Endo-SemiS: Towards Robust Semi-Supervised Image Segmentation for Endoscopic Video

Hao Li[*1] HAO.LI.1@VANDERBILT.EDU
Daiwei Lu[1] DAIWEI.LU@VANDERBILT.EDU
Xing Yao[1] XING.YAO@VANDERBILT.EDU
Nicholas Kavoussi[2] NICHOLAS.L.KAVOUSSI@VUMC.ORG
Ipek Oguz[1] IPEK.OGUZ@VANDERBILT.EDU
[1] *Vanderbilt University*
[2] *Vanderbilt University Medical Center*

Editors: Under Review for MIDL 2026

Abstract

In this paper, we present **Endo-SemiS**, a semi-supervised segmentation framework for providing reliable segmentation of endoscopic video frames with limited annotation. Endo-SemiS uses 4 strategies to improve performance by effectively utilizing all available data, particularly unlabeled data: (1) Cross-supervision between two individual networks that supervise each other; (2) Uncertainty-guided pseudo-labels from unlabeled data, which are generated by selecting high-confidence regions to improve their quality; (3) Joint pseudo-label supervision, which aggregates reliable pixels from the pseudo-labels of both networks to provide accurate supervision for unlabeled data; and (4) Mutual learning, where both networks learn from each other at the feature and image levels, reducing variance and guiding them toward a consistent solution. Additionally, a separate corrective network that utilizes spatiotemporal information from endoscopy video to improve segmentation performance. Endo-SemiS is evaluated on two clinical applications: kidney stone laser lithotomy from ureteroscopy and polyp screening from colonoscopy. Compared to state-of-the-art segmentation methods, Endo-SemiS substantially achieves superior results on both datasets with limited labeled data. The code is publicly available at https://github.com/MedICL-VU/Endo-SemiS

Keywords: Comprehensive supervision, uncertainty-guided pseudo-label, spatiotemporal

1. Introduction

Endoscopic image segmentation poses unique challenges, including large variations in image quality and appearance, which may be caused by motion blur, fluctuating lighting conditions, and often fluid-filled environments (Setia et al., 2023), as well as domain shifts (Ali et al., 2023). These effects are illustrated in Fig. 1, which shows blur, bleeding, debris, occlusions, and cross-site or cross-device appearance changes in ureteroscopy and colonoscopy images. The limited availability of manual labels further complicates the task.

Semi-supervised learning (SSL) approaches provide a potential solution by effectively leveraging information from unlabeled data (Sohn et al., 2020; Chen et al., 2021; Luo et al., 2022a,b; Yang et al., 2023; Tarvainen and Valpola, 2017; Wang et al., 2024). These methods construct supervision signals for unlabeled samples from the predictions of the model

* Corresponding author

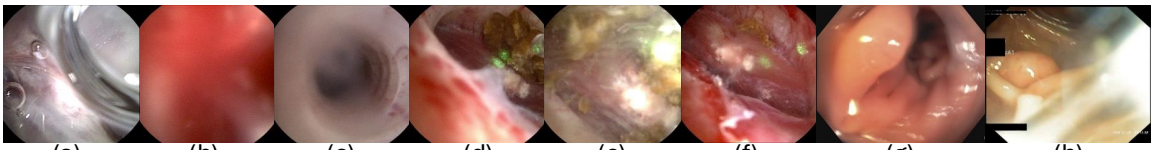

Figure 1: Challenging ureteroscopy (a–f) and colonoscopy (g–h) images for segmentation. (a) irrigation; (b) bleeding; (c) motion blur; (d) early ablation; (e) mid ablation; (f) late ablation. (g) and (h) are from the public dataset (Ali et al., 2023).

itself. A key approach to achieving this is enforcing consistency constraints (Tarvainen and Valpola, 2017), either through uncertainty-guided self-regularization (Sohn et al., 2020; Yang et al., 2023; Luo et al., 2022b; Wang et al., 2024; Tarvainen and Valpola, 2017) or cross-supervision (Chen et al., 2021; Luo et al., 2022a) to improve the quality and reliability of pseudo-labels.

Based on these principles, SSL can be broadly categorized into single-network and dual-network frameworks. Single-network approaches enforce consistency under perturbations and regularize pseudo-labels based on uncertainty. (Sohn et al., 2020; Yang et al., 2023; Wang et al., 2024). However, single model-based method tends to persist in its incorrect predictions, leading to error accumulation. Dual-network approaches maintain two networks that exchange pseudo-labels for cross-supervision (Chen et al., 2021; Luo et al., 2022a) to mitigate confirmation bias (Arazo et al., 2020). Building on this, numerous studies in medical imaging have achieved excellent segmentation performance (Luo et al., 2022a,b; Wang et al., 2023; Yu et al., 2019; Lei et al., 2022).

These existing SSL methods have some limitations: **(1)** Single-network methods lack model-level consistency, which makes them struggle with high-uncertainty samples. **(2)** Methods that either use the entire uncertainty map or apply a fixed uncertainty threshold treat many unreliable regions as confident, leading to false positives and overfitting to incorrect pseudo-labels. **(3)** Cross-supervision methods do not explicitly model uncertainty and struggle to filter out unreliable pseudo-labels. Since each model generates pseudo-labels independently, confirmation bias may occur when both models make similar wrong predictions.

In this paper, we propose **Endo-SemiS**, a semi-supervised segmentation method to address the limitations of existing approaches in endoscopic imaging with robust outcomes. Specifically, to address each of these limitations: **(1)** Endo-SemiS adopts a cross-supervision framework (see Fig. 2(a)) to prevent biased learning (Chen et al., 2022) and uses naive U-Net models to ensure real-time clinical applicability (Wei et al., 2021; Luo et al., 2019) rather than relying on transformer-based models that may require heavy computation (Luo et al., 2022a; Wang et al., 2024). **(2)** To obtain reliable pseudo-labels for unlabeled data, a critical step in SSL (Wu et al., 2021), we leverage both aleatoric and epistemic uncertainty (see Fig. 2(b)). Unlike existing fixed-threshold approaches (Sohn et al., 2020; Luo et al., 2022b), a dynamic thresholding mechanism is applied per uncertainty map, ensuring that only high-confidence regions contribute to pseudo-label supervision. **(3)** To achieve accurate and consistent supervision, we introduce a joint pseudo-labeling strategy as shown in Fig. 2(c),

where supervision is guided by the predictions in the lowest uncertainty regions identified by both networks, and pixels that are classified as uncertain are excluded. **(4)** We design multi-level mutual learning (see Fig. 2(d)) between networks to further mitigate confirmation bias and improve consistency between networks for producing reliable pseudo-labels. Our main contributions are:

- We propose an uncertainty-guided pseudo-labeling approach within a cross-supervision framework, which dynamically filters out unreliable regions for each image and provides more reliable segmentation supervision from unlabeled endoscopic frames.
- We introduce a consistency-focused learning framework with joint pseudo-label supervision and multi-level mutual learning. The more reliable prediction between the two networks is selected as supervision, while mutual learning reduces unnecessary prediction variance in confident regions and leads to more stable pseudo-labels.
- We design a plug-and-play correction model that uses spatiotemporal information from video to refine segmentation and can be easily integrated into other frameworks.

We validate Endo-SemiS on kidney stone laser lithotripsy as a challenging primary task and on polyp screening across different centers to demonstrate generalizability. Our comprehensive evaluation shows consistent improvements over state-of-the-art semi-supervised and fully supervised methods.

2. Methods

We begin with a semi-supervised segmentation dataset D, which consists of limited labeled data $\{x_l, y_l\}$ and a large amount of unlabeled data x_u, where x and y represent the input images and their annotations, respectively.

2.1. Preliminaries

Generic pseudo-label learning. The generic pseudo-label learning (Bellver Bueno et al., 2019) for a single network (referred to as Generic) first trains the model f, with forward pass $f(\cdot)$ on $\{x_l, y_l\}$ and applies it to x_u to generate $\tilde{y_u} = f(x_u)$, binarized as pseudo-labels, which can then be used as additional supervision. This can be described as

$$L = L_s + L_p \tag{1}$$

where L_s and L_p denote the supervised and pseudo-supervised loss for $\{x_l, y_l\}$ and $\{x_u, \tilde{y_u}\}$.

Cross-supervision. Endo-SemiS employs two individual U-Nets without sharing weights (Ronneberger et al., 2015) to achieve cross-supervision signals, as shown in Fig. 2(a). For a given input $x \in (x_l, x_u)$, the supervision can be simply extended from Generic (Eq. 1) as:

$$L_p(x) = L_p(f_1(x), \tilde{y_2}) + L_p(f_2(x), \tilde{y_1}) \tag{2}$$

where L_p represents the cross-supervision applied to a single network using the pseudo-label generated by the other model. The subscripts mark the corresponding processing of each model. Note, $f(x)$ denotes the raw logit map in this work. For brevity, we include it in the loss function term, as it can be converted to probabilities within the loss.

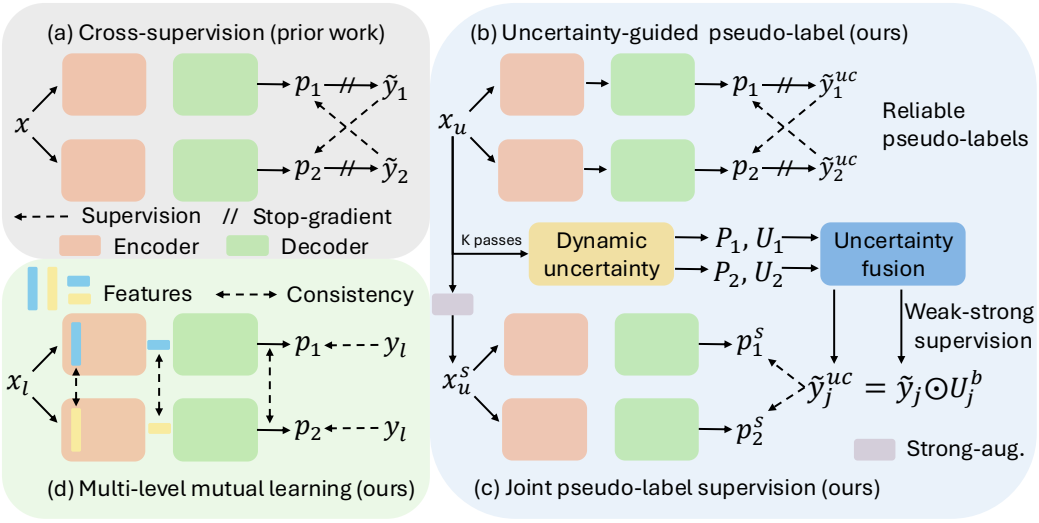

Figure 2: The proposed framework adapts the widely used cross-supervision baseline (a) with uncertainty-guided supervision to obtain reliable pseudo-labels (b–c), and further incorporates multi-level mutual learning (d) to improve cross-network consistency. Panels (b–c) (in blue) operate only on unlabeled data x_u, whereas (d) is applied only to labeled data x_l. The two networks share the same architecture but are optimized independently. y, $\tilde{y}$, and $\tilde{y}^{uc}$ denote the ground-truth mask, the raw pseudo-label, and the uncertainty-guided pseudo-label, respectively. $\odot$ denotes the Hadamard (element-wise) product, and U^b is the binary mask from uncertainty map U. x_u^s represents a strongly intensity-augmented version of x_u.

2.2. Uncertainty-guided pseudo-label

Uncertainty is introduced into the framework to mitigate confirmation bias (Fig. 2(b)). *We hypothesize that uncertainty estimates allow us to identify unreliable pseudo-label regions and exclude them from supervision, so that training focuses on reliable areas.*

Aleatoric uncertainty. We adopt the widely used weak-to-strong augmentation strategy (Sohn et al., 2020). Each unlabeled image x_u first undergoes geometric augmentations, referred to as weak augmentation, and x_u is further modified using intensity-based augmentations to obtain a strongly augmented image x_u^s. The corresponding pseudo-label $\tilde{y_u}$ is used to supervise the prediction from x_u^s. We also leverage CutMix (Yun et al., 2019) augmentation on x_u and x_u^s to further increase the robustness and segmentation performance.

Epistemic uncertainty. The use of cross-supervision allows us to strategically place Monte Carlo dropout (Kendall and Gal, 2017) layers after each convolutional layer in the decoder to introduce randomness into our framework and further improve segmentation performance (Yu et al., 2019). Specifically, as shown in Fig. 3(a), each unlabeled data x_u is passed through the network multiple times to estimate entropy-based uncertainty. The final output probability map is computed as $P = \frac{1}{K}\sum_{k=1}^{K} p_k$, where p_k denotes probability map

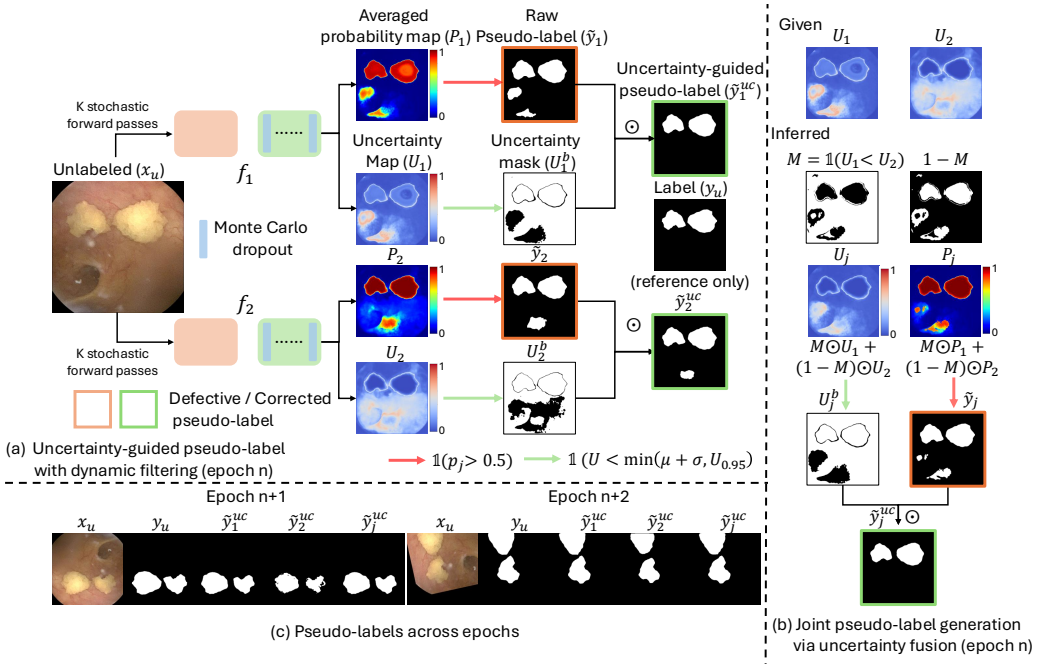

Figure 3: (a) For an unlabeled image x_u, uncertainty-guided pseudo-labels $\tilde{y}_1^{uc}$ and $\tilde{y}_2^{uc}$ (green boxes) are obtained by dynamically filtering the raw pseudo-labels $\tilde{y}_1$ and $\tilde{y}_2$, leading to cleaner supervision. The label y_u of the unlabeled image is shown for reference only. (b) M chooses the lower-uncertainty prediction at each pixel to obtain the joint pseudo-label $\tilde{y}_j^{uc}$ for more reliable supervision by correcting residual defects in $\tilde{y}_2^{uc}$ from (a). (c) Compared with the pseudo-labels at epoch n in (a), the $\tilde{y}_1^{uc}$, $\tilde{y}_2^{uc}$ and $\tilde{y}_j^{uc}$ at epochs $n+1$ and $n+2$ become cleaner and more consistent with y_u, indicating the effectiveness of (a) and (b).

in each forward pass $f(x_u)$ and we set $K = 5$. The entropy-based epistemic uncertainty (U) is derived as: $U = \frac{1}{K}\sum_{k=1}^{K} h(p_k)$, with $h(p_k) = -p_k \log p_k - (1-p_k)\log(1-p_k)$.

Dynamic filtering. Unlike previous works that use a fixed threshold (Sohn et al., 2020), the entire uncertainty map (Luo et al., 2022b) or quantile-based selection (Yu et al., 2019; Yang et al., 2023), we use a dynamic and data-driven thresholding strategy. Given the U, the threshold is set as $T = \min[\mu(U) + \sigma(U), U_{0.95}]$, where μ, σ and $U_{0.95}$ denote the mean, standard deviation and 95^{th} percentile, respectively. Our adaptive thresholding approach effectively handles long-tail distributions and noisy predictions, yielding a more reliable uncertainty-based binary mask $U_{x_u}^b = \{U(x_u) < T\}$ (see Fig. 3(a)). The final uncertainty-guided pseudo-label for x_u is then formulated as $\tilde{y}_u^{uc} = \tilde{y}_u \odot U_{x_u}^b$.

2.3. Joint pseudo-label supervision

Even with the incorporation of uncertainty estimates, the pseudo-labels may still be too noisy to provide appropriate supervision for harder samples. Most existing methods solely

rely on the $\tilde{y_u}$ from each network for supervision, which may not be sufficient. To address this, *our hypothesis is that joint supervision can effectively refine pseudo-labels by leveraging complementary information from both networks, providing more reliable supervision for challenging samples.*

As shown in Fig. 3(b), the joint pseudo-label $\tilde{y}_j^{uc}$ is constructed in three steps: (1) Given the uncertainty maps U_1 and U_2 from the two networks in Endo-SemiS, we create a binary mask $M = \mathbb{1}(U_1 < U_2)$ that selects the more confident prediction at each pixel, where $\mathbb{1}$ denotes the indicator function. (2) Using this mask, we form the joint probability $P_j = M \odot P_1 + (1 - M) \odot P_2$ and obtain the raw pseudo-label $\tilde{y}_j$ by thresholding P_j at 0.5, while the joint uncertainty map is defined as $U_j = M \odot U_1 + (1 - M) \odot U_2$. (3) Finally, we apply the dynamic filtering scheme to U_j to obtain the binary uncertainty mask U_j^b and compute the final uncertainty-guided joint pseudo-label as $\tilde{y}_j^{uc} = \tilde{y}_j \odot U_j^b$.

For an unlabeled image x_u and its strongly augmented version x_u^s, we extend the cross-supervision loss in Eq. 2 to a weak–strong setting, where pseudo-labels are generated from the weak augmented image (see Sec. 2.2) and used to supervise the strongly augmented image. The pseudo-supervised loss $L_p(x_u, x_u^s)$ is defined as:

$$L_p(x_u, x_u^s) = \underbrace{L_p\big(f_1(x_u), \tilde{y}_{u,2}^{uc}\big) + L_p\big(f_2(x_u), \tilde{y}_{u,1}^{uc}\big)}_{\text{uncertainty-guided cross-supervision}} + \underbrace{L_p\big(f_1(x_u^s), \tilde{y}_j^{uc}\big) + L_p\big(f_2(x_u^s), \tilde{y}_j^{uc}\big)}_{\text{joint pseudo-label supervision}} \quad (3)$$

2.4. Multi-level mutual learning

Individual networks may independently learn different representations, which can cause divergence and inconsistencies in their predictions. If one network is consistently wrong, it can bias the other network and propagate errors. We propose a multi-level mutual learning approach to mitigate this variability by aligning the learning trajectories of both models and promoting consistency in their predictions. Although it does not guarantee correctness on unlabeled data, it reduces randomness and stabilizes the learning process, making models less likely to reinforce extreme errors.

We use the labeled data to apply mutual learning between the two networks. This encourages similarity at both the encoders and the decoders. The consistency from encoder and bottleneck features helps align feature representations and reduce variability in learned embeddings. Unlike previous work, which enforces the similarity between the probability maps (Zhang et al., 2018), we enforce prediction consistency at the decoder level by aligning the logit maps of the networks, which is particularly important when generating pseudo-labels. Since pseudo-labels are filtered based on confidence thresholds, mutual learning stabilizes training by reducing prediction variance between networks, making the pseudo-label selection process more reliable.

For a labeled image x_l, let f_1^e, f_1^b, f_1 and f_2^e, f_2^b, f_2 denote the first encoder feature maps, bottleneck features, and logit maps of the two networks, respectively. The multi-level mutual learning loss is defined as:

$$L_m(x_l) = L_{\text{ssim}}\big(f_1^e, f_2^e\big) + \tfrac{1}{2}\big(L_{\text{KL}}(p \parallel q) + L_{\text{KL}}(q \parallel p)\big) + 2\, L_{\text{mse}}\big(f_1, f_2\big) \quad (4)$$

Total objective function. For labeled and unlabeled data, the total objectives are:

$$L(x_l) = L_s(x_l) + 0.5\, L_p(x_l) + 0.5\, L_m(x_l), \quad L(x_u) = 0.5\, L_p(x_u, x_u^s) \quad (5)$$

2.5. Spatiotemporal (ST) correction at frame level

Segmentations produced on semi-supervised frames may exhibit frame-level inconsistencies due to the lack of temporal information, which appear as isolated false positive (FP) or false negative (FN) frames. As a post-processing step, we leverage the inherent spatiotemporal information in video clips, and introduce a separate correction model at frame level to mitigate false positive FP and FN frames. We denote the n^{th} test frame by x_n and its predicted binary segmentation mask by $\tilde{y}_n$. For each frame x_n, we define R_n as the total number of foreground pixels in $\tilde{y}_n$. Our key assumption is that adjacent frames should not exhibit large discrepancies in R_n. For a given frame x_n, we enforce temporal consistency by correcting FP frames when $R_n > 0$ and $R_{n-1} = R_{n+1} = 0$. Similarly, we classify x_n as a FN frame when $R_n = 0$ and $R_{n-1} > r$ and $R_{n+1} > r$. We set $r = \frac{1}{4}HW$, where H and W denote the frame height and width.

To refine the predictions, we train a separate correction model that operates on a local temporal window. Given frames $x_{n-2}, \ldots, x_{n+2}$ and their corresponding predicted masks, we concatenate them along the channel dimension and use this as input to predict a refined segmentation for the central frame x_n. During training, random corruptions are introduced to the masks with basic morphological operations or by setting them to zero. We use the MSE loss to enforce spatiotemporal consistency; the total loss is:

$$L = L_s(x_n, y_n) + 0.25 \sum_{k\in\{-1,1\}} L_{\text{MSE}}(x_n, y_{n+k}) + 0.1 \sum_{k\in\{-2,2\}} L_{\text{MSE}}(x_n, y_{n+k}) \tag{6}$$

This formulation allows the network to leverage spatiotemporal information while preventing it from overly dominating the training process, thereby accommodating potential variations between frames. For inference, the model only takes adjacent frames as inputs.

3. Experiments

Kidney stone dataset. This in-house dataset (Deol et al., 2024) consists of 38 fiberoptic and 98 digital endoscopy videos. We extracted frames at 3 FPS, resulting in a total of 21,718 labeled frames. We partitioned the data at the video-level, yielding approximately a 75/5/20% split for training/validation/testing. While all videos contain kidney stones, some individual frames may not. This adds complexity to the segmentation task, as it also introduces an implicit detection challenge. The dataset exhibits large variation in image quality due to the complex in vivo environment during surgery (Fig. 1). The images are resized into 256×256.

Polyp colonoscopy dataset. PolypGen (Ali et al., 2023) is a publicly available multi-center dataset with 1,537 single-labeled frames (discrete sampling) and 2,225 sequence-labeled frames (short clips) collected from six different imaging centers. Following the benchmark study (Ali et al., 2023), we use data from centers 1–5 for training and test on center 6. We resize images to 512×512.

Implementation details. During training, we set the L_s as naive binary cross entropy loss with a batch size of 16 for 200 epochs. The initial learning rate is 10^{-4} with a cosine curve decay to 5^{-4}. Our study was conducted on an NVIDIA A6000.

Compared methods. We compare to several state-of-the-art semi-supervised segmentation methods, including Generic (Bellver Bueno et al., 2019), AllSpark (Wang et al., 2024),

Table 1: Kidney results ($mean \pm stdev.$, in %) with **10% labeled data**. Bold indicates the **best**. The horizontal sections show: supervised (gray), semi-supervised with single network (blue), cross-supervised (lavender), and supervised with 100% labeled data, i.e., upper bound (green). Our method achieved the highest Dice score, sensitivity, F1, and accuracy.

	Pixel-level			Image-level			
Methods	Dice	Sensitivity	Specificity	Pre.	Rec.	F1	Acc.
U-Net	80.5±32.1	88.6±22.0	95.4±8.4	88.7	95.3	92.8	90.1
nnU-Net	79.5±33.8	85.9±27.4	95.5±9.1	90.1	91.1	90.6	87.6
Generic	78.5±31.7	86.1±25.7	92.3±13.9	90.7	95.3	92.9	90.5
AllSpark	77.0±31.2	88.0±24.8	89.3±18.0	94.7	92.8	93.8	91.7
UPRC	80.7±31.4	84.0±27.3	96.4±7.8	92.9	94.6	93.7	91.6
FixMatch	81.9±31.7	89.8±22.4	94.3±10.9	89.7	**96.5**	93.0	90.5
UniMatch	85.5±27.6	89.4±23.2	95.5±8.9	94.3	96.4	95.4	91.7
Mean Teacher	82.2±31.2	84.1±28.6	96.6±8.5	95.6	90.5	93.0	91.1
CPS	85.2±28.0	88.8±22.8	95.8±8.8	94.0	96.1	95.0	93.4
Cross Teaching	85.6±28.7	87.6±26.5	**96.7±7.4**	**96.5**	92.6	94.8	92.9
Endo-SemiS (Ours)	**87.6±26.4**	**91.1±21.5**	96.0±8.4	95.0	96.1	**95.6**	**94.1**
Upper bound U-Net	85.3±29.2	89.0±24.5	96.5±8.2	94.4	94.2	94.3	92.5
Upper bound nnU-Net	85.5±28.5	89.3±24.5	96.0±8.6	92.4	93.3	92.9	90.5

UPRC (Luo et al., 2022b), FixMatch (Sohn et al., 2020), UniMatch (Yang et al., 2023), Mean Teacher (Tarvainen and Valpola, 2017), Cross-Pseudo Supervision (CPS) (Chen et al., 2021) and Cross Teaching (Luo et al., 2022a). Further details on the category classification of the compared methods are provided in Appendix A. We implemented these methods with their official code repositories. We evaluated performance with Dice, FN, and FP ratios at the pixel level, and precision, recall, F1-score, and accuracy at the image level.

Segmentation performance. The quantitative results of the kidney stone dataset using 10% labeled data are shown in Tab. 1. The Generic model underperforms compared to supervised learning, which highlights the critical role of pseudo-label quality in semi-supervised segmentation. In contrast, the results of Mean Teacher, UniMatch, and FixMatch show that incorporating external uncertainty improves segmentation, especially for UniMatch where epistemic uncertainty is also leveraged. The results of AllSpark indicate that transformer-based method struggles for kidney stone segmentation, where image quality is variable (Fig. 4). Cross-supervision methods (lavender) achieve better performance than single-network-based methods (blue), demonstrating better generalizability. Endo-SemiS achieves substantially superior performance across most metrics compared to these SOTA semi-supervised methods. Notably, it even outperforms supervised methods trained on full labeled data (green).

Consistency analysis. In Tab. 2, we present consistency results in two aspects: (1) robustness across different ratios of labeled training data, and (2) consistency between

Table 2: Dice (%) on kidney dataset with various labeled data ratios. "-1" and "-2" denote individual networks for cross-supervision. ST: spatiotemporal correction. Bold indicates the best in each category. Lavender denotes the cross-supervised methods.

Methods	1%	5%	10%	30%	100%
U-Net	74.9±34.1	77.8±34.5	80.5±32.1	82.0±32.0	85.3±29.2
nnU-Net	76.4±34.3	78.0±34.5	79.5±33.8	82.1±31.6	85.5±28.5
Generic	69.4±37.3	76.5±34.3	78.5±31.7	83.4±29.6	-
CPS-1	82.9±30.5	84.7±28.8	85.2±28.0	85.7±27.7	-
Cross Teaching-1	77.1±32.4	80.1±32.2	85.6±28.7	86.5±27.6	-
Ours-1	**86.5±27.6**	**87.5±26.4**	**87.6±26.4**	**87.9±26.1**	-
CPS-1+ST	83.8±29.5	85.3±28.1	85.7±27.4	86.2±27.1	-
Ours-1+ST	**87.1±27.1**	**87.8±26.3**	**88.1±25.7**	**88.2±25.8**	-
Performance variability in cross-supervised segmentation (±Dice in %)					
CPS-2	-1.0	+1.9	-0.7	+0.6	-
Cross Teaching-2	-11.4	-13.6	-4.0	-4.6	-
Ours-2	**-0.9**	**+0.1**	**0.0**	**-0.2**	-

models within the framework. Endo-SemiS maintains stable performance across different ratios, demonstrating particularly robust performance when labeled data is extremely limited (only 1%). The performance of the two cross-supervised models of our framework is more consistent and reliable than the compared methods. Considering the challenging visibility conditions in kidney stone surgery (Fig. 4), consistency is crucial to performance because inaccurate pseudo-labels can severely degrade segmentation results. Finally, we observe that our ST corrective model improves performance across all label ratios.

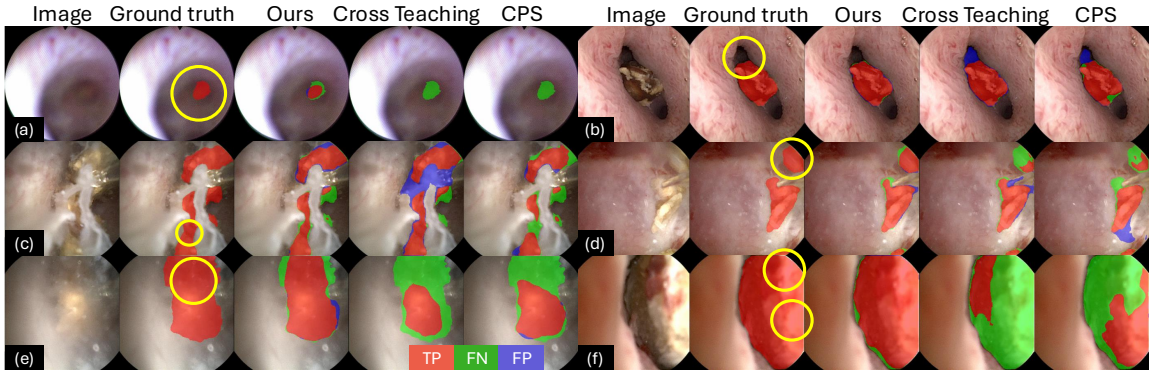

Figure 4: Qualitative kidney stone results (10% labeled data). Yellow circles highlight poor visibility areas. (a) fiberoptic frames, (b) digital frames, (c) fluid distortions, (d) motion blur, (e) debris during stone ablation, and (f) illumination changes. Enlarged, high-resolution views are shown in Fig. 5.

Table 3: Dice (%) for ablation study on kidney dataset with 10% labeled data. AU, EU: aleatoric/epistemic uncertainty. JPS: joint pseudo-label supervision. ML-D: mutual learning in decoder. ML-EB: mutual learning in encoder and bottleneck. Bold denotes the best.

	baseline (Chen et al., 2021)	+ AU	+ EU	+ JPS	+ ML-D	+ ML-EB
Endo-SemiS-1	85.2	86.2	86.9	**87.8**	87.2	87.6
Endo-SemiS-2	84.5	86.4	87.2	86.8	87.5	**87.6**

Table 4: Quantitative results (%) in polyp dataset with 10% labeled data. The left and right parts show the results for single and sequence frames, respectively.

	Single frame data			Sequence frame data		
Methods	Dice	Sensitivity	Specificity	Dice	Sensitivity	Specificity
U-Net	75±**30**	73±31	**100±1**	64±38	64±38	**100±1**
Endo-SemiS-1	76±34	75±34	**100±1**	69±39	67±39	**100±1**
Endo-SemiS-2	**79±30**	77±31	**100±1**	**71±37**	70±37	**100±2**
Upper bound U-Net	**79±30***	**79±31**	99±2	69±**37***	**74±35**	99±2

* denotes our implementation; benchmark (Ali et al., 2023) results are 79% and 66%.

Ablation analysis. Tab. 3 shows the ablation study, where CPS is used as the baseline method, and the improvements for each added component are shown. Importantly, joint pseudo-label supervision (JPS) yields a larger improvement, which indicates that it effectively removes uncertain regions and generates high-quality pseudo-labels for supervision, especially for strong augmented images. Although mutual learning slightly decreases the performance, it improves consistency.

Generalizability analysis. We also evaluate the proposed Endo-SemiS on the polyp segmentation task (Tab. 4). The results using only 10% labeled data show that Endo-SemiS outperforms supervised methods (U-Net) and reaches the upper bound (single) as well as surpasses it (sequence). Furthermore, the performance of the two models is consistent, showing the robustness of our approach to the domain shift between different imaging sites.

4. Conclusion

In this study, we propose **Endo-SemiS** for robust endoscopic segmentation using semi-supervised learning. It uses an uncertainty-guided pseudo-label strategy, cross- and joint-supervision, and mutual learning. We demonstrate the state-of-the-art performance of Endo-SemiS in experiments on two endoscopy datasets with varying image quality. The proposed spatiotemporal corrective model can further improve the segmentation performance.

Acknowledgments

This work was supported in part by the National Institutes of Health (R21DK133742) and Vanderbilt Institute for Surgery and Engineering (VISE) Seed Grant. Daiwei Lu is supported by NIH F31DK143735-01.

References

Sharib Ali, Debesh Jha, Noha Ghatwary, Stefano Realdon, Renato Cannizzaro, Osama E Salem, Dominique Lamarque, Christian Daul, Michael A Riegler, Kim V Anonsen, et al. A multi-centre polyp detection and segmentation dataset for generalisability assessment. *Scientific Data*, 10(1):75, 2023.

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020.

Míriam Bellver Bueno, Amaia Salvador Aguilera, Jordi Torres Viñals, and Xavier Giró Nieto. Budget-aware semi-supervised semantic and instance segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019*, pages 93–102, 2019.

Baixu Chen, Junguang Jiang, Ximei Wang, Pengfei Wan, Jianmin Wang, and Mingsheng Long. Debiased self-training for semi-supervised learning. *Advances in Neural Information Processing Systems*, 35:32424–32437, 2022.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2613–2622, 2021.

Ekamjit S Deol, Daiwei Lu, Ipek Oguz, and Nicholas L Kavoussi. Mp07-15 real-time kidney stone segmentation during distinct ureteroscopic tasks using a computer vision model. *Journal of Urology*, 211(5S):e110, 2024.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Tao Lei, Dong Zhang, Xiaogang Du, Xuan Wang, Yong Wan, and Asoke K Nandi. Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network. *IEEE transactions on medical imaging*, 42(5):1265–1277, 2022.

Huiyan Luo, Guoliang Xu, Chaofeng Li, Longjun He, Linna Luo, Zixian Wang, Bingzhong Jing, Yishu Deng, Ying Jin, Yin Li, et al. Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study. *The Lancet Oncology*, 20(12):1645–1654, 2019.

Xiangde Luo, Minhao Hu, Tao Song, Guotai Wang, and Shaoting Zhang. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In *International conference on medical imaging with deep learning*, pages 820–833. PMLR, 2022a.

Xiangde Luo, Guotai Wang, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Dimitris N Metaxas, and Shaoting Zhang. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Medical Image Analysis*, 80: 102517, 2022b.

Olaf Ronneberger, Philipp Fischer, and Thomas 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*, pages 234–241. Springer, 2015.

Shaan A Setia, Zachary A Stoebner, Chase Floyd, Daiwei Lu, Ipek Oguz, and Nicholas L Kavoussi. Computer vision enabled segmentation of kidney stones during ureteroscopy and laser lithotripsy. *Journal of Endourology*, 37(4):495–501, 2023.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Haonan Wang, Qixiang Zhang, Yi Li, and Xiaomeng Li. Allspark: Reborn labeled features from unlabeled in transformer for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3627–3636, 2024.

Jiacheng Wang, Hao Li, Han Liu, Dewei Hu, Daiwei Lu, Keejin Yoon, Kelsey Barter, Francesca Bagnato, and Ipek Oguz. Ssl2 self-supervised learning meets semi-supervised learning: multiple clerosis segmentation in 7t-mri from large-scale 3t-mri. In *Medical Imaging 2023: Image Processing*, volume 12464, pages 134–144. SPIE, 2023.

Jun Wei, Yiwen Hu, Ruimao Zhang, Zhen Li, S Kevin Zhou, and Shuguang Cui. Shallow attention network for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 699–708. Springer, 2021.

Yicheng Wu, Minfeng Xu, Zongyuan Ge, Jianfei Cai, and Lei Zhang. Semi-supervised left atrium segmentation with mutual consistency training. In *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part II 24*, pages 297–306. Springer, 2021.

Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7236–7246, 2023.

Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, proceedings, part II 22*, pages 605–613. Springer, 2019.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4320–4328, 2018.

Appendix A. Compared methods

We compare our method with several state-of-the-art semi-supervised approaches (Wang et al., 2024; Sohn et al., 2020; Yang et al., 2023; Tarvainen and Valpola, 2017; Luo et al., 2022b; Chen et al., 2021; Luo et al., 2022a). These methods cover both single-network (Wang et al., 2024; Sohn et al., 2020; Yang et al., 2023; Tarvainen and Valpola, 2017) and cross-supervision (Chen et al., 2021; Luo et al., 2022a) frameworks, with and without transformer backbones (Luo et al., 2022a; Wang et al., 2024). They focus on different uncertainty modeling strategies, including aleatoric (Wang et al., 2024; Sohn et al., 2020; Yang et al., 2023) and epistemic (Yang et al., 2023; Tarvainen and Valpola, 2017; Luo et al., 2022b) uncertainty, and combine confidence-based pseudo-labeling (Sohn et al., 2020; Yang et al., 2023; Chen et al., 2021; Luo et al., 2022a; Wang et al., 2024) with uncertainty-guided self-consistency (Tarvainen and Valpola, 2017; Luo et al., 2022b). For completeness, we summarize the main characteristics of each method below.

- **AllSpark** (Wang et al., 2024): Single-network transformer-based semi-supervised semantic segmentation method built on a standard pseudo-labeling baseline. It inserts an AllSpark bottleneck between the encoder and decoder, where channel-wise cross-attention and a class-wise semantic memory reconstruct labeled features from unlabeled features to strengthen supervision. It was published at *CVPR* 2024.

- **Uncertainty-Rectified Pyramid Consistency (URPC)** (Luo et al., 2022b): It is a single-network pyramid-prediction framework for semi-supervised medical image segmentation. The model produces multi-scale predictions and, for unlabeled data, enforces consistency between each scale and their average prediction. Pixel-wise uncertainty is estimated from the discrepancy among scales in a single forward pass and is used both to weight the pyramid consistency loss and to impose an uncertainty-minimization regularizer, enabling more reliable use of unlabeled images. It was published at *Medical Image Analysis* 2022.

- **FixMatch** (Sohn et al., 2020): Single-network method with a CNN backbone that combines consistency regularization and pseudo-labeling. For each unlabeled image, it takes the prediction on a weakly augmented view, keeps it as a qulified pseudo-label only if its confidence exceeds a fixed threshold, and trains the model to match this pseudo-label on a strongly augmented view of the same image. It was published at *NeurIPS* 2020.

- **UniMatch** (Yang et al., 2023): Single-network method with a CNN backbone that revisits FixMatch for semi-supervised semantic segmentation. It maintains weak-strong consistency using fixed confidence-thresholded pseudo-labels from the weakly augmented image, and introduces unified perturbations that operate at both the image level (strong augmentations) and the feature level (dropout), together with two strongly augmented images guided by the same weak prediction, to better exploit the perturbation space. It was published at *CVPR* 2023.

- **Mean Teacher** (Tarvainen and Valpola, 2017): Teacher-Student framework with a single concolutional neural network (CNN) backbone. The student is trained on

labeled data, and an exponential moving average (EMA) of the student weights defines the teacher. For unlabeled data, a consistency loss enforces that the student prediction matches the teacher prediction under stochastic perturbations. This can be viewed as reducing epistemic uncertainty. It was published at *NeurIPS* 2017.

- **Cross Pseudo Supervision (CPS)** (Chen et al., 2021): Cross-supervision semi-supervised semantic segmentation framework in which two segmentation networks with the same architecture but different initializations are trained jointly. For both labeled and unlabeled images, the prediction from each network is used as a pseudo label to supervise the other, enforcing prediction consistency and effectively expanding the training data. It was published at *CVPR* 2021.
- **Cross Teaching between CNN and Transformer (Cross Teaching)** (Luo et al., 2022a): Cross-supervision semi-supervised segmentation framework that pairs a CNN (UNet) and a Transformer (Swin-UNet). On unlabeled images, each network takes the prediction from the other network as a pseudo-label and is optimized with a cross-teaching Dice loss, providing implicit consistency while exploiting the complementary local and long-range representations of CNNs and transformers. It was published at *MIDL* 2022.

Appendix B. Qualitative results

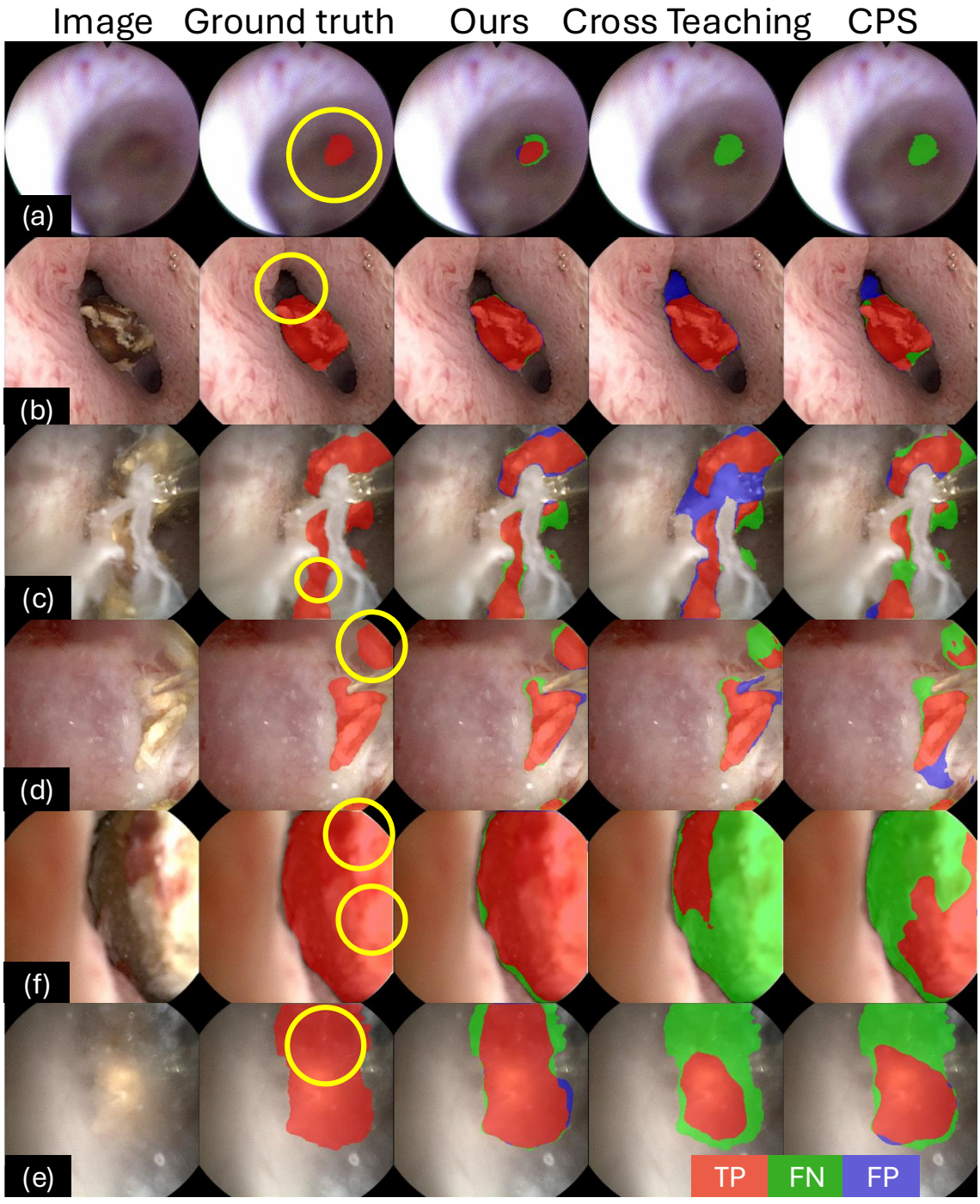

Figure 5: Qualitative kidney stone results (10% labeled data), larger version of Fig. 4 for better visualization. Yellow circles highlight poor visibility areas. (a) fiberoptic frames, (b) digital frames, (c) fluid distortions, (d) motion blur, (e) debris during stone ablation, and (f) illumination changes.