Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Revisiting instrument segmentation: learning from decentralized surgical sequences with various imperfect annotations

Zhou Zheng[1] | **Yuichiro Hayashi**[1] | **Masahiro Oda**[2,1] | **Takayuki Kitasaka**[3] | **Kensaku Mori**[1,2,4]

[1]Graduate School of Informatics, Nagoya University, Chikusa-ku, Nagoya, Aichi, Japan

[2]Information Strategy Office, Information and Communications, Nagoya University, Chikusa-ku, Nagoya, Aichi, Japan

[3]School of Information Science, Aichi Institute of Technology, Yagusa-cho, Toyota, Aichi, Japan

[4]Research Center for Medical Bigdata, National Institute of Informatics, Chiyoda-ku, Tokyo, Japan

Correspondence
Zhou Zheng
Email: zzheng@mori.m.is.nagoya-u.ac.jp
Kensaku Mori
Email: kensaku@is.nagoya-u.ac.jp

Abstract

This paper focuses on a new and challenging problem related to robotic instrument segmentation. We aim to learn a generalizable model from distributed surgical datasets with various imperfect annotations for instrument segmentation. Under practical conditions, curating a large-scale surgical dataset for centralized learning is usually impeded due to data silos and privacy issues. Besides, local clients, such as hospitals or medical institutes, may hold datasets with a diverse array of imperfect annotations. These datasets can include scarce annotations (a large portion of samples are unlabeled), noisy labels prone to errors, and scribble annotations with less precision. Federated learning (FL) has emerged as an attractive paradigm for developing statistical models with these locally distributed datasets. However, its potential in instrument segmentation has yet to be fully investigated. Moreover, the problem of learning from various imperfect annotations in an FL setup is rarely studied, even though it presents a more practical and beneficial scenario. In this work, we rethink instrument segmentation under such a setting. We propose an effective FL framework where local clients utilize their imperfectly annotated data to train local models while a central server carries out model updating and aggregation to coordinate these clients to prepare a global model. We simulated this novel problem setting with an assumption that each client owns a single type of annotation (scarce, noisy, or sparse annotation) and conducted validation using the EndoVis17 dataset. Our method presented a potential solution for this issue. Notably, our approach surpassed centralized learning under various imperfect annotation settings. Our method established a foundational benchmark, and future work can build upon it by considering each client owning various annotations, aligning closer with real-world complexities.

KEYWORDS
robotic instrument, segmentation, imperfect annotations, federated learning

1 | INTRODUCTION

Robust and accurate surgical instrument segmentation serves as the cornerstone for potential applications such as instrument tracking and augmented reality within the domain of robotic minimally invasive surgery. Deep learning-based methodologies, especially the deep convolutional neural networks, have been state-of-the-art solutions for this task, driving a series of significant advancements in this field [1,2].

Conventionally, establishing a high-performing, generalizable model for instrument segmentation is contingent on centralized learning [3,4,5]. This process necessitates the collection of large-scale surgical videos/sequences for training. However, practical implementation is often limited due to stringent privacy and confidentiality-related regulations that restrict the sharing and collecting sensitive surgical data [6]. In light of these limitations of centralized learning, federated learning (FL) is emerging as an appealing alternative, enabling multiple clients or institutes to collaboratively prepare a global model without sharing local datasets. However, the application of FL in the context of instrument segmentation lacks exploration.

Moreover, the problem of learning from various imperfect annotations [7] in an FL setup is hardly investigated, even though it presents a more practical and advantageous scenario. For instance, due to the costly and time-consuming annotation process,

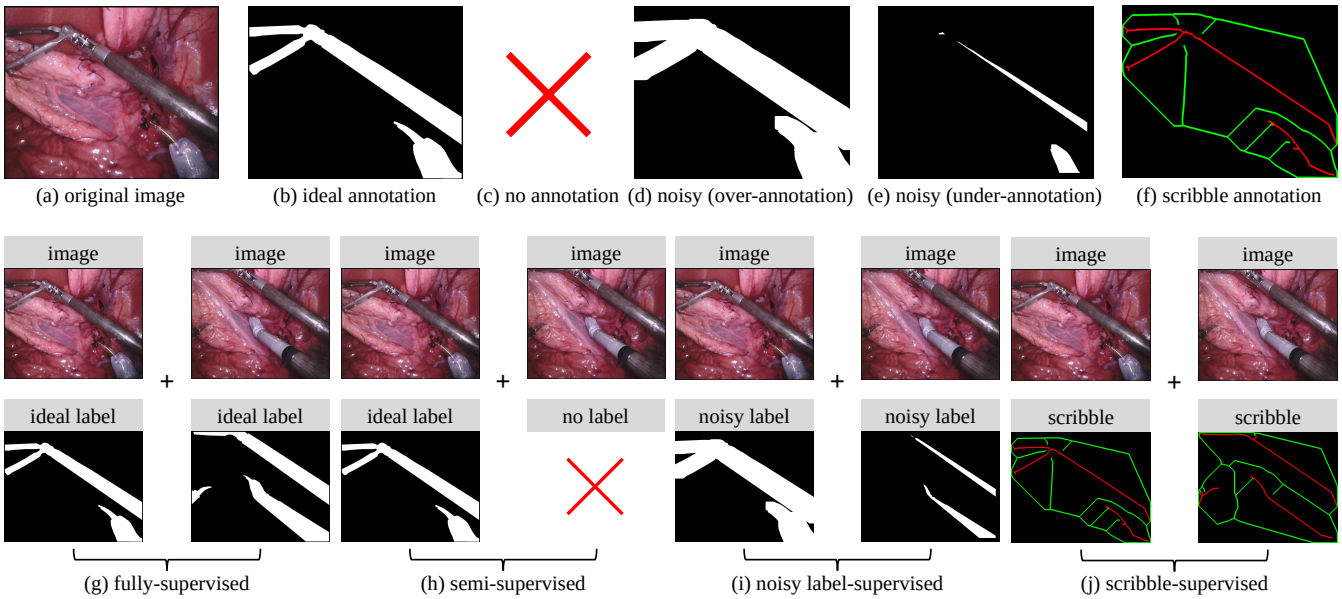


FIGURE 1 Illustration of various annotations and comparison of segmentation paradigms with various annotations. Note that endoscopic images are from EndoVis17[1]. (a) Original image. (b) Ideal annotation (ground truth). (c) Scarce annotations, where a large of images have no annotation. (d)-(e) Noisy annotations, where we simulate two types of noisy labels, i.e., over-annotation of foreground using dilation operation and under-annotation of foreground using erosion operation. (f) Scribble annotation, where only a small part of foreground (in **red**) and background (in **green**) pixels are annotated, leaving a large portion (in **black**) unlabeled. Note that we further dilate the skeleton lines for better visualization. (g) Fully-supervised segmentation utilizes ideal annotations for learning. (h) Semi-supervised segmentation leverages a combination of a small proportion of labeled data and a large quantity of unlabeled data. (i) Noisy label-supervised segmentation aims to minimize the adverse impact of label noise during model training. (j) Scribble-supervised segmentation capitalizes on sparse supervision signals.

a large portion of images might remain unlabeled at some clients[8,9], resulting in scarce annotations. Some clients might then resort to using weak annotations, e.g., scribble annotations[10,11]. In addition, noisy annotations are sometimes unavoidable due to inter-observer variability, as the annotation process is inherently subjective[12,13]. Examples of the ideal, scarce, noisy†, and scribble‡ annotations are illustrated in Fig. 1(a) through Fig. 1(f). As local clients could face various imperfect annotations, this complexity implies that local clients need to maximally utilize these imperfect annotations to contribute an effective local model for global model preparation.

To this end, we start a first attempt to study a more challenging problem related to instrument segmentation. We aim to learn instrument segmentation from distributed surgical sequences with various imperfect annotations. Instrument segmentation in real-world scenarios inherently introduces complexities, particularly when each client possesses multiple annotation types. These diverse annotations carry unique challenges and demand tailored strategies, elevating the intricacy of the problem. In our study, we proceed with an assumption that each client has just one type of annotation. Such an assumption allows us to focus on a specific and rudimentary condition of this problem. By tackling the challenge from this perspective, we can explore a potential solution and provide an initial baseline. Subsequent research can expand on this foundation by gradually integrating additional complexities, such as accommodating multiple annotation types per client, to strive for a more comprehensive approach that better reflects real-world scenarios.

From the above perspectives, we propose a practical FL framework to handle this problem and demonstrate a potential solution. Specifically, we adopt existing mainstream methodologies, e.g., semi-supervised learning[15], noisy-label learning[13], and scribble-supervised learning[16], and show that these specific methods can be integrated to our framework to endow local

† We simulate noisy annotations with dilation and erosion operations, following the work[12].
‡ We simulate scribble annotations by skeletonizing[14] the ground truth.

clients the ability to deal with imperfect annotations. Furthermore, we maintain a central server to coordinate these local clients to prepare a generalizable model using these distributed datasets with imperfect annotations. Our contributions are as follows:

- We reevaluate instrument segmentation from the perspective of a new and more challenging problem setting. We aim to create a generalizable model for instrument segmentation using distributed datasets with imperfect annotations.
- We present a novel and effective framework to address this new problem by unifying semi-, noisy label-, and scribble-supervised segmentation and FL. We show that the careful integration of existing advanced techniques can provide successful solutions for instrument segmentation, taking into account both data availability and imperfect annotations.
- We conduct experiments with the EndoVis17 dataset [1] to present our method's efficacy. Our approach surpasses centralized learning under various imperfect annotation settings, underscoring its potential to tackle this innovative challenge. Our method can serve as an initial baseline in addressing this problem.

2 | RELATED WORK

In this paper, different from existing instrument segmentation methods, we unify semi-, noisy label-, and scribble-supervised segmentation and FL to address instrument segmentation while considering data privacy and imperfect annotation-related issues, leading to a novel and more challenging application. In the following, we review related literature on instrument segmentation, learning segmentation from imperfect annotations, and FL.

Instrument segmentation. Instrument segmentation in robotic surgery [1,17] plays an important role in enhancing surgical procedures. Over the past few years, it has attracted increasing interest, particularly with the advent and rise of robotic platforms such as da Vinci®. Deep convolutional neural networks (CNNs) have emerged as the dominant solutions, surpassing traditional schemes by delivering automatic and highly accurate results. There is a line of advanced CNN models [2,3,4,5,18,19,20] showing promising performance in this task. Nevertheless, most existing methods are introduced under the centralized learning setting, where a large-scale dataset is collected for training. In reality, data centralization is often limited due to privacy-related issues, especially for medical data like surgical sequences. By contrast, the approach of learning from distributed datasets, which aligns more with practical scenarios, remains less explored. In our work, we revisit the challenge of instrument segmentation within a practical FL setting.

Learning segmentation from imperfect annotations. In real-world scenarios, image datasets are often accompanied by imperfect labels, such as scarce, noisy, and scribble annotations [7]. The quality of annotations is a crucial factor influencing the performance of learned models, making it imperative to leverage these inferior annotations. For instance, semi-supervised segmentation [8,9,21] endeavors to make the most use of unlabeled data. Scribble-supervised segmentation [10,11] makes an effort to learn a model from sparse supervision signals. Besides, other efforts like designing a noise-tolerance loss function [13] and correcting noise labels during training [12] have been studied to handle label noises. Fig. 1(g) through Fig. 1(j) show the comparison among different segmentation diagrams. Although numerous attempts have been made to deal with inferior annotations, handling various imperfect annotations simultaneously within a more pragmatic FL framework is more practical but remains unexplored.

Federated learning. Federated learning (FL) [22,23,24] is a machine learning mechanism that enables multiple clients or devices to collaboratively learn a statistical model while keeping their data localized, effectively addressing privacy concerns in sensitive domains like healthcare. Despite the widespread application of FL in medical imaging [25,26,27], learning from various imperfect annotations remains unexplored in FL, where we need to tackle the challenges related to imperfect annotations of local datasets. We posit that investigating this problem is crucial, as it presents a practical and more challenging situation.

3 | METHOD

3.1 | Overview

Assuming there are K clients, denoted as $\{C_i\}_{i=1}^{K}$, where each client C_i holds a private dataset. Ideally, each dataset is expected to contain images and the corresponding ideal annotations (ground truth). However, in reality, local datasets may

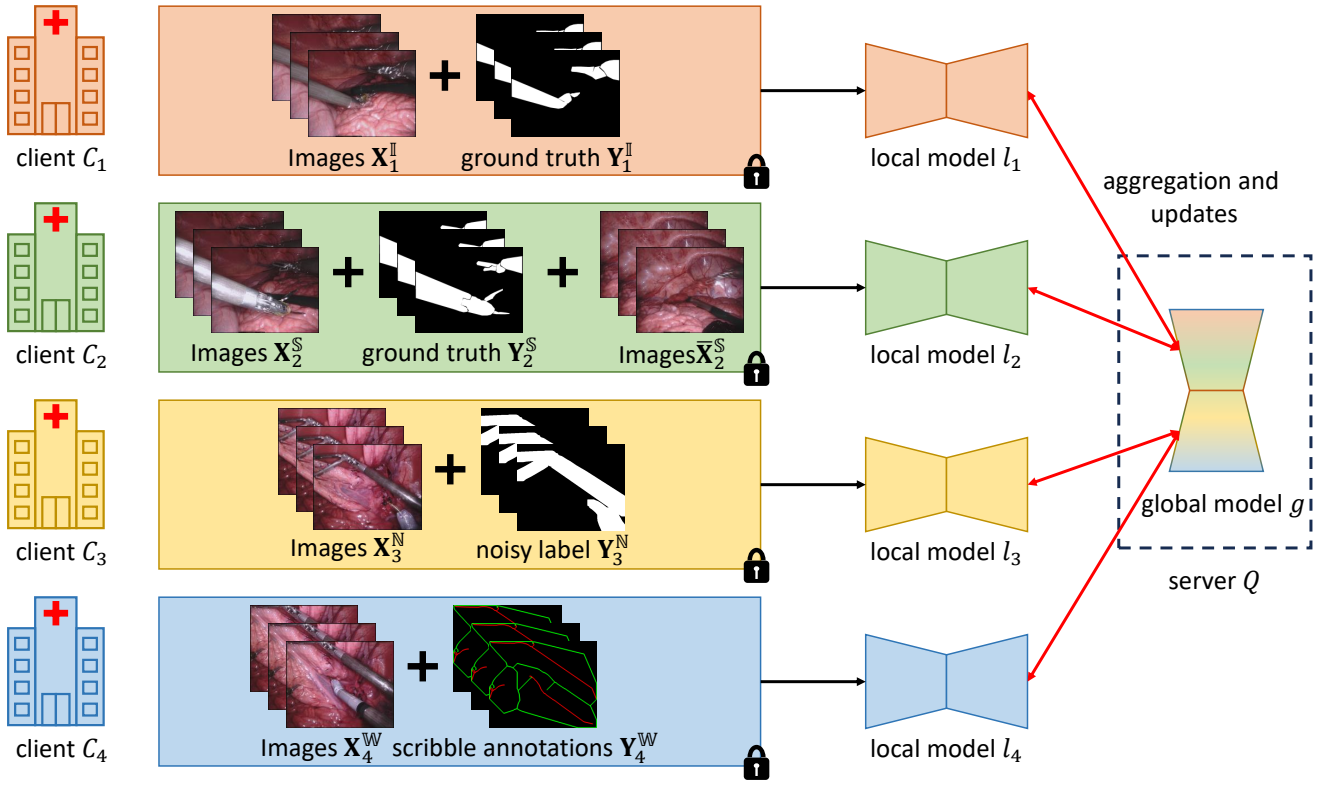


FIGURE 2 Flowchart of proposed framework. This illustration depicts our solution to the novel and practical problem regarding data silos and imperfect annotations within instrument segmentation from surgical sequences, with an assumption that each local client only possesses a dataset with one type of annotation. For clarity, we feature four local clients $\{C_i\}_{i=1}^4$ alongside a central server Q. Local clients $\{C_i\}_{i=1}^4$ signify distinct data repositories, and their datasets reflect diverse annotations, i.e., ideal annotations (ground truth), scarce annotations (a large portion of images are unlabeled), noisy annotations, and scribble annotations. Local clients adopt specific strategies, i.e., fully-, semi-, noisy label-, and scribble-supervised learning, to prepare their respective local models $\{l_i\}_{i=1}^4$ by making use of their distinctive datasets $\mathbb{I}_1 = \{\mathbf{X}_1^{\mathbb{I}}, \mathbf{Y}_1^{\mathbb{I}}\}$, $\mathbb{S}_2 = \{\mathbf{X}_2^{\mathbb{S}}, \mathbf{Y}_2^{\mathbb{S}}, \bar{\mathbf{X}}_2^{\mathbb{S}}\}$, $\mathbb{N}_3 = \{\mathbf{X}_3^{\mathbb{N}}, \mathbf{Y}_3^{\mathbb{N}}\}$, and $\mathbb{W}_4 = \{\mathbf{X}_4^{\mathbb{W}}, \mathbf{Y}_4^{\mathbb{W}}\}$. The central server Q manages model aggregation and updates with the FedAvg scheme[22], guiding local clients toward a generalizable global model g.

come with imperfect annotations, such as scarce, noisy, and scribble annotations. Our goal is to learn a generalizable model with these distributed datasets with imperfect annotations.

In our study, we represent $\mathbb{I}_i = \{\mathbf{X}_i^{\mathbb{I}}, \mathbf{Y}_i^{\mathbb{I}}\}$ as the local dataset consisting of images $\mathbf{X}_i^{\mathbb{I}}$ and the corresponding ideal annotations $\mathbf{Y}_i^{\mathbb{I}}$. For the local dataset with scarce annotations, namely, the dataset comprising a small part of labeled images and a large part of unlabeled images, we denote it as $\mathbb{S}_i = \left\{\mathbf{X}_i^{\mathbb{S}}, \mathbf{Y}_i^{\mathbb{S}}, \bar{\mathbf{X}}_i^{\mathbb{S}}\right\}$, where $\mathbf{X}_i^{\mathbb{S}}$ and $\mathbf{Y}_i^{\mathbb{S}}$ represent the labeled images and the corresponding ground truth, and $\bar{\mathbf{X}}_i^{\mathbb{S}}$ indicates the unlabeled images. We further represent $\mathbb{N}_i = \{\mathbf{X}_i^{\mathbb{N}}, \mathbf{Y}_i^{\mathbb{N}}\}$ as the local dataset containing images $\mathbf{X}_i^{\mathbb{N}}$ and the related noisy annotations $\mathbf{Y}_i^{\mathbb{N}}$, $\mathbb{W}_i = \{\mathbf{X}_i^{\mathbb{W}}, \mathbf{Y}_i^{\mathbb{W}}\}$ as the local dataset comprising images $\mathbf{X}_i^{\mathbb{W}}$ and the related scribble annotations $\mathbf{Y}_i^{\mathbb{W}}$.

Following the spirit of FL, we maintain a central server Q to coordinate these local clients $\{C_i\}_{i=1}^K$ and their local models $\{l_i(\cdot, \mathbf{w}_i^l)\}_{i=1}^K$ to develop a global model $g(\cdot, \mathbf{w}^g)$, where $\mathbf{w}$ denotes model weights.

The flowchart of our framework is depicted in Fig. 2. For simplicity and ease of understanding, four local clients $\{C_i\}_{i=1}^4$ and one central server Q are illustrated. Within the framework, local clients $\{C_i\}_{i=1}^4$ train their respective local models $\{l_i\}_{i=1}^4$ using their private datasets $\mathbb{I}_1 = \{\mathbf{X}_1^{\mathbb{I}}, \mathbf{Y}_1^{\mathbb{I}}\}$, $\mathbb{S}_2 = \{\mathbf{X}_2^{\mathbb{S}}, \mathbf{Y}_2^{\mathbb{S}}, \bar{\mathbf{X}}_2^{\mathbb{S}}\}$, $\mathbb{N}_3 = \{\mathbf{X}_3^{\mathbb{N}}, \mathbf{Y}_3^{\mathbb{N}}\}$, and $\mathbb{W}_4 = \{\mathbf{X}_4^{\mathbb{W}}, \mathbf{Y}_4^{\mathbb{W}}\}$. These datasets come with various annotations, i.e., ideal annotations (ground truth), scarce annotations (a large part of samples remain unlabeled), noisy annotations, and scribble annotations. The central server Q orchestrates the process of model aggregation and updates, coordinating local clients to develop a generalizable global model g.

3.2 | Local training with various annotations

Learning from ideal annotations. For a client C_i holding a dataset $\mathbb{I}_i = \left\{\mathbf{X}_i^{\mathbb{I}}, \mathbf{Y}_i^{\mathbb{I}}\right\}$ that has ideal annotations (ground truth), such as client C_1 in Fig. 2, we can simply perform fully-supervised learning to train the local model. During this process, we input images $\mathbf{X}_i^{\mathbb{I}}$ to the local model, and obtain probability maps $\hat{\mathbf{Y}}_i^{\mathbb{I}}$. We can then calculate the supervised loss, e.g., the cross-entropy (CE) loss $\mathcal{L}_{ce}\left(\hat{\mathbf{Y}}_i^{\mathbb{I}}, \mathbf{Y}_i^{\mathbb{I}}\right)$ between the probability maps $\hat{\mathbf{Y}}_i^{\mathbb{I}}$ and the ground truth $\mathbf{Y}_i^{\mathbb{I}}$. Thus, in this setting, the training objective $\mathcal{L}_i^{\mathbb{I}}$ can be expressed as

$$\mathcal{L}_i^{\mathbb{I}} = \mathcal{L}_{ce}\left(\hat{\mathbf{Y}}_i^{\mathbb{I}}, \mathbf{Y}_i^{\mathbb{I}}\right) = -\sum_{j=1}^{J} \mathbf{Y}_{i,j}^{\mathbb{I}} \log\left(\hat{\mathbf{Y}}_{i,j}^{\mathbb{I}}\right), \tag{1}$$

where J is the number of classes, and $\mathbf{Y}_{i,j}^{\mathbb{I}}$ and $\hat{\mathbf{Y}}_{i,j}^{\mathbb{I}}$ represent the j-th channel of the ground truth $\mathbf{Y}_i^{\mathbb{I}}$ and probability maps $\hat{\mathbf{Y}}_i^{\mathbb{I}}$.

Learning from scarce annotations. A local client C_i such as C_2 in Fig. 2 may hold a dataset $\mathbb{S}_i = \left\{\mathbf{X}_i^{\mathbb{S}}, \mathbf{Y}_i^{\mathbb{S}}, \bar{\mathbf{X}}_i^{\mathbb{S}}\right\}$ with scarce annotations. We employ the well-regarded mean-teacher (MT)[15] framework to train the local model $l_i\left(\cdot, \mathbf{w}_i^l\right)$. Specifically, we sustain an additional teacher model $l_i\left(\cdot, \bar{\mathbf{w}}_i^l\right)$ alongside the local model. $l_i\left(\cdot, \bar{\mathbf{w}}_i^l\right)$ shares the same architecture as $l_i\left(\cdot, \mathbf{w}_i^l\right)$ but maintains the exponential moving average (EMA) weights of $l_i\left(\cdot, \mathbf{w}_i^l\right)$. During training, for the labeled data, images $\mathbf{X}_i^{\mathbb{S}}$ are fed into $l_i\left(\cdot, \mathbf{w}_i^l\right)$ to get the probability maps $\hat{\mathbf{Y}}_i^{\mathbb{S}}$, and the CE loss $\mathcal{L}_{ce}\left(\hat{\mathbf{Y}}_i^{\mathbb{S}}, \mathbf{Y}_i^{\mathbb{S}}\right)$ is then calculated between $\hat{\mathbf{Y}}_i^{\mathbb{S}}$ and the ground truth $\mathbf{Y}_i^{\mathbb{S}}$. For the unlabeled data, images $\bar{\mathbf{X}}_i^{\mathbb{S}}$ are input into $l_i\left(\cdot, \mathbf{w}_i^l\right)$ and $l_i\left(\cdot, \bar{\mathbf{w}}_i^l\right)$ to get probability maps $\bar{\mathbf{Y}}_i^{\mathbb{S}}$ and $\tilde{\mathbf{Y}}_i^{\mathbb{S}}$. We then impose a consistency loss, such as the Mean Square Error (MSE) loss $\mathcal{L}_{mse}\left(\bar{\mathbf{Y}}_i^{\mathbb{S}}, \tilde{\mathbf{Y}}_i^{\mathbb{S}}\right)$, between $\bar{\mathbf{Y}}_i^{\mathbb{S}}$ and $\tilde{\mathbf{Y}}_i^{\mathbb{S}}$. Thus, the training objective $\mathcal{L}_i^{\mathbb{S}}$ in this setting is written as

$$\mathcal{L}_i^{\mathbb{S}} = \mathcal{L}_{ce}\left(\hat{\mathbf{Y}}_i^{\mathbb{S}}, \mathbf{Y}_i^{\mathbb{S}}\right) + \alpha\mathcal{L}_{mse}\left(\bar{\mathbf{Y}}_i^{\mathbb{S}}, \tilde{\mathbf{Y}}_i^{\mathbb{S}}\right) = -\sum_{j=1}^{J} \mathbf{Y}_{i,j}^{\mathbb{S}} \log\left(\hat{\mathbf{Y}}_{i,j}^{\mathbb{S}}\right) + \alpha \sum_{j=1}^{J} \left|\bar{\mathbf{Y}}_{i,j}^{\mathbb{S}} - \tilde{\mathbf{Y}}_{i,j}^{\mathbb{S}}\right|^2, \tag{2}$$

where J indicates the number of classes, namely, $\mathbf{Y}_{i,j}^{\mathbb{S}}$, $\hat{\mathbf{Y}}_{i,j}^{\mathbb{S}}$, $\bar{\mathbf{Y}}_{i,j}^{\mathbb{S}}$, and $\tilde{\mathbf{Y}}_{i,j}^{\mathbb{S}}$, respectively, indicate the j-th channel of $\mathbf{Y}_i^{\mathbb{S}}$, $\hat{\mathbf{Y}}_i^{\mathbb{S}}$, $\bar{\mathbf{Y}}_i^{\mathbb{S}}$, and $\tilde{\mathbf{Y}}_i^{\mathbb{S}}$. α is a trade-off hyperparameter whose value progressively increases from 0 to 0.1, guided by a time-dependent function $\alpha(t) = 0.1\exp\left[-5\left(1-\frac{t}{T}\right)^2\right]$. Here, t represents the current training epoch, and T denotes the maximum training epoch. Besides, a strong augmentation method, CutMix[28], is introduced during training to enhance the performance further.

Learning from noisy annotations. A client C_i, such as C_3 in Fig. 2, may have a dataset $\mathbb{N}_i = \left\{\mathbf{X}_i^{\mathbb{N}}, \mathbf{Y}_i^{\mathbb{N}}\right\}$ with noisy annotations. To train a local model from this noisy dataset, we employ the noise-robust Dice (NR-Dice) loss $\mathcal{L}_{nd}$[13], a generalization of Dice loss and Mean Absolute Error (MAE) loss. Assuming the probability maps for images $\mathbf{X}_i^{\mathbb{N}}$ output by the local model are denoted by $\hat{\mathbf{Y}}_i^{\mathbb{N}}$, we optimize the training objective $\mathcal{L}_i^{\mathbb{N}}$ of local model by combining the CE loss $\mathcal{L}_{ce}$ and the NR-Dice loss $\mathcal{L}_{nd}$, defined as

$$\mathcal{L}_i^{\mathbb{N}} = \mathcal{L}_{ce}\left(\hat{\mathbf{Y}}_i^{\mathbb{N}}, \mathbf{Y}_i^{\mathbb{N}}\right) + \beta\mathcal{L}_{nd}\left(\hat{\mathbf{Y}}_i^{\mathbb{N}}, \mathbf{Y}_i^{\mathbb{N}}\right) = -\sum_{j=1}^{J} \mathbf{Y}_{i,j}^{\mathbb{N}} \log\left(\hat{\mathbf{Y}}_{i,j}^{\mathbb{N}}\right) + \beta \sum_{j=1}^{J} \frac{\left|\mathbf{Y}_{i,j}^{\mathbb{N}} - \hat{\mathbf{Y}}_{i,j}^{\mathbb{N}}\right|^{\gamma}}{\left(\mathbf{Y}_{i,j}^{\mathbb{N}}\right)^2 + \left(\hat{\mathbf{Y}}_{i,j}^{\mathbb{N}}\right)^2 + \epsilon}, \tag{3}$$

in which J is the number of classes, and $\mathbf{Y}_{i,j}^{\mathbb{N}}$ and $\hat{\mathbf{Y}}_{i,j}^{\mathbb{N}}$ represent the j-th channel of the noisy labels $\mathbf{Y}_{i,j}^{\mathbb{N}}$ and probability maps $\hat{\mathbf{Y}}_{i,j}^{\mathbb{N}}$. β and $\gamma \in [1,2]$ are hyperparameters, and ϵ is a small constant to avoid zero-division. We set γ to 1.5, and ϵ to 10^{-5}. We gradually decrease the value of β from 0.1 to 0 with a time-based function $\beta(t) = 0.1\left\{1-\exp\left[-5\left(1-\frac{t}{T}\right)^2\right]\right\}$, where t indicates the current training epoch and T is the maximum training epoch.

Learning from scribble annotations. For a client C_i like C_4 in Fig. 2 that possess a dataset $\mathbb{W}_i = \left\{\mathbf{X}_i^{\mathbb{W}}, \mathbf{Y}_i^{\mathbb{W}}\right\}$, in which scribble annotations provide much fewer supervision signals than the ideal annotations, we adopt the partial cross-entropy (pCE)[16] for training, similar to the work[10] for weakly-supervised instrument segmentation. Let the probability maps yielded by the local model for images $\mathbf{X}_i^{\mathbb{W}}$ be $\hat{\mathbf{Y}}_i^{\mathbb{W}}$. We calculate the training objective $\mathcal{L}_i^{\mathbb{W}}$ as the pCE loss $\mathcal{L}_{pce}\left(\hat{\mathbf{Y}}_i^{\mathbb{W}}, \mathbf{Y}_i^{\mathbb{W}}\right)$ between $\hat{\mathbf{Y}}_i^{\mathbb{W}}$ and $\mathbf{Y}_i^{\mathbb{W}}$, written as

$$\mathcal{L}_i^{\mathbb{W}} = \mathcal{L}_{pce}\left(\hat{\mathbf{Y}}_i^{\mathbb{W}}, \mathbf{Y}_i^{\mathbb{W}}\right) = -\sum_{j=1}^{J} \sum_{h=1}^{H} \mathbf{Y}_{i,j,h}^{\mathbb{W}} \log\left(\hat{\mathbf{Y}}_{i,j,h}^{\mathbb{W}}\right), \tag{4}$$

where J is the number of classes, and H denotes the annotated pixel numbers. $\hat{\mathbf{Y}}_{i,j,h}^{\mathbb{W}}$ and $\mathbf{Y}_{i,j,h}^{\mathbb{W}}$ denote the probability value and label value of the h-th pixel in the j-th channel of probability maps $\hat{\mathbf{Y}}_i^{\mathbb{W}}$ and scribble annotations $\mathbf{Y}_i^{\mathbb{W}}$.

TABLE 1 Detailed information on dataset setup for local clients. $\mathcal{I}$: ideal annotation type. $\mathcal{S}$: scarce annotation type. $\mathcal{N}$: noisy annotation type. $\mathcal{W}$: scribble annotation type. ξ: labeled sample ratio. δ: noisy label ratio.

client	annotation type	sequence	training sample	testing sample	notation
C_1	$\mathcal{I}$	$\{1,2\}$	first 225 frames of $\{1,2\}$	last 75 frames of $\{1,2,3,4,5,6,7,8\}$	-
C_2	$\mathcal{S}$	$\{3,4\}$	first 225 frames of $\{3,4\}$		$\xi = \{0.2, 0.4\}$
C_3	$\mathcal{N}$	$\{5,6\}$	first 225 frames of $\{5,6\}$		$\delta = \{0.2, 0.4, 0.6, 0.8\}$
C_4	$\mathcal{W}$	$\{7,8\}$	first 225 frames of $\{7,8\}$		-

3.3 | Federated learning module

Following the standard FL paradigm, we involve a central server Q for model aggregation and updating. Concretely, at each global round r, local clients $\{C_i\}_{i=1}^{K}$ upload local model weights $\{\mathbf{w}_i^l\}_{i=1}^{K}$ to the central server Q after local training, then the central server Q aggregates local weights to update the global model weight $\mathbf{w}^g$. Afterward, local clients download the global model weights, assign them to their local models, and then fine-tune them with their local datasets. By repeating this procedure until convergence, we can develop a generalizable global model $g(\cdot, \mathbf{w}^g)$. In our study, we adopt the FedAvg scheme[22] to update $\mathbf{w}^g$, written as

$$\mathbf{w}_{r+1}^g = \sum_{i=1}^{K} \frac{U_i}{\sum_{i=1}^{K} U_i} \mathbf{w}_{i,r}^l, \tag{5}$$

where U_i denotes the dataset size of client C_i.

4 | EXPERIMENTS AND RESULTS

4.1 | Experimental setup

Dataset and metric. We validated our method on the publicly available endoscopic dataset EndoVis17, provided by the 2017 robotic instrument segmentation challenge[1]. The EndoVis17 dataset consists of 10 sequences from abdominal porcine procedures along with the corresponding ground truth for binary, multi-class, and multi-part segmentation tasks. All images are in a 1920×1080 pixel resolution. Our study focused on the binary segmentation task. We utilized the former 8 sequences, denoted as $\{1,2,3,4,5,6,7,8\}$. As suggested by the work[1], we used the first 225 frames from each sequence for training and the remaining 75 frames for testing. We employed the Intersection Over Union (IoU) as the evaluation metric.

Problem simulation. For the FL framework, we established a central server Q and four local clients $\{C_i\}_{i=1}^{4}$. Concretely, we allocated the first 225 frames of sequences $\{1,2\}$, $\{3,4\}$, $\{5,6\}$, and $\{7,8\}$ to C_1, C_2, C_3, and C_4, respectively. We constructed the test dataset using the last 75 frames of sequences $\{1,2,3,4,5,6,7,8\}$. To mimic scenarios with scarce annotations[§], we treated 20% and 40% of the data as labeled data, leaving the remaining 80% and 60% as unlabeled data for C_2. To generate noisy annotations[¶] for C_3, we utilized dilation and erosion operations to create noisy annotations for 20%, 40%, 60%, and 80% of the samples. The operations were carried out with a 5×5 all-ones matrix kernel and a variable number of iterations, denoted as $\tau \in [15, 20]$. For C_4, we used the skeletonization method[14] on the ground truth to obtain scribble annotations. Detailed information on data setup for local clients is described in Table 1. For the sake of simplicity, we represent the ideal, scarce, noisy, and scribble annotation types as $\mathcal{I}$, $\mathcal{S}$, $\mathcal{N}$, and $\mathcal{W}$.

Implementation details. We preprocessed the data by cropping all images from the pixel at the (320, 28) to achieve a pixel resolution of 1280×1024, as described in the approach[2]. Subsequently, we resized all images to 320×256 by a factor of 16. We adopted the U-Net model[18] as the backbone for training. We trained local models using the Adam optimizer with a learning rate of 10^{-4}. The batch size was set to 16. We ran the local training for 1 epoch and repeated the global round 200 times. We employed horizontal and vertical flips (with a probability of 0.5 to flip) as data augmentation strategies. Alongside our method, we also executed experiments for `standalone` (local training) and `centralization` (centralized training) for comparison. For `standalone`, we conducted local training with 100 epochs for local clients using their private datasets. For

[§] We define ξ as the ratio of labeled images. For instance, $\xi = 0.2$ signifies that 20% of the images are labeled, leaving the remaining 80% unlabeled.

[¶] δ is used to represent the ratio of noisy annotations. For example, $\delta = 0.2$ denotes that 20% of the images contain noisy annotations, while the other 80% have perfect annotations.

TABLE 2 Quantitative results in IoU (%) of `standalone` under various data settings. $\mathcal{I}$: ideal annotation type. $\mathcal{S}$: scarce annotation type. $\mathcal{N}$: noisy annotation type. $\mathcal{W}$: scribble annotation type. ξ: labeled sample ratio. δ: noisy label ratio.

method	C_1: $\mathcal{I}$	C_2: $\mathcal{I}$	C_2: $\mathcal{S}$ ($\xi = 0.2$)	C_2: $\mathcal{S}$ ($\xi = 0.4$)	C_3: $\mathcal{I}$	C_3: $\mathcal{N}$ ($\xi = 0.2$)	C_3: $\mathcal{N}$ ($\xi = 0.4$)	C_3: $\mathcal{N}$ ($\xi = 0.6$)	C_3: $\mathcal{N}$ ($\xi = 0.8$)	C_4: $\mathcal{I}$	C_4: $\mathcal{W}$
standalone	71.02 ± 2.29	79.40 ± 0.63	79.65 ± 0.45	79.29 ± 0.93	81.56 ± 0.26	78.45 ± 1.32	76.59 ± 1.69	69.56 ± 0.97	73.82 ± 1.82	73.87 ± 0.88	64.09 ± 0.40

TABLE 3 Quantitative results in IoU (%) of `centralization` and `ours` under various data settings. $\mathcal{I}$: ideal annotation type. $\mathcal{S}$: scarce annotation type. $\mathcal{N}$: noisy annotation type. $\mathcal{W}$: scribble annotation type. ξ: labeled sample ratio. δ: noisy label ratio.

method	C_1: $\mathcal{I}$ C_2: $\mathcal{I}$ C_3 : $\mathcal{I}$ C_4 : $\mathcal{I}$	C_1: $\mathcal{I}$ C_2: $\mathcal{S}$ ($\xi = 0.2$) C_3: $\mathcal{N}$ ($\delta = 0.2$) C_4 : $\mathcal{W}$	C_1: $\mathcal{I}$ C_2: $\mathcal{S}$ ($\xi = 0.2$) C_3: $\mathcal{N}$ ($\delta = 0.4$) C_4 : $\mathcal{W}$	C_1: $\mathcal{I}$ C_2: $\mathcal{S}$ ($\xi = 0.2$) C_3: $\mathcal{N}$ ($\delta = 0.6$) C_4 : $\mathcal{W}$	C_1: $\mathcal{I}$ C_2: $\mathcal{S}$ ($\xi = 0.2$) C_3: $\mathcal{N}$ ($\delta = 0.8$) C_4 : $\mathcal{W}$	C_1: $\mathcal{I}$ C_2: $\mathcal{S}$ ($\xi = 0.4$) C_3: $\mathcal{N}$ ($\delta = 0.2$) C_4 : $\mathcal{W}$	C_1: $\mathcal{I}$ C_2: $\mathcal{S}$ ($\xi = 0.4$) C_3: $\mathcal{N}$ ($\delta = 0.4$) C_4 : $\mathcal{W}$	C_1: $\mathcal{I}$ C_2: $\mathcal{S}$ ($\xi = 0.4$) C_3: $\mathcal{N}$ ($\delta = 0.6$) C_4 : $\mathcal{W}$	C_1: $\mathcal{I}$ C_2: $\mathcal{S}$ ($\xi = 0.4$) C_3: $\mathcal{N}$ ($\delta = 0.8$) C_4 : $\mathcal{W}$
centralization	**86.90 ± 0.79**	83.65 ± 0.86	82.79 ± 0.55	81.53 ± 0.58	80.28 ± 0.65	82.90 ± 1.27	84.08 ± 0.38	80.95 ± 1.34	81.20 ± 0.21
ours	86.37 ± 0.31	**84.86 ± 0.61**	**84.13 ± 0.12**	**83.87 ± 0.20**	**84.29 ± 0.35**	**84.99 ± 0.41**	**84.83 ± 0.18**	**84.19 ± 0.59**	**84.74 ± 0.82**

`centralization`, we pooled all local datasets, constructed individual data loaders for these datasets, and performed centralized training by iterating through these data loaders with 200 epochs. Each data loader with a specific annotation type corresponded to a specific scheme from Section 3.2. In our study, all experiments were conducted three times with different random seeds.

4.2 | Experimental results

Quantitative results. We first show the quantitative results of `standalone` for each client, as shown in Table 2. C_1 trains the local model using its dataset with ideal annotations and achieves an IoU score of 71.02%. For C_2, C_3, and C_4 holding datasets with imperfect annotations, the upper-bound results obtained with ideal annotations are also reported for reference. We can note that imperfect annotations can negatively affect the model performance compared to ideal annotations. C_2 makes use of its dataset with scarce annotations to train its local model and realizes IoU scores of 79.40%, 79.65%, and 79.29%, respectively, when using ideal annotations, scarce annotations with labeled data ratios of 0.2 and 0.4. C_3 brings IoU scores of 81.56%, 78.45%, 76.59%, 69.56%, and 73.82%, respectively, under the settings of training with ideal annotations, noisy label ratios of 0.2, 0.4, 0.6, and 0.8. C_4 learns its local model from scribble annotations and achieves an accuracy of 64.09% in IoU, while the upper-bound accuracy obtained by training with ideal annotations is 73.87% in IoU.

Table 3 shows the comparison between `ours` and `centralization` under various data settings. Firstly, we can observe that `ours` and `centralization` consistently improves `standalone` under various dataset settings, providing evidence that taking advantage of more data helps improve model accuracy and generalization ability.

Interestingly, when comparing `centralization` to `ours`, we note that `centralization` surpasses `ours` by about 0.53% in IoU under the setting where both train models with datasets with ideal annotations, but realizes much worse performance under the other settings where imperfect annotations are involved. The case of `centralization` outperforming `ours` using ideal annotations is as anticipated since we believe `centralization` should benefit from the consolidation of all available data, allowing the model to learn a more comprehensive representation of the data. However, the decrease in performance of `centralization` with various imperfect annotations presents an intriguing phenomenon.

We hypothesize that the inconsistency of annotation types and their corresponding learning strategies could be a potential explanation. In the centralized setup, the model might be overly influenced by the majority trend in the data, which could lead to a "rich get richer" effect, often ignoring or misrepresenting the minority class or outliers. This effect can boost performance when ideal annotations are used due to the high consistency of the data. However, with the introduction of various imperfect annotations, the data becomes inconsistent and diverse. Moreover, each type of imperfect annotation requires a different learning strategy, leading to inconsistent strategies for different data loaders. The single centralized model may not adapt well to the diverse learning strategies, thus struggling with inconsistent or conflicting information, which results in a decline in performance. Contrarily, due to its distributed spirit, `Ours` with FL might be more robust against such inconsistencies. In `Ours`, each local client learns a model based on its local data, and these local models are then combined in a global model. This process allows each client to focus on their specific data subset, accommodating the local characteristics and imperfections of the data. Besides, the aggregation of these models might provide a better balance between different learning strategies, leading to enhanced performance.

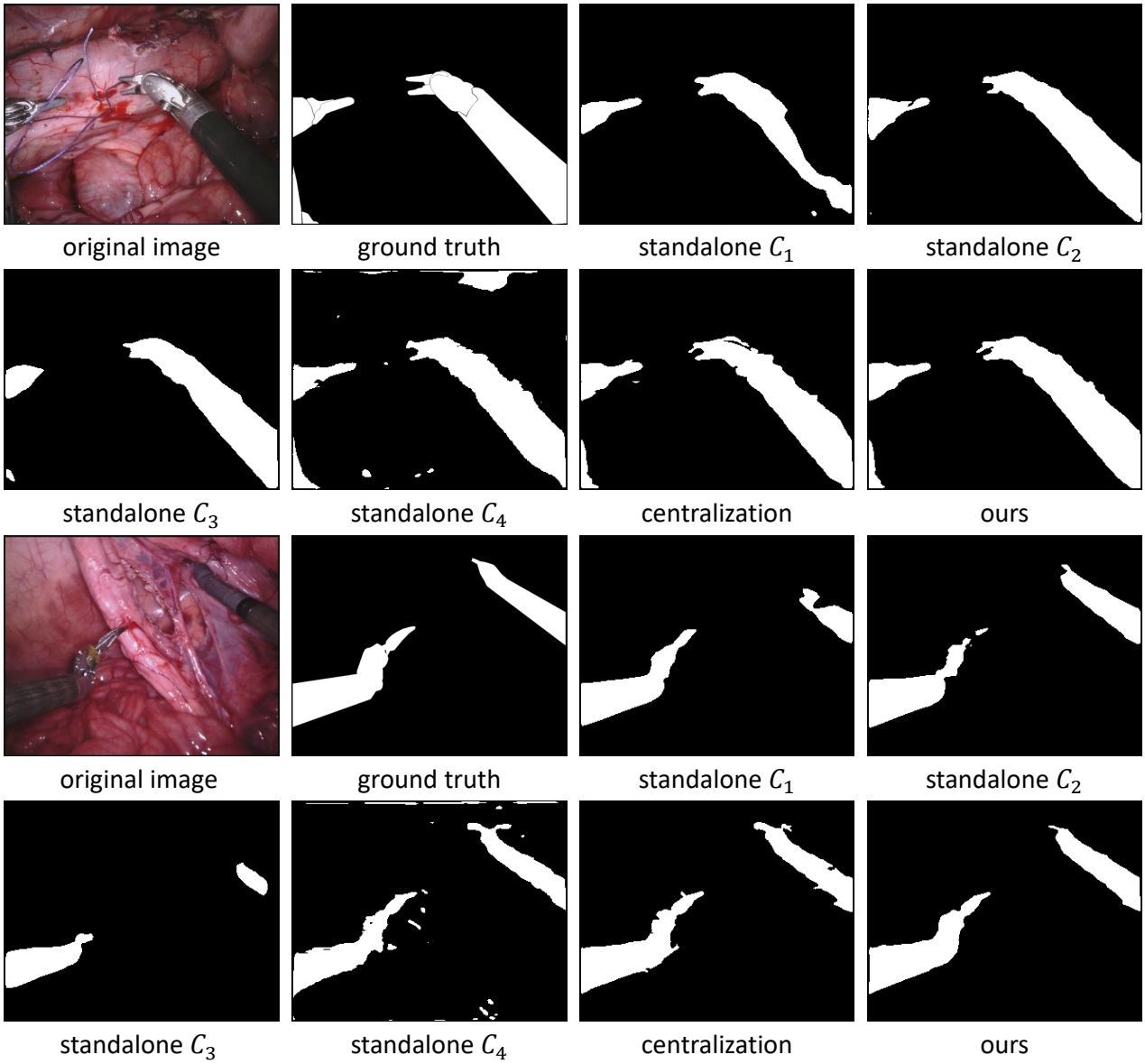
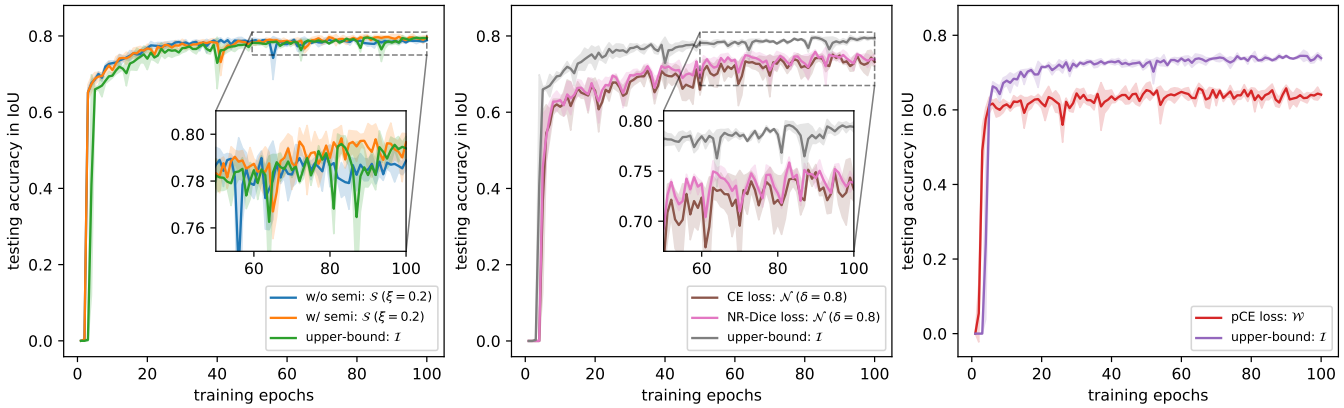


FIGURE 3 Qualitative visualization results. Cases are from the settings of C_1: $\mathcal{I}$, C_2: $\mathcal{S}$ ($\xi = 0.2$), C_3: $\mathcal{N}$ ($\delta = 0.8$), and C_4: $\mathcal{W}$. $\mathcal{I}$: ideal annotation type. $\mathcal{S}$: scarce annotation type. $\mathcal{N}$: noisy annotation type. $\mathcal{W}$: scribble annotation type. ξ: labeled sample ratio. δ: noisy label ratio.

Qualitative results. In addition to the quantitative results, we present a qualitative analysis of different methods, as shown in Fig. 3. We showcase two examples from the settings of C_1: $\mathcal{I}$, C_2: $\mathcal{S}$ ($\xi = 0.2$), C_3: $\mathcal{N}$ ($\delta = 0.8$), and C_4: $\mathcal{W}$. We can note that our method produces results considerably closer to the ground truth, even under this challenging data situation. It demonstrates that our method can effectively learn from distributed datasets with imperfect annotations and generate accurate predictions.

4.3 | Analytical studies

Effectiveness in handing various imperfect annotations. We first evaluated the efficacy of adopted methods, i.e., the MT framework [15], NR-Dice loss [13], and pCE loss [16] in handling imperfect annotations, since local clients should contribute effective local models for global model preparation. We show the cases of `standalone` for C_2, C_3, and C_4 in Fig. 4, where the testing



(a) standalone C_2 (b) standalone C_3 (c) standalone C_4

FIGURE 4 Ablation study on effectiveness of adopted methods of local clients in handling various imperfect annotations. $\mathcal{I}$: ideal annotation type. $\mathcal{S}$: scarce annotation type. $\mathcal{N}$: noisy annotation type. $\mathcal{W}$: scribble annotation type. ξ: labeled sample ratio. δ: noisy label ratio.

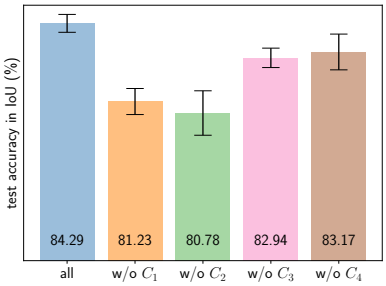


FIGURE 5 Ablation study on contribution of each client for global model preparation under the settings of C_1: $\mathcal{I}$, C_2: $\mathcal{S}$ ($\xi = 0.2$), C_3: $\mathcal{N}$ ($\delta = 0.8$), and C_4: $\mathcal{W}$.

TABLE 4 Ablation study on impact of various noise levels simulated with different ranges for dilation/erosion iteration τ and different values of noisy label ratio δ. $\mathcal{I}$: ideal annotation type. We chose $\tau \in [15, 20]$ for noisy label simulation in our study.

ranges	$\delta = 0.2$	$\delta = 0.4$	$\delta = 0.6$	$\delta = 0.8$
$\tau \in [1, 5]$	81.05 ± 0.87	81.00 ± 0.34	80.32 ± 1.05	80.40 ± 0.68
$\tau \in [5, 10]$	79.98 ± 1.21	78.72 ± 1.11	78.49 ± 0.43	78.36 ± 1.29
$\tau \in [10, 15]$	79.42 ± 1.26	77.70 ± 1.12	73.94 ± 0.85	74.82 ± 1.76
$\tau \in [15, 20]$	78.45 ± 1.32	76.59 ± 1.69	69.56 ± 0.97	73.82 ± 1.82
upper-bound: $\mathcal{I}$	81.56 ± 0.26			

accuracy in IoU of each training epoch is recorded. We can observe that the adopted methods present successful solutions. Specifically, the MT scheme improves the baseline when training with very scarce annotations ($\xi = 0.2$), as shown in Fig. 4(a). Besides, the NR-Dice loss shows better noise tolerance than the CE loss in a very high noise level setting ($\delta = 0.8$), as illustrated in Fig. 4(b). In addition, the pCE loss presents its potential solution in learning with scribble annotations, as depicted in Fig. 4(c). It achieves an accuracy of about 64.09% in IoU, which is about 9.78% lower than the upper-bound accuracy.

Contribution of each client. We then analyzed the contribution of each client for global model preparation under the settings of C_1: $\mathcal{I}$, C_2: $\mathcal{S}$ ($\xi = 0.2$), C_3: $\mathcal{N}$ ($\delta = 0.8$), and C_4: $\mathcal{W}$. As shown in Fig. 5, we can observe that combining all clients achieves the best performance while excluding any single client leads to accuracy degradation. The global model's performance decreases by approximately 3.06%, 3.51%, 1.35%, and 1.12% in IoU when excluding C_1, C_2, C_3, and C_4, respectively. This demonstrates that each client contributes to the overall performance.

Choice of noise level for noisy annotation simulation. We further studied the impact of various noise levels simulated with different ranges for dilation/erosion iteration τ and distinct values of noisy label ratio δ. Larger values for both τ and δ lead to an increased level of noise. As expected, our experimental results, shown in Table 4, demonstrate a decline in model accuracy in correlation with increasing noise levels. It is well-accepted that a small degree of dilation/erosion, e.g., with $\tau \in [1, 5]$, can better mimic the uncertainty and ambiguity often experienced during the annotation process, especially near the boundaries of objects. Interestingly, our experiments indicate that the model's performance remains relatively stable even with smaller iteration ranges, suggesting that the model is resilient to a certain degree of annotation noise in binary instrument segmentation

since it is a relatively easy task. However, in this study, we intentionally opted for a more extensive iteration range for our noisy simulations, i.e., $\tau \in [15, 20]$. We believe that this decision allows us to thoroughly evaluate and highlight the robustness of our method under more challenging conditions where the noise level is substantially increased, thus providing a more rigorous and convincing demonstration of its capabilities.

5 | DISCUSSION AND CONCLUSION

We aim to learn a generalizable model for instrument segmentation from decentralized surgical sequences with various imperfect annotations. In practice, surgical datasets are usually highly siloed with individual hospitals or medical institutions due to privacy concerns. Besides, these datasets often come with various imperfect annotations. Most existing methods ignore these realities and focus on centralized and well-annotated data, limiting their applicability. This problem relates to data availability and quality issues in the real world and calls for more robust, widely applicable models for robotic surgery.

Our method unifies the semi-, noisy label-, and scribble-supervised segmentation paradigms and the FL scheme into a single framework. Our method handles data privacy issues and effectively learns from imperfectly annotated data, accommodating diverse annotation scenarios. Our method outperformed `standalone` and `centralization` under various imperfect annotation settings, demonstrating its successful solution for this new and challenging problem. We posit that the variation in annotation types and their respective learning strategies might account for the diminished performance in centralized learning. However, this hypothesis warrants deeper exploration, including examining intermediate training patterns and offering a more detailed analysis of the relationship between different strategies and performance outcomes.

Despite its efficacy, our method has several limitations that warrant future investigation. Firstly, a gap exists between our method and its real-world implementations. One aspect of this gap stems from our use of simulated imperfect annotations, such as noisy annotations and scribble annotations. Although imperfect annotation simulation is widely applied in existing methods [12,29], authenticating our model with real-world imperfect annotations will further fortify the validity of our approach. Additionally, our method depended on a simplified assumption that each client possesses only a single type of annotation. Future research should address the challenges associated with each client having multiple annotation types to better reflect the intricacies of real-world scenarios, thus ensuring a deeper understanding and a more comprehensive solution. Secondly, our method relied on the conventional FedAvg scheme [22]. However, considering that each client possesses distinct datasets and diverse types of imperfect annotations, the performance of local models can exhibit significant variance. This raises the question of whether naive aggregation with FedAvg in our study would inadvertently impact the robustness and fairness of the global model [30,31]. Therefore, it is essential to investigate the effects of local model performance variance on the global model in our context and to ascertain whether more tailored aggregation strategies could yield improved performance. Thirdly, our method was only evaluated with binary instrument segmentation. To assess the generalization ability of our method, further validation with other instrument segmentation tasks and additional surgical datasets is also required.

ACKNOWLEDGMENTS

This work was supported by the JSPS KAKENHI Grant Numbers 21K19898 and 17H00867, the JST CREST Grant Number JPMJCR20D5, and the JST Moonshot R&D Grant Number JPMJMS2214-07, Japan.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

1. Allan M, Shvets A, Kurmann T, et al. 2017 Robotic Instrument Segmentation Challenge. *arXiv preprint arXiv:1902.06426.* 2019.
2. Shvets AA, Rakhlin A, Kalinin AA, Iglovikov VI. Automatic Instrument Segmentation in Robot-Assisted Surgery Using Deep Learning. In: 2018 17th IEEE International Conference on Machine Learning and Applications. IEEE. 2018:624-628
3. Iglovikov V, Shvets A. TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation. *arXiv preprint arXiv:1801.05746.* 2018.
4. Jin Y, Cheng K, Dou Q, Heng PA. Incorporating Temporal Prior from Motion Flow for Instrument Segmentation in Minimally Invasive Surgery Video. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer. 2019:11768,440–448.
5. Pakhomov D, Premachandran V, Allan M, Azizian M, Navab N. Deep Residual Learning for Instrument Segmentation in Robotic Surgery. In: Machine Learning in Medical Imaging. Springer. 2019:11861,566–573.
6. Maier-Hein L, Vedula SS, Speidel S, et al. Surgical Data Science for Next-Generation Interventions. *Nature Biomedical Engineering.* 2017;1(9):691–696.

7. Tajbakhsh N, Jeyaseelan L, Li Q, Chiang JN, Wu Z, Ding X. Embracing Imperfect Datasets: A Review of Deep Learning Solutions for Medical Image Segmentation. *Medical Image Analysis.* 2020;63:101693.
8. Zhao Z, Jin Y, Gao X, Dou Q, Heng PA. Learning Motion Flows for Semi-Supervised Instrument Segmentation from Robotic Surgical Video. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Springer. 2020:12263,679–689.
9. Yang H, Shan C, Bouwman A, Dekker LRC, Kolen AF, With dPHN. Medical Instrument Segmentation in 3D US by Hybrid Constrained Semi-Supervised Learning. *IEEE Journal of Biomedical and Health Informatics.* 2022;26(2):762-773.
10. Yang Z, Simon R, Linte C. A Weakly Supervised Learning Approach for Surgical Instrument Segmentation from Laparoscopic Video Sequences. In: Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling. SPIE. 2022:120341U.
11. Fuentes-Hurtado F, Kadkhodamohammadi A, Flouty E, Barbarisi S, Luengo I, Stoyanov D. EasyLabels: Weak Labels for Scene Segmentation in Laparoscopic Videos. *International Journal of Computer Assisted Radiology and Surgery.* 2019;14:1247–1257.
12. Xue C, Deng Q, Li X, Dou Q, Heng PA. Cascaded Robust Learning at Imperfect Labels for Chest X-ray Segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Springer. 2020:12266,579–588.
13. Wang G, Liu X, Li C, et al. A Noise-Robust Framework for Automatic Segmentation of COVID-19 Pneumonia Lesions from CT Images. *IEEE Transactions on Medical Imaging.* 2020;39(8):2653–2663.
14. Zhang TY, Suen CY. A Fast Parallel Algorithm for Thinning Digital Patterns. *Communications of the ACM.* 1984;27(3):236–239.
15. Tarvainen A, Valpola H. Mean Teachers are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results. *arXiv preprint arXiv:1703.01780.* 2018.
16. Tang M, Djelouah A, Perazzi F, Boykov Y, Schroers C. Normalized Cut Loss for Weakly-Supervised CNN Segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE. 2018:1818–1827.
17. Ross T, Reinke A, Full PM, et al. Robust Medical Instrument Segmentation Challenge 2019. *arXiv preprint arXiv:2003.10299.* 2020.
18. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer. 2015:9351,234–241.
19. Wang A, Islam M, Xu M, Ren H. Rethinking Surgical Instrument Segmentation: A Background Image Can Be All You Need. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. Springer. 2022:13437,355–364.
20. Yang L, Gu Y, Bian G, Liu Y. TMF-Net: A Transformer-Based Multiscale Fusion Network for Surgical Instrument Segmentation from Endoscopic Images. *IEEE Transactions on Instrumentation and Measurement.* 2022;72:1–15.
21. French G, Laine S, Aila T, Mackiewicz M, Finlayson G. Semi-Supervised Semantic Segmentation Needs Strong, Varied Perturbations. *arXiv preprint arXiv:1906.01916.* 2019.
22. McMahan B, Moore E, Ramage D, Hampson S, Arcas yBA. Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Artificial Intelligence and Statistics. PMLR. 2017:1273–1282.
23. Kairouz P, McMahan HB, Avent B, et al. Advances and Open Problems in Federated Learning. *Foundations and Trends® in Machine Learning.* 2021;14(1–2):1–210.
24. Zhang C, Xie Y, Bai H, Yu B, Li W, Gao Y. A Survey on Federated Learning. *Knowledge-Based Systems.* 2021;216:106775.
25. Li W, Milletarì F, Xu D, et al. Privacy-Preserving Federated Brain Tumour Segmentation. In: Machine Learning in Medical Imaging. Springer. 2019:11861,133–141.
26. Rieke N, Hancox J, Li W, et al. The Future of Digital Health with Federated Learning. *NPJ Digital Medicine.* 2020;3(1):119.
27. Shen C, Wang P, Roth HR, et al. Multi-Task Federated Learning for Heterogeneous Pancreas Segmentation. In: Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-Preserving Machine Learning. Springer. 2021:12969,101–110.
28. Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE. 2019:6023–6032.
29. Gao F, Hu M, Zhong ME, et al. Segmentation Only Uses Sparse Annotations: Unified Weakly and Semi-Supervised Learning in Medical Images. *Medical Image Analysis.* 2022;80:102515.
30. Li T, Hu S, Beirami A, Smith V. Ditto: Fair and Robust Federated Learning Through Personalization. In: International Conference on Machine Learning. PMLR. 2021:6357–6368.
31. Jiang M, Roth HR, Li W, et al. Fair Federated Medical Image Segmentation via Client Contribution Estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE. 2023:16302–16311.