = 1, \cdots, N_l \\ \tilde{y}_{k-N_l}^{(n)}, & k = N_l + 1, \cdots, N_l + N_u \end{cases} \tag{5}$$

During the n-th training, use a unified loss function $\mathcal{L}$ to calculate the loss in the merged dataset.

$$\mathcal{L}^{(n)} = \frac{1}{N_l + N_u} \sum_{k=1}^{N_l+N_u} \mathcal{L}\left(f_{\theta^{(n)}}\left(x_k\right), \hat{y}_k^{(n)}\right) \tag{6}$$

The loss function is the combination of the Dice loss and the cross-entropy loss like the supervised loss in the initial stage:

$$\mathcal{L} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{C E}} \tag{7}$$

3. EXPERIMENTS

3.1. Datasets

The dataset consists of transvaginal ultrasound images captured after bladder emptying, with patients positioned in a low Fowler's position. A curved transducer with a 2–10 MHz vaginal probe is used for imaging, and operators are instructed to minimize post-processing artifacts while adjusting parameters such as gain and frequency at their discretion. The segmentation task focuses on identifying and delineating the anterior lip and posterior lip of the cervix from transvaginal ultrasound images. In the dataset, the training set contains 50 image-mask pairs along with 450 unlabeled images. The testing set for the validation phase has 90 samples, while the dataset for the final evaluation includes 300 samples.

Table 1. Performance comparison between Semi-CervixSeg and supervised learning methods in the validation phase.↑↓ denotes the higher (lower) the better. − means the experiment is not conducted. **Bold** represents the best result.

Methods	Type	Dice(↑)	HD(↓)
R50-UNet	Supervised	0.7782	56.54
PVT-EMCAD-B2	Supervised	0.8268	48.06
Semi-CervixSeg	Semi-Supervised	**0.9117**	**41.92**

3.2. Implementation Details

All experiments are performed on PG500-216(V-100) with 32 GB of memory. The resolution of the input images is 384×384. The total training epochs are 300, and the batch size is 8. The initial learning rate is 1e-3, and the minimum learning rate is 0. Augmentation methods for labeled and unlabeled data include horizontal flipping, vertical flipping, and random 90-degree rotation multiple times. AdamW [21] optimizer and CosineAnnealingLR [22] scheduler are used. Average Dice and Hausdorff Distance (HD) are used as evaluation metrics.

Table 2. Performance comparison of different training configurations in the validation phase.

Id	Initial Stage	Stage I	Stage II	Stage III	Stage IV	Dice(↑)	HD(↓)
0	R50-UNet	-	-	-	-	0.7996	61.20
1	PVT-EMCAD-B2	-	-	-	-	0.8467	52.69
2	PVT-EMCAD-B2	R50-UNet	-	-	-	0.8469	54.76
3	PVT-EMCAD-B2	PVT-EMCAD-B2	-	-	-	0.8831	**40.89**
4	RWKV-UNet	PVT-EMCAD-B2	-	-	-	0.8930	50.93
5	RWKV-UNet	RWKV-UNet	PVT-EMCAD-B2	-	-	0.9076	51.75
6	RWKV-UNet	RWKV-UNet	PVT-EMCAD-B2	PVT-EMCAD-B2	-	0.9085	41.85
7	RWKV-UNet	RWKV-UNet	PVT-EMCAD-B2	PVT-EMCAD-B2	PVT-EMCAD-B2	**0.9117**	41.92

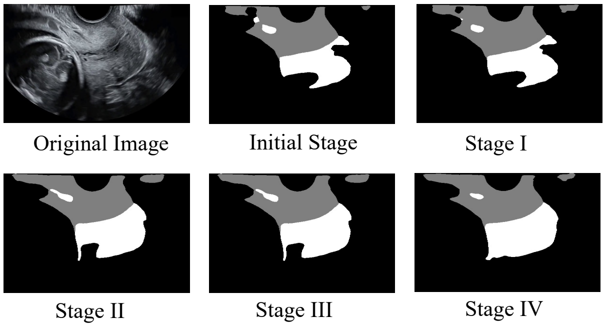

Fig. 2. Example pseudo-labels generated by models trained in the initial stage, and subsequent Stage I, Stage II, Stage III and Stage IV.

3.2.1. *Initial Stage*

We split the labeled data into a training set and a validation set with a ratio of 9:1 and train different models in the initial stage. In addition to RWKV-UNet, we also experiment with some other models, such as PVT-EMCAD-B2 and ResNet50-UNet (R50-UNet), as baselines. For experiments training PVT-EMCAD-B2, the deep supervision strategy is used, which employs multi-level feature supervision by adding losses from different upsampling stages:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4 \tag{8}$$

3.2.2. *Subsequent Stages*

We split the labeled data into a training set and a validation set with a ratio of 9:1. The training set of the labeled and unlabeled data is merged into a new training set. The best models on the validation set are used to generate pseudo-labels for the next stage or to get the final results. In addition to RWKV-UNet and PVT-EMCAD-B2, we also experiment with some other models, such as R50-UNet. For experiments training PVT-EMCAD-B2, the deep training strategy is used.

3.3. Results

Example pseudo-labels generated by models trained in the initial stage, and subsequent Stage I, Stage II, Stage III and Stage IV can be seen in Figure 2, from which we can see that with the iterative refinement process, the quality of the pseudo-labels has been significantly improved, thus helping with training segmentation models.

Table 1 compares the performance of our method and two supervised learning methods on the dataset in the validation phase. The experiments show that our method can effectively utilize unlabeled data and far outperforms the supervised learning methods. The experimental results in our framework in the validation phase are shown in Table 2.

By comparing Table 1 and 2, and analyzing the performances of R50-UNet and PVT-EMCAD-B2, with and without using consistency loss, the effectiveness of the initial stage can be demonstrated. Results in Table 2 demonstrate that the process of conducting multiple rounds of training, repeatedly generating and refining pseudo-labels, can significantly improve segmentation accuracy. Besides, analysis of Experiments 3 and 4 in Table 2 demonstrate that the RWKV-UNet architecture exhibits superior performance during the initial training phase than PVT-EMCAD-B2.

4. CONCLUSION

In this paper, we propose a multi-stage semi-supervised learning framework, *Semi-CervixSeg*, to address the challenging task of cervical segmentation in ultrasound images. By integrating contrastive learning, consistency regularization, and a multi-view random augmentation strategy, our method effectively utilizes unlabeled data. The progressive refinement of pseudo-labels through multi-stage training further improves segmentation performance. Experimental results on the FUGC ISBI 2025 Challenge Dataset demonstrate that our approach achieves remarkable performance.

Limitations and Future Work. Due to the limited amount of annotated data, the validation set consistently consists of only 5 images, making it difficult to select the optimal model. Further exploration could be carried out on how to rationally divide the dataset with a small amount of annotated data or how to utilize unannotated data for model selection. Additionally, the design of the consistency loss function has not been fully explored, and more attempts can be made in this regard in the future.

5. ACKNOWLEDGMENTS

This research is supported by the project "Research on the Core Controller of the New Generation of Mobile Robots Based on Vision-Language Models (State Key Laboratory)".

References

[1] Jieyun Bai et al., "A framework for computing angle of progression from transperineal ultrasound images for evaluating fetal head descent using a novel double branch network," *Frontiers in physiology*, 2022.

[2] Yaosheng Lu et al., "The jnu-ifm dataset for segmenting pubic symphysis-fetal head," *Data in Brief*, 2022.

[3] Jieyun Bai et al., "Pubic symphysis-fetal head segmentation from transperineal ultrasound images," 2023.

[4] Jieyun Bai et al., "Intrapartum ultrasound grand challenge 2024," Apr. 2024.

[5] Zhanhong Ou et al., "Rtseg-net: A lightweight network for real-time segmentation of fetal head and pubic symphysis from intrapartum ultrasound images," *Comput. Biol. Medicine*, 2024.

[6] Zhensen Chen et al., "Fetal head and pubic symphysis segmentation in intrapartum ultrasound image using a dual-path boundary-guided residual network," *IEEE J. Biomed. Health Informatics*, 2024.

[7] Ruiyu Qiu et al., "Psfhsp-net: an efficient lightweight network for identifying pubic symphysis-fetal head standard plane from intrapartum ultrasound images," *Medical Biol. Eng. Comput.*, 2024.

[8] Zhensen Chen, Zhanhong Ou, Yaosheng Lu, and Jieyun Bai, "Direction-guided and multi-scale feature screening for fetal head-pubic symphysis segmentation and angle of progression calculation," *Expert Syst. Appl.*, 2024.

[9] Zihao Zhou, Yaosheng Lu, Jieyun Bai, Víctor M. Campello, Fan Feng, and Karim Lekadir, "Segment anything model for fetal head-pubic symphysis segmentation in intrapartum ultrasound image analysis," *Expert Syst. Appl.*, 2025.

[10] Jieyun Bai et al., "Psfhs challenge report: Pubic symphysis and fetal head segmentation from intrapartum ultrasound images," *Medical Image Analysis*, 2025.

[11] Aaron van den Oord et al., "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[12] Ting Chen et al., "A simple framework for contrastive learning of visual representations," 2020.

[13] Xinkai Zhao et al., "Cross-level contrastive learning and consistency constraint for semi-supervised medical image segmentation," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[14] Yuexiang Li et al., "Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*. Springer, 2020, pp. 614–623.

[15] Huifeng Yao, Xiaowei Hu, and Xiaomeng Li, "Enhancing pseudo label quality for semi-supervised domain-generalized medical image segmentation," in *Proceedings of the AAAI conference on artificial intelligence*, 2022, vol. 36, pp. 3099–3107.

[16] Jiawei Su, Zhiming Luo, Sheng Lian, Dazhen Lin, and Shaozi Li, "Mutual learning with reliable pseudo label for semi-supervised medical image segmentation," *Medical Image Analysis*, vol. 94, pp. 103111, 2024.

[17] Juntao Jiang, Jiangning Zhang, Weixuan Liu, Muxuan Gao, Xiaobin Hu, Xiaoxiao Yan, Feiyue Huang, and Yong Liu, "Rwkv-unet: Improving unet with long-range cooperation for effective medical image segmentation," 2025.

[18] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy, "Dataset of breast ultrasound images," *Data in brief*, vol. 28, pp. 104863, 2020.

[19] Wenhai Wang et al., "Pvt v2: Improved baselines with pyramid vision transformer," *Computational visual media*, vol. 8, no. 3, pp. 415–424, 2022.

[20] Md Mostafijur Rahman, Mustafa Munir, and Radu Marculescu, "Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11769–11779.

[21] I Loshchilov, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[22] Ilya Loshchilov and Frank Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.