IEEE Transactions on Medical Robotics and Bionics, Available online 29 August 2025
DOI:10.1109/tmrb.2025.3604103


# S4RoboFormer: Scribble-Supervised Surgical Robotic Segmentation Transformer via Augmented Consistency Training

Ziyang Wang, Tianxiang Chen, Zi Ye, Yiyuan Ge, Zhihao Chen, Jiabao Li, Yifan Zhao *Senior Member, IEEE,*

***Abstract*—Advancements in deep learning for surgical instrument segmentation have notably improved the proficiency, safety, and efficacy of minimally invasive robotic surgeries. The effectiveness of deep learning, however, is contingent upon the availability of large datasets for training, which are often associated with substantial annotation costs. Given the dynamic nature of surgical robots, scribble-based labeling emerges as a more viable and cost-effective alternative to traditional pixel-wise dense labeling. This paper introduces the Scribble-Supervised Surgical Robotic Segmentation Transformer (S4RoboFormer), designed to mitigate the challenges posed by resource-intensive annotations. S4RoboFormer incorporates a Vision Transformer (ViT)-based U-shaped segmentation network, enhanced with a specialized Weakly-Supervised Learning (WSL) strategy that comprises consistency training through (i) data-based perturbation using a data-mixed interpolation technique, and (ii) network-based perturbation via a self-ensembling strategy. This methodology promotes uniform predictions across different levels of perturbation under conditions of limited-signal supervision. S4RoboFormer outperforms existing state-of-the-art baseline WSL frameworks with both convolutional neural network(CNN)- and ViT-based segmentation networks on a pre-processed public dataset. The code of S4RoboFormer, all baseline methods, pre-processed data, and scribble simulation algorithm are all made publicly available at https://github.com/ziyangwang007/CV-WSL-Robot.**

***Index Terms*—Surgical AI, Image Segmentation, Vision Transformer, Minimally Invasive Surgery.**

## I. Introduction

Robotic-assisted surgery has revolutionized complex surgical procedures by significantly enhancing precision, control, and adaptability [19], [1]. These advancements offer the potential for reduced patient trauma, expedited recovery, and improved surgical outcomes. A critical aspect of these robotic systems is the accurate segmentation of surgical instruments and tissues, which is essential for their optimal operation [35], [7], [14], [8].

The introduction of surgical robotic segmentation was initially marked by the Endoscopic Vision Instrument Segmentation and Tracking Challenge in 2015 [2]. Since then, researchers have enriched this field by developing a variety of datasets [1], [6], [18]. Leveraging these datasets, foundational segmentation networks such as UNet [21] have emerged, building on the Fully Convolutional Network (FCN) [16] and offering a tailored approach for biomedical image segmentation. UNet's architecture facilitates precise semantic segmentation, promoting the development of various UNet adaptations aimed at distinct segmentation challenges across diverse medical imaging domains [27], [5], [33]. These adaptations have been consistently applied to surgical instrument segmentation, yielding notable enhancements [22], [10]. Recent ViT architecture, which outperforms CNN due to modeling global dependencies, has also been explored with UNet, such as TransUNet [5], SwinUNet [4]. Despite the promising capabilities of deep learning architectures, particularly CNN and ViT, in surgical tool segmentation, their dependency on large, annotated datasets presents a significant hurdle, both in terms of cost and time. In response, Weakly-Supervised Learning (WSL) segmentation methods have emerged as a compelling alternative, reducing the reliance on extensively labeled data by utilizing either unannotated or weakly annotated data [12], [28], [34], [29], [32].

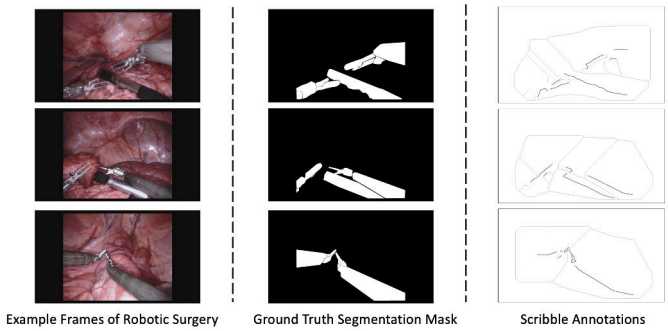


Fig. 1. Example Frames During Robotic Surgery (Left Column), Traditional Segmentation Annotations of Surgical Robots (Middle Column), and Our Proposed Scribble Annotations (Right Column). This figure illustrates the contrast between the dense, pixel-wise annotations typically used in surgical instrument segmentation and the more efficient, clinician-friendly scribble annotations, which can be utilized for S4RoboFormer training.

Figure 1 visually contrasts video frames from robotic surgeries with conventional dense segmentation annotations, and our proposed scribble annotations, which is a much more practical annotation fashion for clinicians while remaining challenging for network training. To address the

Ziyang Wang is with Department of Computer Science, University of Oxford, UK (ziyangwang@ieee.org).
Tianxiang Chen is with the School of Cyber Science and Technology, University of Science and Technology of China, China.
Ye Zi is with the Institute of Intelligent Software, China.
Yiyuan Ge, Zhihao Chen are with Computer School, Beijing Information Science & Technology University, China.
Jiabao Li is with School of Artificial Intelligence, Anhui Polytechnic University, China.
Yifan Zhao is with Centre for Life-cycle Engineering and Management, Cranfield University, UK.

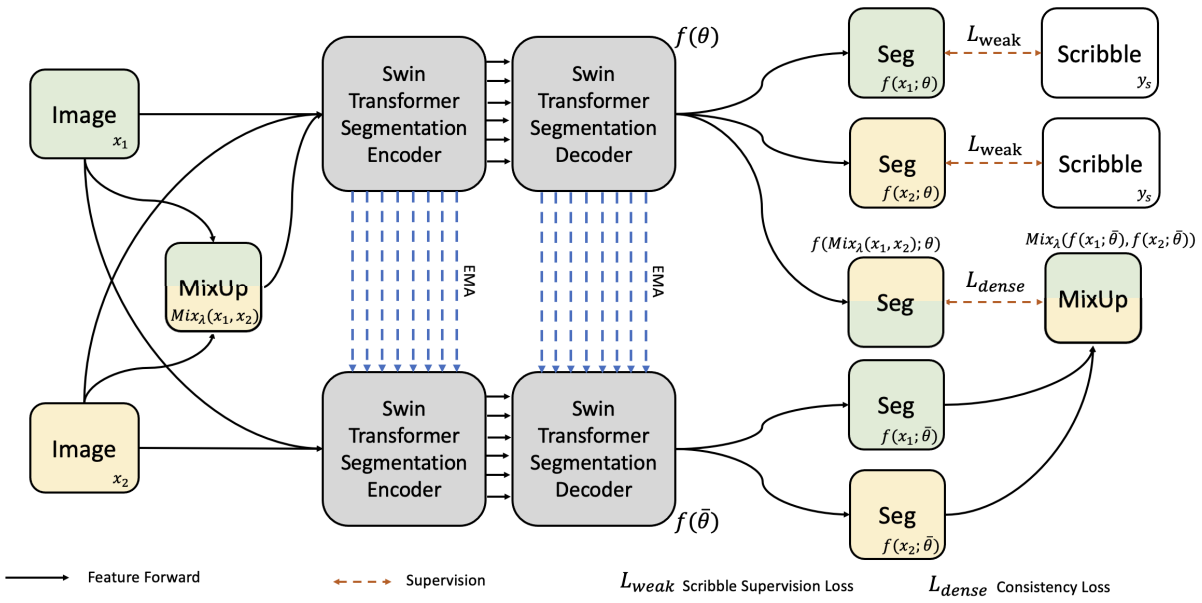


Fig. 2. The Framework of S4RoboFormer for Surgical Robotic Instrument Segmentation with Scribble Annotation.

challenges of sparse signal supervision, we propose the **S**cribble-**S**upervised **S**urgical Robotic **S**egmentation Transformer, named S4RoboFormer. A weakly-supervised learning framework with scribble annotations as the primary source of supervision is proposed. Scribble annotations, which provide sparse and imprecise labels compared to dense pixel-wise labels, require the model to infer information from limited data points. Although our method is primarily weakly-supervised, it integrates concepts from semi-supervised learning, such as consistency regularization. This integration is crucial for leveraging the unlabeled portions of the images, which are abundant due to the nature of scribble annotations. Here, consistency regularization aids in stabilizing the model's predictions across the variably labeled and unlabeled pixels, effectively using the unlabeled data to boost learning efficiency and model robustness.

To the best of our knowledge, S4RoboFormer is the first work to explore ViT for surgical robotic segmentation with WSL scribble-supervised fashion. The contributions are considered to be fourfold:

- Integration of Vision Transformer with scribble-supervised learning for enhanced surgical robotic segmentation.
- A data perturbation strategy employing mixed-interpolation augmentation to facilitate consistency training, enhancing the robustness and reliability of the segmentation process.
- A network perturbation scheme that leverages network self-ensembling to bolster consistency training, further improving the model's performance under various conditions.
- We introduce a pre-processed challenging but more practical dataset specifically tailored for scribble-supervised studies in surgical robotic segmentation for public research.

## II. Approach

In our proposed study of surgical robotic segmentation with scribble supervision, we define the scribble-labeled training dataset as $\mathcal{D}_{train} \in [\boldsymbol{X}, \boldsymbol{Y}_{\text{s}}]^{h\times w}$, where $\boldsymbol{X}$ is the original image, and $\boldsymbol{Y}_{\text{s}} \in [0, 1, \text{None}]$ is the scribble label. Here, 0 and 1 indicate the sparse annotations of surgical robots and the background, respectively, while None signifies the absence of annotation information on the corresponding pixels. Notably, over 90% of the pixels in the training set are annotated with None. The testing set, denoted as $\mathcal{D}_{test} \in [\boldsymbol{X}, \boldsymbol{Y}_{\text{gt}}]^{h\times w}$, where $\boldsymbol{Y}_{\text{gt}} \in [0, 1]$, represents precise pixel-level dense annotations. S4RoboFormer consists of two networks with the same architecture, denoted as $f(\theta)$ and $f(\bar{\theta})$, where $f(\bar{\theta})$ is updated by Exponential Moving Average (EMA) [25] based on $\theta$, aligned with the network self-ensembling scheme. The inference process of S4RoboFormer for given input frames is represented as $f(\theta) : \boldsymbol{X} \mapsto \boldsymbol{Y}_{\text{i}}$, with $\boldsymbol{Y}_{\text{i}}$ being a dense inference, which can potentially serve as a pseudo-label $\boldsymbol{Y}_{\text{p}}$ to enhance supervision signals for retraining the network. The weakly-supervised loss via scribble is represented as $\mathcal{L}_{\text{weak}} : (\boldsymbol{Y}_{\text{i}}, \boldsymbol{Y}_{\text{s}}) \mapsto \mathbb{R}$, and the consistency-aware loss via dense pseudo-label is denoted as $\mathcal{L}_{\text{dense}} : (\boldsymbol{Y}_{\text{i}}, \boldsymbol{Y}_{\text{p}}) \mapsto \mathbb{R}$. The overall objective of the training process is to minimize the total loss by updating the parameters $\theta$ of the network $f$. Performance evaluation is conducted by measuring the difference between $(\boldsymbol{Y}_{\text{i}}, \boldsymbol{Y}_{\text{gt}}) \mapsto \mathbb{R}$. The framework of S4RoboFormer is illustrated in Figure 2.

### A. Training Objective

The overarching training objective of S4RoboFormer is detailed as:

$$\mathcal{L}_{total} = \mathcal{L}_{weak}(y_i, y_s) + \lambda \mathcal{L}_{dense}(y_i, y_p), \tag{1}$$

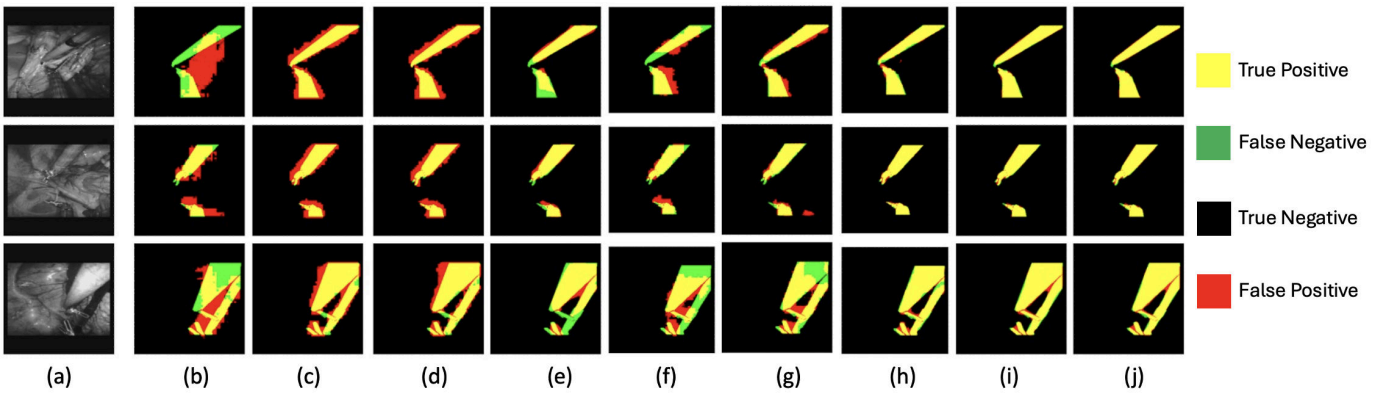


Fig. 3. Three Example Frames During Robotic Surgery with da Vinci Robots, and the Corresponding Segmentation Inference of S4RoboFormer and each of Baseline Method Against Ground Truth. The annotations within the figure provide clearer differentiation between correctly segmented regions (yellow), over-segmentation (red), and under-segmentation (green). (a) Input Raw Images, (b) TV+CNN, (c) TV+ViT, (d) USTM+CNN, (e) pCE+CNN, (f) pCE+ViT, (g) USTM+ViT, (h) Proposed S4RoboFormer, (i) Fully Supervised ViT, (j) Fully Supervised CNN.

where $\lambda$ is with an exponential ramp-up function $\lambda = e^{-5\times(1-t/t_{\max})^2}$, $t$ is the iteration during training, and $t_{\max}$ is the number of total iterations. This setting enables the whole framework initialized by sparse signal scribbles and then gradually move focus to dense signal consistency training [11]. To mitigate the difficulties associated with sparse-signal scribble supervision, we employ a modified CrossEntropy $CE$ function that focuses exclusively on annotated pixels, effectively disregarding unlabeled ones. This adaptation results in a partially supervised segmentation loss. Specifically, we introduce the Partial Cross-Entropy $pCE$, which utilizes only scribble annotations during training of the networks, denoted as $\mathcal{L}_{\text{weak}}$, as delineated in the following Equation 2,

$$\mathcal{L}_{\text{weak}} = -\sum_{i\in\Omega_L}\sum_{k} y_{\text{s}}[i,k]\log(y_{\text{i}}[i,k]), \tag{2}$$

where $i$ represents the index of a given pixel, and $\Omega_L$ denotes the ensemble of pixels that have been annotated with scribbles. The variable $k$ corresponds to the index of a particular class, and the notation $[i,k]$ signifies the likelihood of the $i$-th pixel being of the $k$-th class. Once the S4RoboFormer is initialized, the prior knowledge could be utilized to expand the dataset from sparse signal to dense signal via pseudo label $\boldsymbol{Y}_{\text{p}}$, where the $\mathcal{L}_{\text{dense}}$ is detailed as $\mathcal{L}_{\text{dense}} = \text{CE}(\boldsymbol{Y}_{\text{i}}, \boldsymbol{Y}_{\text{p}}) + \text{Dice}(\boldsymbol{Y}_{\text{i}}, \boldsymbol{Y}_{\text{p}})$ where CE and Dice demonstrate the CrossEntropy-based and the Dice Coefficient-based difference measures, respectively.

### B. Consistency under Network-based Perturbation

Inspired by consistency regularization techniques in semi-supervised learning (SSL), which aim to ensure that a network produces stable outputs when subjected to various perturbations on unlabeled data [17], [20], [3], [30], we propose an enhanced consistency training strategy incorporating both data and network perturbations. To improve feature regularization in networks supervised by sparse signals, we refine $f(\theta)$ through a network self-ensembling method. This method employs the Exponential Moving Average (EMA) procedure [25], producing two networks, $f(\theta)$ and $f(\bar{\theta})$, as illustrated by the group of blue dashed lines ⇢ in Figure 2.

We introduce Gaussian noise to the input data to achieve perturbation. The inference of $f(\theta)$ is represented as $\boldsymbol{Y}_{\text{i}} = f(\boldsymbol{X} + Noise; \theta)$, where $\boldsymbol{X}$ represents the input frames from robotic surgery, and $\boldsymbol{Y}_{\text{i}}$ denotes the prediction generated by $f(\theta)$. In the EMA procedure, $f(\theta)$ learns directly from data with scribble annotations $\boldsymbol{Y}_{\text{s}}$ and dense signal pseudo labels $\boldsymbol{Y}_{\text{p}}$, while $f(\bar{\theta})$ updates its parameters according to the EMA rule, as $\bar{\theta}_i = \alpha\theta_i + (1-\alpha)\theta_{i-1}$. Here, $\alpha \in [0,1]$ is a balance weight for the update, and $i$ indicates the i-th step in the training process. The predictions from $f(\bar{\theta})$ are generally more accurate than those from $f(\theta)$, making them well-suited for use as pseudo labels [25].

### C. Consistency under Data-based Perturbation

Following the [26], to regularize SSL by encouraging consistent predictions, a procedure of feature interpolation is involved and shown in Equation 3,

$$f(\alpha X_1 + (1-\alpha)X_2) = \alpha f(X_1) + (1-\alpha)f(X_2), \tag{3}$$

where $X_1, X_2$ are two separate feature as the input of network $f$, and $\alpha$ is a hyper-parameter weight for mixed-interpolation. Equation 3 suggests that the inference of a mixed input image should be similar to the mixed inference given two separate images. Given the MixUp operation in a previous SSL study for the pixel-level data augmentation [36], which is illustrated in Equation 4,

$$Mix_\lambda(a,b) = \lambda \times a + (1-\lambda)\times b, \tag{4}$$

where $a, b$ are both data features, and the $\lambda$ is a weight setting factor when mixing two data. The $Mix$ operation is applied directly to each pixel within images and their corresponding inferences. Our ViT-based segmentation network, denoted as $f$, is designed to provide consistent predictions for interpolated unlabeled pixels within images. The goal of the consistency

training is to achieve reliable and uniform segmentation for image pixels that are interpolated from existing data points. This objective is encapsulated in the following Equation 5,

$$f(Mix_\lambda(X_1, X_2); \theta) \approx Mix_\lambda(f(X_1; \overline{\theta}), f(X_2; \overline{\theta})), \quad (5)$$

where $X_1$ and $X_2$ are two separate images. The $\approx$ operation is achieved by updating parameters $\theta$ of network $f$ by minimizing the loss $\mathcal{L}_{\text{dense}}$, which is briefly sketched in Figure 2.

## III. Experiments

**Dataset** The original dataset is from MICCAI Robotic Instrument Segmentation Challenge 2017 [1]. The dataset consists of $8 \times 225$-frame robotic surgical videos, captured at 2 Hz (8 videos and each with 255 frames). The Intuitive Surgical Robotic Instruments have been manually labelled as various types of instruments. In the WSL experiment, we considered a binary instrument segmentation task, where each frame was separated into surgical robot or a background class. All the frames are normalized to [0,1] and resized to $224 \times 224$. Scribble annotations, which require annotators to draw lines over the regions of interest rather than meticulously label every pixel, significantly reduce the time and effort involved in preparing training data. This method not only speeds up the annotation process but also reduces the cognitive load on annotators, making it especially suitable for medical applications where expert time is valuable. To ensure a fair evaluation, the scribble labels are generated based on published dense annotation data by a published data pre-processing technique following common studies [31].

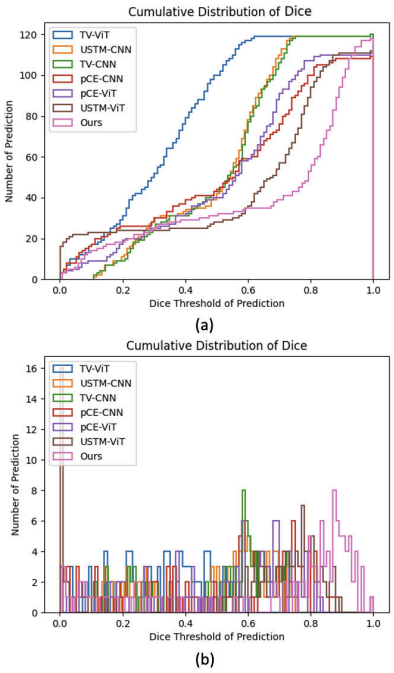


Fig. 4. The Distribution of Dice-Coefficient for Each Segmentation Inference Against Ground Truth on Test Set.

**Implementation** Our implementation was developed using Python, and PyTorch. The hardware setup comprised an Nvidia GeForce RTX 3090 GPU and an Intel Core i9-10900K processor. The S4RoboFormer model underwent training over 30,000 iterations with a batch size of 24. SGD is selected as the optimizer, with an initial learning rate of 0.01, a momentum of 0.9, and a weight decay of 0.0001. This experimental setup was consistently applied across all baseline methods without modifications. The computational cost of solely training each network backbone is reported in Table I.

TABLE I
The Computational Cost of Backbone Network on Nvidia RTX 3090.

| | $\text{FLOPS}_{10^9}$ | $\text{Par}_{10^6}$ | $\text{Memory}_{GB}$ |
|---|---|---|---|
| CNN | 1.9694 | 1.0809 | 0.5 |
| ViT | 8.6929 | 41.3415 | 1.4 |

**Baseline Methods** The baseline for WSL strategies include pCE [23], USTM [15], TV [9], ScribbleSup [13], and CRF [24], respectively. For a fair and comprehensive comparison of segmentation networks, the segmentation backbone network for all baseline weakly-supervised frameworks and S4RoboFormer are with CNN-based UNet [21], Swin Vision Transformer-based UNet, SwinUNet [4]. Fully supervised learning with segmentation backbone networks are also validated as the upper-boundary performance.

**Qualitative Results** Examples of inference against ground truth of all methods are sketched in Figure 3 as qualitative results, including three randomly selected raw images with corresponding predicted inference against the published ground truth, where Yellow, Red, Green and Black represent True Positive, False Positive, False Negative and True Negative inferences at pixel level, respectively. The qualitative results in Figure 3 demonstrate that S4RoboFormer is more likely to predict True Positive and True Negative results on pixel level, especially on the boundary of the region of interest. S4RoboFormer achieves very similar results compared to fully supervised learning.

**Quantitative Results** The direct quantitative comparison experiments between S4RoboFormer and other baseline methods are conducted with a variety of evaluation metrics, including Dice, Accuracy (ACC), Precision (PRE), Sensitivity (SEN), and Specificity (SPE). The results of the average performance on test set are detailed in Table II where the reported numbers are the higher, the better, indicating with $\uparrow$. The best performance except the fully supervised learning is highlighted with **Bold**. We have conducted independent two-sample t-tests to compare the performance metrics across different models. The p-values obtained from these tests ($p < 0.0005$) indicate that the improvements in performance metrics by S4RoboFormer are statistically significant when compared to the baseline methods.

**Ablation Study** We have conducted an ablation study to evaluate the specific contributions of the various components of S4RoboFormer. This study systematically evaluates the impact of the ViT backbone, the network-based perturbation and data-based perturbation for consistency training. The detailed results are introduced in Table III.

TABLE II
THE DIRECT COMPARISON OF WEAKLY-SUPERVISED FRAMEWORKS ON THE TEST SET.

| WSL | Net | Dice↑ | ACC↑ | PRE↑ | SEN↑ | SPE↑ |
|---|---|---|---|---|---|---|
| pCE [23] | ViT | 0.6396 | 0.9082 | 0.5110 | 0.8547 | 0.9139 |
| USTM [15] | ViT | 0.6482 | 0.9154 | 0.5367 | 0.8181 | 0.9256 |
| TV [9] | ViT | 0.6092 | 0.8907 | 0.4620 | 0.8940 | 0.8904 |
| ScribbleSup [13] | ViT | 0.6389 | 0.8955 | 0.4968 | 0.8948 | 0.8956 |
| CRF [24] | ViT | 0.6892 | 0.8956 | 0.5605 | 0.8947 | 0.8957 |
| pCE [23] | CNN | 0.6298 | 0.9032 | 0.4955 | 0.8641 | 0.9073 |
| USTM [15] | CNN | 0.6193 | 0.9004 | 0.4870 | 0.8502 | 0.9057 |
| TV [9] | CNN | 0.6258 | 0.9052 | 0.5016 | 0.8315 | 0.9130 |
| ScribbleSup [13] | CNN | 0.6692 | 0.8966 | 0.5345 | 0.8946 | 0.8969 |
| CRF [24] | CNN | 0.7088 | 0.8967 | 0.5870 | 0.8944 | 0.8971 |
| **Ours** | | **0.7294** | **0.9271** | **0.6087** | **0.9099** | **0.9292** |
| Fully Supervised CNN | | 0.7811 | 0.9601 | 0.8188 | 0.7469 | 0.9826 |
| Fully Supervised ViT | | 0.7637 | 0.9545 | 0.7556 | 0.7720 | 0.9737 |

TABLE III
THE ABLATION STUDY OF BACKBONE NETWORK & CONSISTENCY TRAINING.

| Net | Data-Perturbation | Network-Perturbation | Dice↑ |
|---|---|---|---|
| CNN | | | 0.6298 |
| CNN | | ✓ | 0.6239 |
| CNN | ✓ | | 0.6864 |
| CNN | ✓ | ✓ | 0.7148 |
| ViT | | | 0.6396 |
| ViT | | ✓ | 0.6749 |
| ViT | ✓ | | 0.6946 |
| **(Ours)** ViT | ✓ | ✓ | **0.7294** |

In addition to reporting the average performance on the test set, we also evaluate the inference for each test set image by analyzing the distribution of Dice Coefficients. Figure 4 provides a visual representation of the model's segmentation performance variability. Figure 4 (a) presents a line chart where the X-axis represents the Dice Coefficient threshold, and the Y-axis indicates the number of inferences with Dice performance below that threshold. A steeper curve suggests that the method is more likely to achieve high Dice values. Additionally, Figure 4 (b) further reports the distribution of Dice Coefficients with histogram. The X-axis represents different Dice score intervals, while the Y-axis indicates the frequency of test images falling within these intervals. A higher frequency in higher Dice score intervals signifies a greater consistency and accuracy of the S4RoboFormer in segmenting surgical instruments accurately across a variety of images.

## IV. CONCLUSION

The S4RoboFormer integrates the advanced capabilities of Vision Transformers with the efficiency of scribble-supervised learning to address the challenges of medical image annotation in robotic-assisted surgeries. By employing scribble annotations, which are derived through a streamlined conversion of dense labels, our approach markedly reduces the resources and time required for data preparation without compromising the quality of the training data. Our dual consistency training enhances network adaptability and learning efficiency, particularly in varied surgical scenarios with sparse signal supervision. The promising results demonstrate the potential for broader application of surgical techniques.

## REFERENCES


[1] M. Allan, A. Shvets, T. Kurmann, Z. Zhang, R. Duggal, Y.-H. Su, N. Rieke, I. Laina, N. Kalavakonda, S. Bodenstedt *et al.*, "2017 robotic instrument segmentation challenge," *arXiv preprint arXiv:1902.06426*, 2019.

[2] S. Bodenstedt, M. Allan, A. Agustinos, X. Du, L. Garcia-Peraza-Herrera, H. Kenngott, T. Kurmann, B. Müller-Stich, S. Ourselin, D. Pakhomov *et al.*, "Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery," *arXiv preprint arXiv:1805.02475*, 2018.

[3] G. Bortsova *et al.*, "Semi-supervised medical image segmentation via learning consistency under transformations," in *MICCAI*. Springer, 2019.

[4] 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," in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[5] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[6] M. Grammatikopoulou *et al.*, "Cadis: Cataract dataset for image segmentation," *arXiv preprint arXiv:1906.11586*, 2019.

[7] S. K. Hasan and C. A. Linte, "U-netplus: A modified encoder-decoder u-net architecture for semantic and instance segmentation of surgical instruments from laparoscopic images," in *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2019, pp. 7205–7211.

[8] M. Islam, Y. Li, and H. Ren, "Learning where to look while tracking instruments in robot-assisted surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 412–420.

[9] M. Javanmardi *et al.*, "Unsupervised total variation loss for semi-supervised deep learning of semantic segmentation," *arXiv preprint arXiv:1605.01368*, 2016.

[10] Y. Jin, K. Cheng, Q. Dou, and P.-A. Heng, "Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22*. Springer, 2019, pp. 440–448.

[11] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[12] M. Lerousseau, M. Vakalopoulou, M. Classe, J. Adam, E. Battistella, A. Carré, T. Estienne, T. Henry, E. Deutsch, and N. Paragios, "Weakly supervised multiple instance learning histopathological tumor segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*. Springer, 2020, pp. 470–479.

[13] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3159–3167.

[14] D. Liu, Y. Wei, T. Jiang, Y. Wang, R. Miao, F. Shan, and Z. Li, "Unsupervised surgical instrument segmentation via anchor generation and semantic diffusion," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference,*

*Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*. Springer, 2020, pp. 657–667.

[15] X. Liu, Q. Yuan, Y. Gao, K. He, S. Wang, X. Tang, J. Tang, and D. Shen, "Weakly supervised segmentation of covid19 infection with scribble annotation on ct images," *Pattern recognition*, vol. 122, p. 108341, 2022.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[17] T. Miyato *et al.*, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE TPAMI*, 2018.

[18] K. B. Ozyoruk *et al.*, "Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos," *MedIA*, 2021.

[19] J. H. Palep, "Robotic assisted minimally invasive surgery," *Journal of minimal access surgery*, 2009.

[20] S. Park *et al.*, "Adversarial dropout for supervised and semi-supervised learning," in *AAAI*, 2018.

[21] 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: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[22] A. A. Shvets, A. Rakhlin, A. A. Kalinin, and V. I. Iglovikov, "Automatic instrument segmentation in robot-assisted surgery using deep learning," in *ICMLA*. IEEE, 2018.

[23] M. Tang, A. Djelouah, F. Perazzi, Y. Boykov, and C. Schroers, "Normalized cut loss for weakly-supervised cnn segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1818–1827.

[24] M. Tang, F. Perazzi, A. Djelouah, I. Ben Ayed, C. Schroers, and Y. Boykov, "On regularized losses for weakly-supervised cnn segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 507–522.

[25] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[26] V. Verma *et al.*, "Interpolation consistency training for semi-supervised learning," *Neural Networks*, 2022.

[27] Z. Wang, N. Dong, and I. Voiculescu, "Computationally-efficient vision transformer for medical image semantic segmentation via dual pseudo-label supervision," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 1961–1965.

[28] Z. Wang *et al.*, "Adversarial vision transformer for medical image semantic segmentation with limited annotations," *BMVC*, 2022.

[29] Z. Wang and C. Ma, "Weak-mamba-unet: Visual mamba makes cnn and vit work better for scribble-based medical image segmentation," *arXiv preprint arXiv:2402.10887*, 2024.

[30] ——, "Dual-contrastive dual-consistency dual-transformer: A semi-supervised approach to medical image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 870–879.

[31] Z. Wang and I. Voiculescu, "Weakly supervised medical image segmentation through dense combinations of dense pseudo-labels," in *MICCAI Workshop on Data Engineering in Medical Imaging*. Springer, 2023, pp. 1–10.

[32] Z. Wang and C. Yang, "Mixsegnet: Fusing multiple mixed-supervisory signals with multiple views of networks for mixed-supervised medical image segmentation," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108059, 2024.

[33] Z. Wang, Z. Zhang, and I. Voiculescu, "Rar-u-net: a residual encoder to attention decoder by residual connections framework for spine segmentation under noisy labels," in *2021 IEEE international conference on image processing (ICIP)*. IEEE, 2021, pp. 21–25.

[34] Z. Yang, R. Simon, and C. Linte, "A weakly supervised learning approach for surgical instrument segmentation from laparoscopic video sequences," in *Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 12034. SPIE, 2022, pp. 412–417.

[35] Y. Yu, Z. Zhao, Y. Jin, G. Chen, Q. Dou, and P.-A. Heng, "Pseudo-label guided cross-video pixel contrast for robotic surgical scene segmentation with limited annotations," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 10 857–10 864.

[36] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

# S4RoboFormer: scribble-supervised surgical robotic segmentation transformer via augmented consistency training

Wang, Ziyang

2025-8



Wang Z, Chen T, Ye Z, et al., (2025) S4RoboFormer: scribble-supervised surgical robotic segmentation transformer via augmented consistency training. IEEE Transactions on Medical Robotics and Bionics, Available online 29 August 2025

https://doi.org/10.1109/tmrb.2025.3604103