Learning Robust Medical Image Segmentation under Mixed-Quality Annotations

Minjae Jeong*, Seungjoo Lee*, Jaejin Lee, and Won Hwa Kim

Pohang University of Science and Technology (POSTECH), Pohang, South Korea
{minjaetidtid, seungjoo612, jlee4923, wonhwa}@postech.ac.kr

Abstract. Learning robust medical image segmentation models is challenging when training datasets contain mixed-quality annotations arising from inter-observer variability and ambiguous boundaries. Although existing methods address noisy supervision through label correction or supervision selection strategies, they rarely integrate both, which limits its effective noise handling. In response, we propose RGSS, **R**eliability-**G**uided **S**upervision for medical image **S**egmentation under annotations of varying qualities. RGSS jointly estimates annotation reliability via co-training and Gaussian Mixture Model (GMM), and adaptively reconstructs supervision targets using a pretrained diffusion-based mask refiner. The refinement strength is adjusted by the estimated noise severity, enabling aggressive correction for heavily corrupted masks while preserving structural details in mildly noisy cases. To mitigate confirmation bias and prior-induced errors, refined masks are further blended with peer-model predictions through epoch-dependent scheduling. Extensive experiments on two public medical segmentation benchmarks with four backbone networks demonstrate that our method consistently improves robustness and segmentation accuracy under varying levels of annotation noise, outperforming existing baselines.

Keywords: Noisy Label Learning · Annotation Refinement · Semantic Segmentation.

1 Introduction

Medical image segmentation is fundamental for computer-aided diagnosis and interpretation, enabling pixel-wise identification of anatomical structures and pathological regions such as tumors [4,12], lesions [18,19], and abnormalities [13, 31]. Recent deep neural networks (DNNs) have achieved strong performance in various segmentation tasks [28]; however, their success depends on large-scale datasets annotated with dense pixel-level masks provided by clinical experts [11]. In practice, even carefully curated datasets often contain imperfect annotations due to inter-observer variability and boundary ambiguity [27]. Consequently, medical segmentation datasets often exhibit mixed-quality supervision [23], posing a challenge for robust learning under imperfect annotations.

* These authors contributed equally to this work.

Related works. Existing noisy-label learning methods can be categorized into *i)* label correction and *ii)* supervision selection. Label correction methods directly refine annotations to improve structural consistency. For example, *SpatialCorrection* [30] models boundary distortions via a Markov process and relieves bias, while *diffusion-based refinement models* [26] iteratively denoise coarse masks through learned diffusion processes. Although these approaches restore geometric coherence, refinement is typically applied uniformly without accounting for varying annotation reliability. In contrast, supervision selection approaches identify unreliable annotations and reduce their influence on learning. While early works [2, 25, 29] largely focused on classification and relied on sample-level confidence estimation, more recent extensions to segmentation [10, 17] perform pixel-level noise detection and model optimization in a semi-supervised manner. Nevertheless, they operate primarily by discarding uncertain regions rather than reconstructing them, overlooking remaining salient structures in noisy regions.

Our proposal. In this regard, we propose a model-agnostic training scheme that iteratively selects low-reliability samples and explicitly refines imperfect annotations. Our method employs a co-training strategy with two segmentation networks. Per-sample losses computed by each network are reciprocally used by the peer model to fit a two-component distribution that separates clean and noisy subsets. While clean labels are used directly, noisy labels are corrected by a diffusion-based refinement strategy, adaptive to per-sample noise severity. Labels are then blended with peer model guidance to mitigate confirmation bias and potential undetected noise. As a result, the proposed framework constructs training targets by integrating co-training guidance and diffusion-based refinement, enabling robust performance under mixed-quality supervision.

Main contributions. Our contributions are summarized as follows: **1)** We propose a novel iterative framework for image segmentation under noisy annotations by refining target masks *without prior knowledge* of annotation reliability. **2)** We estimate per-sample reliability, which modulates a diffusion-based mask refiner that adaptively injects structural priors when generating reliable supervision for noisy samples in a *model-agnostic* manner. **3)** RGSS aggregates refined results and peer model guidance to construct a novel supervision with *robustness under mixed-quality supervision*. Extensive experiments on two public benchmarks with four different backbones demonstrate the robustness of our method under varying levels of annotation noise.

2 Preliminary: Diffusion-based Mask Refinement

Discrete diffusion models (DDMs) (Fig. 1a) [3, 7, 14] are designed for generation and learn forward corruption and reverse transition processes over discrete variables to approximate and sample from the data distribution. They are specifically formulated for sample generation rather than the refinement of existing input, making them less suitable for segmentation mask refinement tasks that require selective correction of erroneous regions while preserving reliable structures.

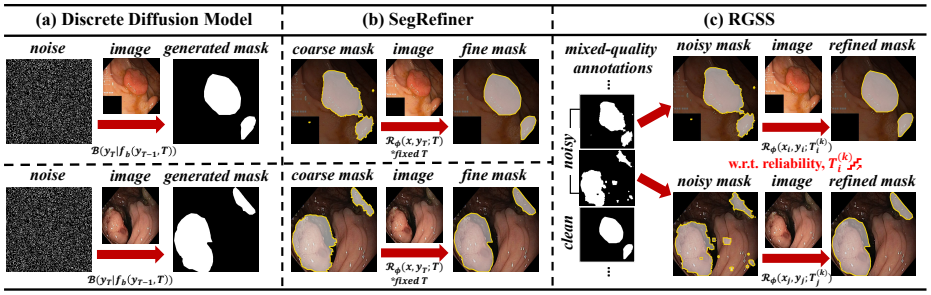

Fig. 1: (a) Mask generation using DDMs. (b) Mask Refinement with SegRefiner. (c) Adaptive Mask Refinement based on sample reliability and noise-intensity.

To address this limitation, SegRefiner [26] proposed a DDM application for mask refinement by introducing a *unidirectional* forward process over pixel-wise discrete states. For training, a fine–coarse mask pair (y_0, y_T) is constructed for input image x, where y_T is a structured corruption of y_0. For pixel (i, j), latent states $s_t^{i,j} \in \{[1, 0], [0, 1]\}$ corresponding to fine and coarse states, the intermediate mask is constructed as $y_t = f(s_t; y_0, y_T)$, where $f(\cdot)$ denotes a deterministic state-to-mask conversion. Likewise, forward and reverse processes become:

$$q(s_t^{i,j} \mid s_{t-1}^{i,j}) = s_{t-1}^{i,j} Q_t, \qquad p_\phi(s_{t-1}^{i,j} \mid s_t^{i,j}) = s_t^{i,j} P_{\phi,t}^{i,j}(x, y_t, t), \tag{1}$$

where Q_t is a state-transition matrix that enforces a unidirectional fine-to-coarse transition, and $P_\phi(x, y_t, t)$ is a reverse state-transition matrix whose entries are parameterized by learned predictions conditioned on x and y_t. At inference time, starting from a given coarse mask y_T, the refined mask is obtained by sequentially applying reverse diffusion over pixel-wise states for $t = T, \ldots, 1$ and constructing y_{t-1} from the resulting states, which is denoted as $\hat{y} = \mathcal{R}_\phi(x, y_T; T)$ as in Fig. 1b.

Our approach. Unlike SegRefiner, which applies the same refinement strategy to all input masks, our method (Fig. 1c) explicitly accounts for mixed-quality annotations. Furthermore, refinement strength is determined by sample-wise noise intensity for adaptive correction to preserve boundary-level details.

3 Method

We present an iterative plug-and-play learning framework for medical image segmentation under mixed-quality annotations, which proceeds as follows: **1)** Two segmentation networks are trained in parallel to estimate per-sample annotation reliability based on their training losses. **2)** To enhance robustness and mitigate confirmation bias, noisy labels are adaptively refined via a diffusion-based module with severity-aware strength, followed by peer-guided blending across all samples. **3)** The models are then optimized using the reconstructed targets in a mutual learning manner, and this procedure is iteratively repeated.

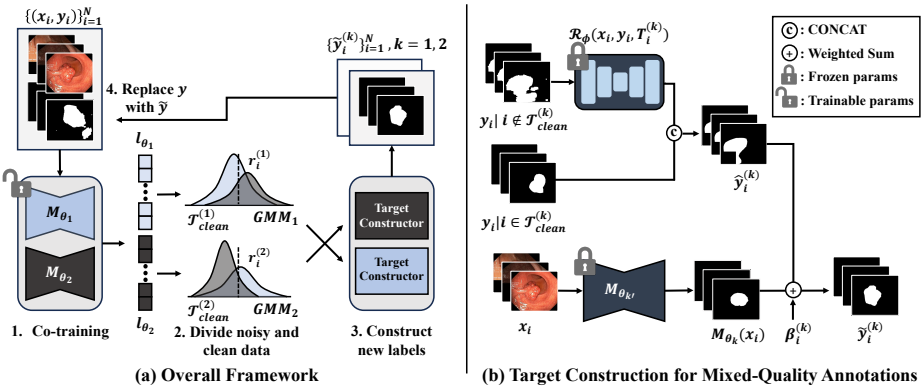

Fig. 2: (a) Overview of RGSS. Models M_{θ_1} and M_{θ_2} iteratively compute per-sample losses l_{θ_k} to estimate annotation reliability $r_i^{(k)}$ via GMM. For model M_{θ_k}, (b) refines noisy labels using $\mathcal{R}_\phi$, adjusted by $r_i^{(k)}$, to generate new supervision $\tilde{y}_i^{(k)}$, which is used to train each model in the subsequent iteration.

Problem Definition. Let $\{(x_i, y_i)\}_{i=1}^N$ be a set of image–annotation pairs, where $x_i \in \mathbb{R}^{H\times W\times C}$ denotes an image and $y_i \in \{0,1\}^{H\times W}$ denotes its corresponding pixel-level annotation. Segmentation masks y_i are provided by human annotators and are therefore prone to annotation noise. Presumably, the dataset consists of a combination of relatively accurate and corrupted annotations, while the reliability of each annotation is unknown. RGSS aims to learn a segmentation model $M_\theta(x) : \mathbb{R}^{H\times W\times C} \to [0,1]^{H\times W}$ that predicts accurate segmentation masks without prior knowledge of the dataset's annotation reliability.

Co-Estimation of Annotation Reliability. To handle corrupted annotations, we adopt a co-training framework that trains two segmentation networks in parallel, mitigating confirmation bias via complementary supervision. Reliability estimation leverages the memorization dynamics of DNNs [1,22,32], where clean samples are fitted earlier and incur lower losses. Accordingly, peer-network losses are used to infer annotation reliability.

Consider two segmentation models with different parameterizations, denoted as M_{θ_1} and M_{θ_2}. Before estimating annotation reliability, we train both models for e_0 epochs to ensure initial convergence on the given data. Formally, for model M_{θ_k}, the per-sample training loss for sample i is defined as $\ell_i^{(k)} = \mathcal{L}(\sigma(M_{\theta_k}(x_i)), y_i)$ where $\mathcal{L}(\cdot)$ denotes the Dice loss and $\sigma(\cdot)$ is the sigmoid activation. Both models are trained directly on the given $\{(x_i, y_i)\}_{i=1}^N$ without any label correction, despite the unknown risks of invalid annotations.

To separate clean and noisy samples in a soft and data-driven manner, we fit a two-component GMM [20] to the per-sample losses. For model M_{θ_k}, we model the empirical distribution of per-sample training losses $\{\ell_i^{(k)}\}_{i=1}^N$ as a mixture of

two clusters corresponding to clean and noisy annotations as

$$p(\ell^{(k)}) = \sum_{c \in \{\text{clean,noisy}\}} \pi_c^{(k)} \mathcal{N}\left(\ell^{(k)} \mid \mu_c^{(k)}, \sigma_c^{(k)}\right), \tag{2}$$

such that $\pi_c^{(k)}$ is the mixing coefficient per cluster and $\mu_c^{(k)}, \sigma_c^{(k)}$ is the mean and standard deviation of cluster c. The parameters in GMM are estimated using the Expectation–Maximization algorithm [9]. Accordingly, each sample is assigned a soft clean probability. To mitigate confirmation bias, the reliability weight used to supervise model M_{θ_k} is computed from the loss distribution of the other model. With such complementary exchange of knowledge, the reliability for sample i when training M_{θ_k} is defined as $r_i^{(k)} = p(c = \text{clean} \mid \ell_i^{(k')})$ for $k' \neq k$.

Target Construction for Mixed-Quality Annotations. Based on estimated clean probabilities, we construct training targets that directly address mixed-quality annotations. Given a threshold τ and reliability weights $r_i^{(k)}$, the clean label subset $\mathcal{J}_{clean}^{(k)} = \{i | r_i^{(k)} \geq \tau\}$ is approximated for model M_{θ_k}.

For samples $\{(x_i, y_i) \mid i \in \mathcal{J}_{clean}^{(k)}\}$, the original labels are retained, as their high $r_i^{(k)}$ indicate reliable supervision. In contrast, samples $\{(x_i, y_i) \mid i \notin \mathcal{J}_{clean}^{(k)}\}$, are regarded as unreliable and require adjustment. We therefore apply a refinement module $\mathcal{R}_\phi$ in Sec. 2 to construct new supervision targets. $\mathcal{R}_\phi$ is pre-trained on given masks y by learning the reverse state transitions $p_\phi(s_{t-1} \mid s_t)$ that invert the mask corruptions using forward diffusion process. In addition, the correction strength of $\mathcal{R}_\phi$ per sample is adaptively adjusted based on the estimated noise severity; we define the step number $T_i^{(k)}$ to be sample-dependent. Specifically, given a step range of $[T_{min}, T_{max}]$, T is defined as $T_i^{(k)} = T_{min} + (1 - r_i^{(k)})(T_{max} - T_{min})$. As a result, the refined annotation $\hat{y_i}^{(k)}$ is denoted as:

$$\hat{y_i}^{(k)} = \begin{cases} y_i, & i \in \mathcal{J}_{clean}^{(k)} \\ \mathcal{R}_\phi(x_i, y_i, T_i^{(k)}), & \text{otherwise} \end{cases} \tag{3}$$

Nonetheless, using $\hat{y}_i^{(k)}$ as the final hard target may induce systematic bias and propagate boundary inaccuracies. For $i \in \mathcal{J}_{clean}^{(k)}$, high $r_i^{(k)}$ does not ensure complete correctness, hence labels may contain subtle annotation errors that can adversely affect learning. Furthermore, while $\mathcal{R}_\phi$ demonstrates strong capability in restoring global structures [24, 26] even under noisy supervision, refinement may not fully resolve local boundary inconsistencies. To address such limitations, we construct soft supervision targets by blending $\hat{y}_i^{(k)}$ with peer prediction $M_{\theta_{k'}}(x_i)$ using a weighted interpolation, parametrized by $\beta_i^{(k)}$.

For clean samples, $\beta_i^{(k)}$ is defined as $r_i^{(k)}$ to ensure that supervision integrates $M_{\theta_{k'}}(x_i)$ based on sample reliability. Conversely, for refined masks, $\beta_i^{(k)}$ is defined as epoch-dependent scheduling $\alpha(e) = \frac{e_{max} - e}{e_{\max} - e_0}$ given maximum epoch e_{max}. This design prioritizes refined $\hat{y}_i^{(k)}$ during early training to suppress major noise as $\mathcal{R}_\phi$

injects a structural prior and progressively increases the contribution of $M_{\theta_{k'}}(x_i)$ as training stabilizes. Thus, the final supervision for model M_{θ_k}, $\tilde{y_i}^{(k)}$ becomes:

$$\tilde{y_i}^{(k)} = \beta_i^{(k)} \hat{y_i}^{(k)} + (1 - \beta_i^{(k)}) M_{\theta_{k'}}(x_i), \quad \beta_i^{(k)} = \begin{cases} r_i^{(k)}, & i \in \mathcal{J}_{clean}^{(k)} \\ \alpha(e), & \text{otherwise} \end{cases} \tag{4}$$

Resulting targets with enhanced supervision are used to update each model, enabling robust learning under mixed-quality annotations.

Training Objective. For $(e_{max} - e_0)$ epochs, the two segmentation models are trained with newly constructed targets $\{(x_i, \tilde{y_i})\}_{i=1}^N$, which are iteratively updated per epoch. For each model, the supervised loss is defined as $\mathcal{L}_{sup}^{(k)} = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(\sigma(M_{\theta_k}(x_i)), \tilde{y_i})$.

4 Experiment

4.1 Experimental Setup

Dataset. Evaluation is based on two public medical segmentation datasets: Kvasir-SEG [15] and ISIC 2017 [8], which respectively consist of endoscopic polyp images and dermoscopic skin lesion images with pixel-level annotations. These datasets contain 1000/2000 image-mask pairs, respectively. We partition Kvasir-SEG into an 8:1:1 ratio for training, validation, and testing, while ISIC 2017 is partitioned in advance, provided by the dataset providers.

Implementation. RGSS uses Adam optimizer with learning rate of 4e-3 for a batch size of 16. Hyperparameters $e_0/e_{max}/T_{min}/T_{max}/\tau$ are set as 20/200/1/7 /0.5 and R_ϕ pretraining follows the settings in [26]. To assess the generalizability of RGSS, we use 4 backbone models: U-Net [21], U-Net++ [33], LinkNet [5], and DeepLabv3 [6].

Baselines & Evaluation. We provide an extensive analysis of RGSS using other notable frameworks: SpatialCorrection(SC) [30], RMD [10] (Extension of [16] to segmentation), and SegRefiner [26]. For fair comparison, SegRefiner is applied iteratively per epoch. Performance is evaluated using mean Intersection-over-Union (mIoU) and mean Dice (mDice). To simulate noise, structured perturbations are added to ground-truth masks for each dataset using random morphological operations at ratios of 0/25/50%. Specifically, we iteratively apply region-wise dilation and erosion with randomly sampled elements (rectangular, circular, or elliptical) of size 3 or 5. Within randomly selected spatial windows, morphological transformations and stochastic pixel flips are performed to generate spatially coherent boundary distortions and shape variations. The process is repeated for up to 200 iterations and ends early once the results achieve an IoU of at least 0.5 with respect to the original mask to avoid extreme corruption that would break semantic integrity.

Table 1: Segmentation performance of RGSS and baseline methods.

Method	Kvasir-SEG [15]						ISIC 2017 [8]					
	0%		25%		50%		0%		25%		50%	
	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice
U-Net [21]	83.35	90.32	78.56	87.19	74.87	84.41	86.66	91.89	81.81	88.39	78.84	85.59
+SC [30]	80.01	88.24	77.34	86.50	76.40	85.81	82.97	89.77	80.79	87.86	79.31	86.21
+RMD [10]	83.49	90.74	80.47	88.91	75.10	84.93	86.74	91.91	83.91	89.83	79.77	87.60
+Segrefiner [26]	83.26	89.05	78.26	87.03	77.17	85.90	86.21	91.51	82.55	89.07	81.22	88.95
+RGSS(Ours)	**85.01**	**91.45**	**83.73**	**90.06**	**80.68**	**86.25**	**88.19**	**93.71**	**86.07**	**91.00**	**83.97**	**89.54**
U-Net++ [33]	84.17	90.76	79.57	87.94	73.16	83.35	87.33	92.30	82.21	88.77	79.60	86.43
+SC [30]	79.99	88.08	78.24	87.06	77.47	87.48	85.39	90.94	81.79	88.69	80.49	86.99
+RMD [10]	84.45	91.01	81.01	88.50	74.27	84.43	87.80	92.79	84.58	90.76	80.32	86.88
+Segrefiner [26]	83.87	90.11	79.64	88.20	78.00	87.56	87.17	92.49	82.30	88.79	81.69	87.94
+RGSS(Ours)	**85.97**	**91.66**	**84.11**	**90.54**	**81.37**	**88.43**	**89.79**	**94.91**	**87.06**	**92.00**	**84.42**	**90.53**
LinkNet [5]	85.58	91.60	78.97	86.94	75.58	85.41	86.69	91.87	81.73	88.44	77.29	85.87
+SC [30]	78.90	87.82	78.47	86.79	77.13	86.62	84.40	90.14	80.20	87.23	79.99	87.09
+RMD [10]	85.72	91.77	81.32	88.01	75.87	85.75	86.88	91.98	84.27	90.12	78.76	86.21
+Segrefiner [26]	84.61	90.52	79.33	87.51	78.34	86.81	86.09	91.44	82.26	88.64	81.81	88.04
+RGSS(Ours)	**86.40**	**91.94**	**84.04**	**90.20**	**82.01**	**88.99**	**88.37**	**93.85**	**86.71**	**91.84**	**84.61**	**90.63**
DeepLabv3+ [6]	83.87	90.65	79.71	87.93	76.41	85.79	87.19	92.25	82.44	88.61	78.92	85.57
+SC [30]	79.87	87.86	79.52	87.60	78.86	86.59	85.15	90.85	82.15	88.43	80.85	88.17
+RMD [10]	84.57	90.88	81.19	88.49	77.07	86.00	87.53	92.61	84.53	90.77	81.41	89.24
+Segrefiner [26]	83.62	90.44	80.64	88.52	79.20	87.39	87.02	92.02	83.48	89.22	82.57	89.75
+RGSS(Ours)	**85.70**	**91.20**	**83.95**	**90.14**	**81.17**	**88.61**	**89.36**	**94.27**	**87.50**	**92.36**	**85.05**	**91.06**

Table 2: Ablation study on the effect of co-training, diffusion step and $\beta_i^{(k)}$. The results were obtained from the Kvasir-SEG experiment for 50% noise setting.

Method			U-Net [21]		U-Net++ [33]	
Co-training	Diffusion step	$\beta_i^{(k)}$	mIoU	mDice	mIoU	mDice
×	×	×	74.87	84.41	73.16	83.35
✓	×	×	75.10	84.93	74.27	84.43
✓	T_{max}	✓	78.76	85.97	79.71	88.03
✓	$T_i^{(k)}$	×	78.15	85.63	79.43	87.89
✓	$T_i^{(k)}$	✓	**80.68**	**86.25**	**81.37**	**88.43**

4.2 Quantitative Results

Comparison with baselines. Tab. 1 shows that RGSS achieves superior performance across all baselines, highlighting strong generalizability and powerful model-agnostic properties. In particular, RGSS yields an average of +2.75% in mIoU and +1.30% in mDice under 50% noise settings compared to the second-best results. Notably, RGSS outperforms other baselines even on the original clean dataset, suggesting that it effectively models intrinsic annotation variability beyond synthetic perturbations. Hence, RGSS generates effective supervision that demonstrates strong robustness compared to other baseline models.

Ablation Study. Tab. 2 emphasizes the impact of co-training, adaptive diffusion, and blending by $\beta_i^{(k)}$. Accordingly, co-training yields average improvements of 0.67% and 0.80% in mIoU and mDice, respectively. The results also demonstrate that diffusion-based refinement significantly improves learning; in particular, rows 3 and 5 show that controlling diffusion step T via noise severity further enhances performance through noise-aware correction. Furthermore, supervision

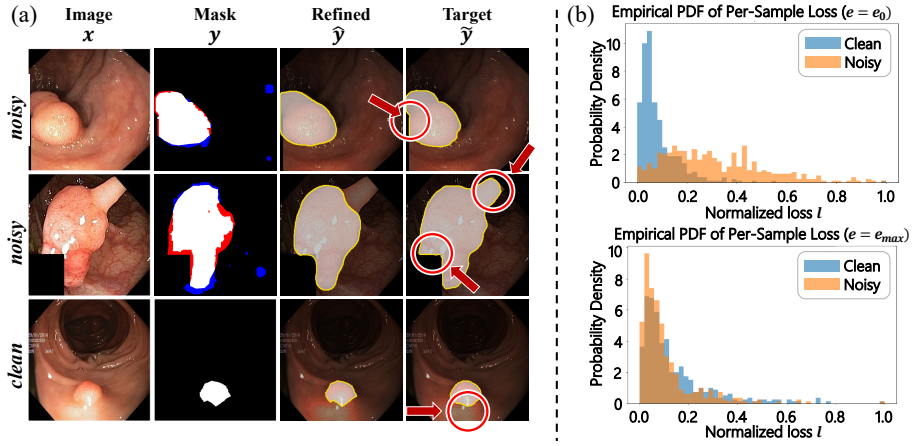

Fig. 3: (a) Visualization of generated labels on Kvasir-SEG. For mask y, RED indicates under-annotated regions, while BLUE indicates over-annotated regions (b) Loss distribution of clean and noisy samples before and after training.

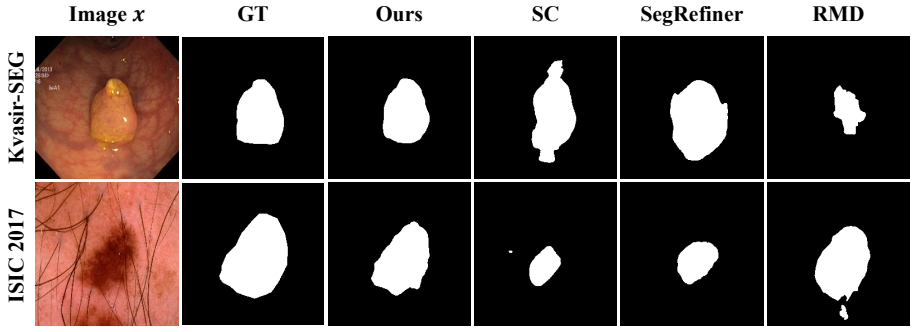

Fig. 4: Segmentation results of different models on Kvasir-SEG and ISIC 2017.

constructed via blending shows superior results, establishing the efficacy of our supervision in alleviating confirmation bias. (Row 4 and 5)

4.3 Qualitative Results

Visualization. Fig. 3a illustrates the efficacy of our framework. Mask y visualizes annotation inconsistencies: red indicates under-annotated regions, while blue indicates over-annotated regions. With $\mathcal{R}_\phi$, $\hat{y}$ eliminates major discrepancies in y, but at the cost of boundary-level details. Our final target $\tilde{y}$ integrates $M_{\theta_{k'}}(x_i)$ to restore fine details and subtle protrusions. Red circles indicate salient details of $\tilde{y}$. Notably, in row 2, $\tilde{y}$ captures a polyp protrusion omitted in y. This suggests that our framework reduces overfitting to noisy labels and promotes predictions aligned with image-level evidence. Such properties facilitate robust boundary perception, contributing to the superior results shown by Fig. 4.

Model Behavior. Fig. 3b depicts the empirical pdf of the per-sample losses for epochs e_0 and e_{max}. The plot demonstrates how the final target $\tilde{y}$ for the noisy subset ultimately aligns with the clean label distribution after e_{max} training.

5 Conclusion

We propose a novel iterative training framework for robust medical segmentation that explicitly accounts for mixed-quality annotations via supervision reconstruction. By introducing a diffusion-based refinement strategy to correct perturbed labels and adequate blending of complementary information via co-training to update training targets, our framework progressively realigns corrupted labels toward a unified clean-label distribution. Extensive evaluation on multiple datasets and backbones confirms the effectiveness of our method.

Acknowledgments. This research was supported by RS-2025-02216257 (50%), RS-2026-25494850 (45%), RS-2019-II191906 (AI Graduate Program at POSTECH, 5%).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arpit, D., et al.: A closer look at memorization in deep networks. In: International Conference on Machine Learning (ICML) (2017)
2. Ashraf, M., et al.: A loss-based patch label denoising method for improving whole-slide image analysis using a convolutional neural network. Scientific Reports **12**(1), 1392 (2022)
3. Austin, J., et al.: Structured denoising diffusion models in discrete state-spaces. Advances in Neural Information Processing Systems (NeurIPS) **34**, 17981–17993 (2021)
4. Biratu, E.S., et al.: A survey of brain tumor segmentation and classification algorithms. Journal of Imaging **7**(9), 179 (2021)
5. Chaurasia, A., Culurciello, E.: Linknet: Exploiting encoder representations for efficient semantic segmentation. In: IEEE Visual Communications and Image Processing (VCIP). pp. 1–4. IEEE (2017)
6. Chen, L.C., et al.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
7. Chen, T., et al.: Berdiff: Conditional Bernoulli diffusion model for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 491–501. Springer (2023)
8. Codella, N.C.F., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC) (2017)
9. Dempster, A.P., et al.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological) **39**(1), 1–22 (1977)

10. Fang, C., et al.: Reliable mutual distillation for medical image segmentation under imperfect annotations. IEEE Transactions on Medical Imaging **42**(6), 1720–1734 (2023)
11. Gao, Y., et al.: Medical image segmentation: A comprehensive review of deep learning-based methods. Tomography **11**(5), 52 (2025)
12. Havaei, M., et al.: Brain tumor segmentation with deep neural networks. Medical Image Analysis **35**, 18–31 (2017)
13. Homayoun, H., Ebrahimpour-Komleh, H.: Automated segmentation of abnormal tissues in medical images. Journal of Biomedical Physics & Engineering **11**(4), 415 (2021)
14. Jeong, M., Cho, H., Jung, S., Kim, W.H.: Uncertainty-aware diffusion-based adversarial attack for realistic colonoscopy image synthesis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 647–658. Springer (2024)
15. Jha, D., et al.: Kvasir-seg: A segmented polyp dataset. In: International Conference on Multimedia Modeling. pp. 451–462. Springer (2020)
16. Li, J., et al.: Dividemix: Learning with noisy labels as semi-supervised learning. International Conference on Learning Representations (ICLR) (2020)
17. Li, P., et al.: Semi-supervised semantic segmentation under label noise via diverse learning groups. pp. 1229–1238 (2023)
18. Malik, M., et al.: Stroke lesion segmentation and deep learning: A comprehensive review. Bioengineering **11**(1), 86 (2024)
19. Mirikharaji, Z., et al.: A survey on deep learning for skin lesion segmentation. Medical Image Analysis **88**, 102863 (2023)
20. Permuter, H., et al.: A study of Gaussian mixture models of color and texture features for image classification and segmentation. Pattern Recognition **39**(4), 695–706 (2006)
21. Ronneberger, O., et al.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 234–241. Springer (2015)
22. Shen, Y., Sanghavi, S.: Learning with bad training data via iterative trimmed loss minimization. In: International Conference on Machine Learning (ICML) (2018)
23. Shi, J., et al.: A survey of label-noise deep learning for medical image analysis. Medical Image Analysis **95**, 103166 (2024)
24. Shuai, X., et al.: A survey of multimodal-guided image editing with text-to-image diffusion models. arXiv preprint arXiv:2406.14555 (2024)
25. Sun, Z., et al.: PNP: Robust learning from noisy labels by probabilistic noise prediction. pp. 5311–5320 (2022)
26. Wang, M., et al.: SegRefiner: Towards model-agnostic segmentation refinement with discrete diffusion process. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
27. Wong, J., et al.: Effects of interobserver and interdisciplinary segmentation variabilities on CT-based radiomics for pancreatic cancer. Scientific Reports **11**(1), 16328 (2021)
28. Xu, Y., et al.: Advances in medical image segmentation: A comprehensive review of traditional, deep learning and hybrid approaches. Bioengineering **11**(10), 1034 (2024)
29. Xue, C., et al.: Robust medical image classification from noisy labeled data with global and local representation guided co-training. IEEE Transactions on Medical Imaging **41**(6), 1371–1382 (2022)

30. Yao, J., et al.: Learning to segment from noisy annotations: A spatial correction approach. In: International Conference on Learning Representations (ICLR) (2023)
31. Zacharaki, E.I., Bezerianos, A.: Abnormality segmentation in brain images via distributed estimation. IEEE Transactions on Information Technology in Biomedicine **16**(3), 330–338 (2011)
32. Zhang, C., et al.: Understanding deep learning requires rethinking generalization. International Conference on Learning Representations (ICLR) (2017)
33. Zhou, Z., et al.: Unet++: A nested U-net architecture for medical image segmentation. In: International workshop on Deep Learning in Medical Image Analysis (DLMIA). pp. 3–11. Springer (2018)