Pattern Recognition 183 (2027) 115009



Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr





Reliability-aware feedback learning for scribble-supervised 3D multi-organ segmentation☆

Xiaoxiang Han [a,1], Yiman Liu [b,1], Jiang Shang [a], Yunhao Bai [c], Yan Wang [c,*], Qi Zhang [a,*]

[a] The SMART (Smart Medicine and AI-based Radiology Technology) Lab, School of Communication and Information Engineering, Shanghai University, Shanghai, 200444, China
[b] Department of Pediatric Cardiology, Shanghai Children's Medical Center, Shanghai Jiao Tong University School of Medicine, Shanghai, 200127, China
[c] Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai, 200241, China

ARTICLE INFO

Keywords:
Weakly-supervised learning
Medical image segmentation
Scribble annotation
Mahalanobis distance

ABSTRACT

Sparse scribbles significantly reduce annotation costs for 3D multi-organ segmentation but often result in unstable feature learning and poor performance in unlabeled regions due to severe under-supervision. These unreliable regions reveal limitations in discriminative representation learning and remain a key obstacle to reliable segmentation. Therefore, we propose a reliability-aware feedback learning framework. Unlike existing methods that rely only on uncertainty estimation or pseudo-label refinement, it explicitly models feature-space reliability and uses reliability-based feedback conditions, enabling more stable representation learning in unreliable regions. Specifically, (1) we represent each class as a multivariate Gaussian in the embedding space and measure sample-to-class discrepancy using Mahalanobis distance. With triplet loss promoting intra-class compactness and inter-class separability, the resulting distance maps provide a reliability estimate. (2) A feature-perturbed decoder produces complementary predictions, which are fused via distance-based reliability masks to generate robust pseudo-labels. (3) The resulting low-reliability discrepancy map, together with the pseudo-labels, is fed back as an additional conditioning signal to focus learning on hard regions. Extensive experiments on two public 3D multi-organ benchmarks demonstrate consistent improvements over existing advanced methods in accuracy, cross-dataset generalization, and noise robustness.

1. Introduction

Computed Tomography (CT) imaging is a key diagnostic tool in modern medicine, in which precise organ segmentation is vital for diagnosis, treatment planning, and intervention. However, training deep learning models for this task typically requires large datasets with time-consuming and costly voxel-wise annotations. For 3D multi-organ CT data, annotating a single scan can take 3–4 h of expert clinician time. To reduce this burden, sparse annotation strategies such as scribbles drawn on a few slices have emerged as a practical alternative. Scribbles leverage clinician expertise during image review, taking only minutes per scan while capturing essential structural details. This work focuses on developing robust segmentation models trained with sparse scribble annotations, aiming for performance comparable to dense supervision at a fraction of the cost [1].

Since sparse scribble annotations cannot accurately represent regions of interest and provide only limited supervisory information, scribble-supervised learning primarily focuses on exploiting information from unlabeled pixels/voxels. Current approaches can be broadly categorized into four paradigms: (1) Pseudo-label generation methods [2–4] with filtering strategies [5], which often suffer from error accumulation and confirmation bias; (2) Loss function optimization techniques [6–8] that struggle to capture complex anatomical relationships; (3) Auxiliary task-based approaches [9,10] that may introduce task-specific biases; and (4) Consistency regularization methods [3,11–14] relying on various perturbations, which frequently fail to distinguish between meaningful variations and noise. While these methods have shown promise, they exhibit three fundamental limitations: First, most approaches lack explicit modeling of feature distribution boundaries, making them vulnerable to overconfidence in under-supervised regions. Second, existing uncertainty estimation techniques [15–17] typically focus on prediction-level uncertainty while ignoring feature-space discriminability. Third, current frameworks cannot dynamically adjust their learning focus based on the model's evolving reliability

☆ Code will be made available at https://github.com/GtLinyer/RAFLS.

* Corresponding authors.
E-mail addresses: hanxx@shu.edu.cn (X. Han), liuyiman@scmc.com.cn (Y. Liu), ywang@cee.ecnu.edu.cn (Y. Wang), zhangq@t.shu.edu.cn (Q. Zhang).

[1] These authors contributed equally to this work, and are co-first authors.

https://doi.org/10.1016/j.patcog.2026.115009
Received 8 July 2025; Received in revised form 14 April 2026; Accepted 21 September 2026
Available online 26 September 2026
0031-3203/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

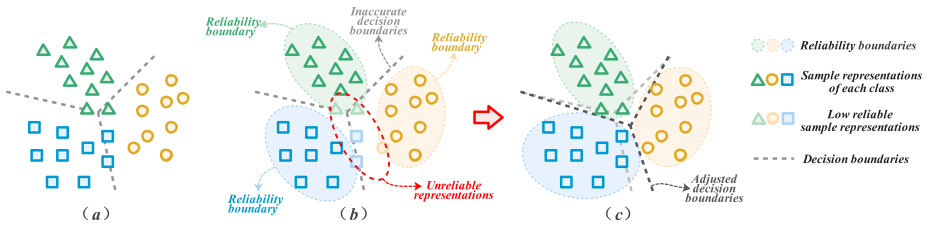

Fig. 1. Illustration of reliability-aware representation learning under sparse scribble supervision. (a) Initial decision boundaries with unreliable sample representations in under-supervised regions. (b) Class-conditional feature distributions showing reliability boundaries that separate confident and ambiguous regions. (c) Adjusted decision boundaries after reliability-guided refinement, where unreliable representations are corrected through feedback learning.

boundaries, leading to either under-utilization of reliable predictions or over-reliance on noisy pseudo-labels. These limitations manifest particularly in complex multi-organ segmentation scenarios where adjacent structures share similar intensity profiles or exhibit intricate topological relationships.

The fundamental challenge in scribble-supervised frameworks stems from the inherent sparsity of annotations, which creates two critical limitations: (1) Feature representation ambiguity in unsupervised regions, where models fail to discern discriminative patterns between adjacent anatomical structures; (2) Decision boundary instability in multi-organ segmentation tasks, particularly for organs with similar intensity profiles or complex topological relationships. These issues typically manifest as inter-class confusion and reduced generalization performance, with errors concentrating in anatomically ambiguous or under-supervised regions. From a feature-space perspective, reliable regions tend to form compact and class-consistent representations, whereas features extracted from poorly supervised areas are often dispersed, noisy, and unstable. Such behavior reflects epistemic and aleatoric uncertainty induced by limited supervision. This observation motivates a reliability-centered formulation of the problem. Instead of categorizing samples based on distribution membership, a more pertinent question is how to explicitly quantify spatial feature reliability within the learned embedding space, and how to leverage this information to guide representation refinement. Specifically, two key questions arise: (1) How can the reliability of feature representations be assessed through distribution-aware modeling under sparse supervision? (2) How can this reliability information be integrated into the training process to focus learning on unreliable regions?

To this end, we propose a reliability-aware feedback learning framework for scribble-supervised segmentation. Inspired by the Information Bottleneck (IB) principle [18], we aim to learn discriminative embeddings that encourage intra-class compactness while enhancing inter-class separability. Under sparse scribble supervision, features from well-annotated regions tend to be compact and consistent, while those from under-supervised or ambiguous regions are often scattered and unstable. We explicitly model feature reliability in the embedding space to guide targeted representation refinement (Fig. 1). Our framework comprises three synergistic components. (1) Reliability-aware Representation Modeling: Class-conditional features are modeled with multivariate Gaussians, and sample-to-class discrepancy is quantified via Mahalanobis distance. Intra-class variance is reduced and inter-class separation is enhanced through a triplet loss, yielding spatially resolved feature reliability. (2) Reliable Pseudo-Supervision: The decoder generates perturbed predictions to select high-confidence pseudo-labels, emphasizing reliable regions while suppressing noisy ones to mitigate error accumulation under sparse supervision. (3) Reliability-guided Feedback Learning: Reliability cues are fed back into the network to adaptively strengthen learning in ambiguous or under-supervised regions, enabling representation refinement. By updating feature distributions and reinforcing learning on challenging regions, the framework enables a progressive refinement process that steadily improves representations without additional supervision or architectural complexity. This enables the model to capture subtle anatomical structures and boundaries, achieving robust segmentation under sparse scribble supervision.

Experiments on the BTCV [19] and WORD [20] 3D multi-organ CT datasets demonstrate that the method significantly improves segmentation accuracy for small and complex organs while exhibiting excellent generalization and robustness performance.

2. Related work

2.1. Scribble-supervised segmentation

Scribble-supervised segmentation is favored for its flexibility and efficiency. Traditional methods like Random Walker [21] and GrowCut [22] existed but performed worse than deep learning (DL) methods. Existing approaches to scribble-supervised segmentation primarily address sparse supervision through three strategies: (1) Pseudo-label propagation, where methods like S2L [2] and DMPLS [1] generate pseudo-labels from scribbles via walk algorithms or dual-network consensus, though error accumulation remains a challenge; (2) Loss function engineering, including partial cross-entropy (pCE) [7] and Mumford-Shah losses [8], which mitigate supervision sparsity by selectively weighting labeled pixels or enforcing geometric constraints; (3) Consistency regularization [23], exemplified by USTM [16] and CPS [24], which leverages input/feature perturbations or cross-pseudo supervision to exploit unlabeled regions. While improving label efficiency, these methods often lack explicit boundary modeling, causing unreliable predictions in under-supervised regions.

2.2. Reliability and uncertainty modeling

Recent advances in reliability and uncertainty modeling have shown promise in mitigating annotation sparsity for medical image segmentation. Classical methods like Monte Carlo Dropout [25] and ensembles [26] quantify prediction-level uncertainty via output variability, yet often overlook latent feature reliability critical under sparse supervision. Out-of-distribution detection further improves model safety by identifying distributional shifts [27,28]. Beyond prediction uncertainty, feature-level approaches address representation ambiguity: Dirichlet-based evidential learning [29] models epistemic uncertainty via higher-order distributions; energy-based models [30] assess confidence through scalar energy functions; contrastive learning [31] enhances discriminability via intra-class compactness and inter-class separation; and gradient-based propagation [32] guides feature refinement through uncertainty transmission. Despite emphasizing representation reliability, most existing methods treat these signals as auxiliary rather than integrating them for spatially adaptive learning.

2.3. Feedback and conditioning mechanisms

Recent advances in feedback and conditioning mechanisms have enhanced the adaptability of medical image segmentation under limited supervision. Prompt-based learning leverages sparse inputs (e.g., points, boxes, text) to guide segmentation: Prompt-DA [33] employs

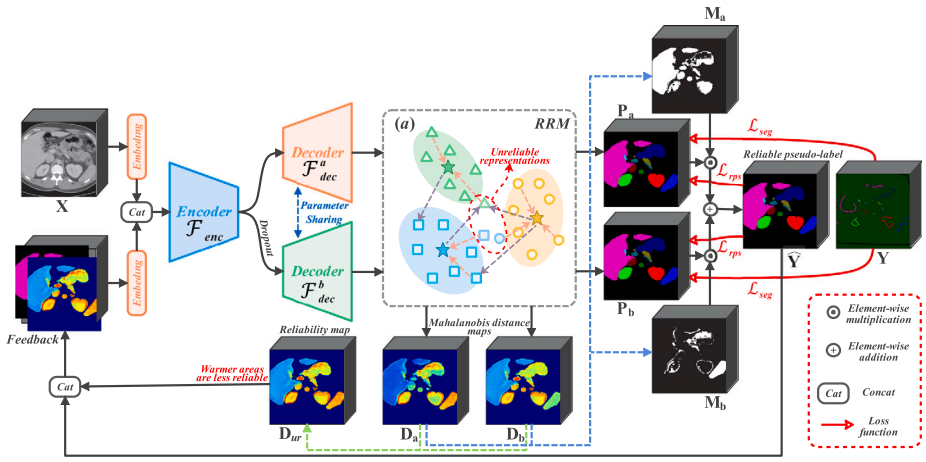

Fig. 2. Architecture of the proposed reliability-aware feedback learning framework. The overall pipeline consists of an encoder and shared decoders with feedback connections. The framework integrates three key components: Reliability-aware Representation Modeling (RRM), Reliable Pseudo-Supervision, and Reliability-guided Feedback Learning. The framework takes input image **X** and sparse scribble annotations **Y**, generates embedding features, computes Mahalanobis distance maps to identify unreliable regions (warmer areas indicate lower reliability), produces reliable pseudo-labels through dual-prediction fusion, and feeds reliability-aware feedback signals back to the network to focus learning on hard regions.

domain-specific prompts for cross-domain feature adaptation; PE-MED [34] iteratively refines results via user clicks and attention-based feedback; and foundation models like SAM [35] modulate latent representations using sparse prompts. Beyond explicit prompting, implicit conditioning approaches feed auxiliary signals back into networks without user interaction: consistency-based frameworks [36] exploit prediction discrepancies for refinement in weak/semi-supervised settings, while generative models [37] synthesize auxiliary representations to aid training. These methods prioritize robustness and consistency over explicit cues. Nevertheless, existing prompt-based approaches remain largely tailored to interactive or foundation-model paradigms and depend on sparse, externally provided inputs.

3. Method

This paper is set against the backdrop of scribble-supervised learning, where sparse annotations represent only a small subset compared to dense annotations. We consider a dataset $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{N}$, which contains 3D volumes of CT scans $\mathbf{X} \in \mathbb{R}^{W\times H\times L}$ and their corresponding sparse annotations $\mathbf{Y} \in \{0, 1, \ldots, K\}^{W\times H\times L}$. The proposed pipeline is illustrated in Fig. 2. We adopt a V-Net [38] backbone and introduce a reliability-aware learning strategy that couples (i) Reliability-aware Representation Modeling, (ii) Reliable Pseudo-Supervision, and (iii) Reliability-guided Feedback Learning. Specifically, we attach two prediction branches with shared weights, where one branch applies dropout to induce feature-level perturbations, producing complementary outputs under the same encoder representation. Based on class-conditional Gaussian modeling in the embedding space, we compute Mahalanobis distance maps to estimate spatial reliability beyond probability confidence. These reliability cues are then used to fuse the complementary predictions and generate more robust pseudo-labels for self-training. Meanwhile, regions with low estimated reliability, together with the pseudo-labels, are transformed into an additional conditioning input and injected back into the network, enabling the model to focus updates on under-supervised and difficult areas while reducing error accumulation.

3.1. Reliability-aware representation modeling

To characterize the reliability of learned feature representations under sparse scribble supervision, we propose a class-conditional Gaussian modeling strategy that enables distribution-aware measurement of representation quality in the embedding space. The proposed design is conceptually inspired by the Information Bottleneck (IB) principle [18], which advocates learning representations that are maximally informative about the target labels while remaining compact. Under sparse supervision, this intuition is particularly relevant: well-supervised regions should yield compact, class-consistent embeddings, while under-supervised or ambiguous regions are expected to produce dispersed, unstable features that are difficult to classify reliably. Directly optimizing the intractable IB objective is avoided because it involves mutual information terms that are computationally prohibitive. Instead, we adopt a practical discriminative surrogate that operationalizes the desiderata through explicit distribution modeling and metric-based regularization. Formally, the IB objective is $\mathcal{L}_{\mathrm{IB}} = I(X; Z) - \beta \cdot I(Z; Y)$. Rather than solving it directly, we focus on two key desiderata: label predictiveness and representation compactness. We instantiate these goals through a discriminative framework built on class-conditional Gaussian distributions.

Under sparse supervision, features extracted from well-annotated regions tend to form compact and class-consistent clusters, whereas features from under-supervised or ambiguous regions are often dispersed and unstable. To capture this behavior, we model each foreground class $c \in \{1, \ldots, K\}$ using a *multivariate Gaussian* distribution in the embedding space:

$$q(Z \mid Y = c) = \mathcal{N}(Z;\, \boldsymbol{\mu}_c,\, \boldsymbol{\Sigma}_c), \tag{1}$$

where $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are trainable class-specific mean and covariance parameters. In practice, we model $\boldsymbol{\Sigma}_c$ as a diagonal covariance matrix to ensure numerical stability and efficient inversion. Because these parameters are optimized over the entire training trajectory, they implicitly aggregate information from the full dataset. We note that this Gaussian assumption does not imply that the true feature distribution is strictly unimodal; instead, it serves as a tractable approximation that enables covariance-aware characterization of class compactness. The background class is excluded from this modeling, as it typically aggregates heterogeneous anatomical structures and acquisition artifacts that cannot be well described by a single distribution. It is still supervised through the segmentation loss.

Given a voxel embedding $\mathbf{z} = f_\theta(X)$ with label (Y = c), we define a distance-based energy score using the Mahalanobis distance:

$$r(Y = c \mid Z = \mathbf{z}) \propto \exp\left(-D_M\left(\mathbf{z},\, \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)\right)\right), \tag{2}$$

where $D_M(\mathbf{z}, \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)) = \frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c^{-1}(\mathbf{z}-\boldsymbol{\mu}_c)$ serves as an unnormalized energy score rather than a calibrated probability. Minimizing the

corresponding negative log-likelihood encourages embeddings to approach their class prototypes in a covariance-normalized sense, leading to the following compactness objective:

$$\mathcal{L}_{comp} \propto -\log \mathcal{N}\left(f_\theta(X); \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\right) = \frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c^{-1}(\mathbf{z}-\boldsymbol{\mu}_c) + \frac{1}{2}\log\left|\boldsymbol{\Sigma}_c\right|. \quad (3)$$

Minimizing $\mathcal{L}_{\text{comp}}$ promotes intra-class compactness and stabilizes feature representations. Importantly, the resulting Mahalanobis distance provides a spatially resolved, distribution-aware signal: embeddings that lie close to their class centers exhibit low distance values and are considered reliable, whereas embeddings that deviate significantly from learned distributions are indicative of under-supervised or poorly learned regions. In this work, we interpret such deviations as a practical proxy for feature-level unreliability, since sparse supervision often leads to insufficiently constrained representations in these regions.

While $\mathcal{L}_{\text{comp}}$ enforces intra-class consistency, it does not explicitly guarantee inter-class separability. This is particularly problematic in multi-organ segmentation, where adjacent structures may share similar intensity patterns. To address this, we introduce a margin-based triplet constraint in the Mahalanobis space. For an anchor embedding $\mathbf{z}$ of class c and a negative class $c^- \neq c$ sampled from other foreground classes within the mini-batch, we enforce:

$$D_M\left(\mathbf{z}, \mathcal{N}(\boldsymbol{\mu}_{c^-}, \boldsymbol{\Sigma}_{c^-})\right) \geq D_M\left(\mathbf{z}, \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)\right) + \alpha, \quad (4)$$

where $\alpha > 0$ is a margin hyperparameter. This constraint enlarges inter-class margins and improves discriminability in the embedding space.

Combining the segmentation objective, intra-class compactness, and inter-class separation, the overall supervised representation modeling loss is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{sup}} = \mathcal{L}_{\text{seg}} &+ \varphi \cdot \mathbb{E}\left[D_M\left(\mathbf{z}, \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)\right)\right] \\ &+ \psi \cdot \mathbb{E}\left[\max\left(0, D_M\left(\mathbf{z}, \mathcal{N}(\boldsymbol{\mu}_{c^-}, \boldsymbol{\Sigma}_{c^-})\right) - D_M\left(\mathbf{z}, \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)\right) + \alpha\right)\right],\end{aligned} \quad (5)$$

where $\mathcal{L}_{\text{seg}}$ is the partial segmentation loss computed on scribble-labeled voxels, and both regularization terms are applied exclusively to annotated foreground voxels. Scalars φ and ψ control the relative strength of each regularizer. Together, the three terms realize the IB intuition in a fully tractable, discriminative manner: the segmentation loss maximizes label predictiveness; the compactness term promotes intra-class concentration; and the triplet loss enforces inter-class separability. The resulting Mahalanobis distance maps provide the reliability signal that drives both pseudo-label generation (Section 3.2) and feedback learning (Section 3.3).

3.2. Reliable pseudo-supervision

To address the inherent limitations of sparse scribble annotations, we develop a robust pseudo-label generation mechanism that exploits prediction diversity induced by stochastic decoding behaviors. By applying different dropout configurations during decoding, the shared decoder produces diverse stochastic prediction views, which are jointly leveraged to improve pseudo-label reliability. Mahalanobis distance maps are employed to quantify prediction reliability and suppress error propagation from unreliable regions.

Let $\mathbf{D}_a$ and $\mathbf{D}_b$ denote the Mahalanobis distance maps obtained from two stochastic forward passes of the shared decoder, corresponding to different dropout states. Although generated by the same decoder with shared parameters, these predictions exhibit distinct characteristics due to stochastic regularization. We construct a reliability mask $\mathbf{M}$ that selects the more reliable prediction at each spatial location:

$$\mathbf{M}_a = \mathbb{I}(\mathbf{D}_a \leq \mathbf{D}_b),\ \mathbf{M}_b = \mathbb{I}(\mathbf{D}_b \leq \mathbf{D}_a), \quad (6)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. This binary masking strategy effectively partitions the prediction space based on relative reliability levels.

The final pseudo-label $\hat{\mathbf{Y}}$ is synthesized through a weighted combination of decoder outputs $\mathbf{P}_a$ and $\mathbf{P}_b$, guided by the reliability masks:

$$\hat{\mathbf{Y}} = \arg\max\left(\mathbf{P}_a \odot \mathbf{M}_a + \mathbf{P}_b \odot \mathbf{M}_b\right), \quad (7)$$

where $\odot$ represents element-wise multiplication. This formulation ensures that each spatial location is supervised by the most reliable prediction available from either decoder.

To enforce prediction consistency between the shared decoders while maintaining their complementary characteristics, we employ a bidirectional consistency loss:

$$\mathcal{L}_{rps} = \ell_{CE}(\sigma(\mathbf{P}_a), \hat{\mathbf{Y}}) + \ell_{CE}(\sigma(\mathbf{P}_b), \hat{\mathbf{Y}}), \quad (8)$$

where ℓ_{CE} denotes the Cross-Entropy loss function, and σ represents the *Softmax* function. This loss term simultaneously: (1) encourages each prediction to align with the consensus pseudo-label in high-confidence regions. (2) preserves decoder-specific characteristics in unreliable areas through differential dropout patterns. (3) facilitates progressive refinement of pseudo-labels through iterative training. During inference, the most reliable predictions are selected as the result.

3.3. Reliability-guided feedback learning

Using the masks $\mathbf{M}_a$ and $\mathbf{M}_b$ obtained in Eq. (6), we can easily extract the Mahalanobis distance map for unreliable regions, denoted as: $\mathbf{D}_{ur} = \mathbf{D}_a \odot \mathbf{M}_b + \mathbf{D}_b \odot \mathbf{M}_a$. We concatenate $\mathbf{D}_{ur}$ and $\hat{\mathbf{Y}}$ along the channel dimension and project them into an embedding space matching the input image dimensions using a lightweight convolutional embedding layer:

$$\text{Ebd}(\cdot) = \text{Conv}_{3\times3}(\text{LeakyReLU}(\text{Conv}_{1\times1}(\cdot))), \quad (9)$$

where the 1×1 convolution reduces channel dimensions and the 3×3 convolution establishes spatial correlations, with LeakyReLU ($\alpha = 0.2$) providing non-linearity. Simultaneously, the input image $\mathbf{X}$ is also embedded and added to the former, then re-fed into the network, expressed as follows:

$$\mathbf{Q} = \mathcal{F}_{enc}(\text{Ebd}(\mathbf{X}) + \text{Ebd}(\text{Cat}(\hat{\mathbf{Y}}, \mathbf{D}_{ur})); \boldsymbol{\Theta}_{enc}), \quad (10)$$

$$\mathbf{P}_a = \mathcal{F}^a_{dec}\left(\mathbf{Q}, \boldsymbol{\Theta}_{dec}\right),\ \mathbf{P}_b = \mathcal{F}^b_{dec}\left(\text{Dropout}(\mathbf{Q}), \boldsymbol{\Theta}_{dec}\right), \quad (11)$$

where $\mathcal{F}_{enc}(\cdot, \boldsymbol{\Theta}_{enc})$ denotes the encoder, $\boldsymbol{\Theta}_{enc}$ represents its trainable parameters, $\mathcal{F}^a_{dec}$ and $\mathcal{F}^b_{dec}$ denote two decoder branches sharing parameters, $\text{Cat}(\cdot)$ stands for the concatenation operation along the channel dimension, and $\text{Ebd}(\cdot)$ indicates embedding operation.

To enable the network to develop distinct sensitivity patterns for homologous features, thereby generating diverse prompts, we design slightly different loss functions for the two outputs of the shared decoders:

$$\mathcal{L}_{seg} = \frac{1}{3}\left[\ell_{pCE}(\sigma(\mathbf{P}_a), \mathbf{Y}) + \ell_{pCE}(\sigma(\mathbf{P}_b), \mathbf{Y}) + \ell_{pDice}(\mathbf{P}_b, \mathbf{Y})\right], \quad (12)$$

where ℓ_{pCE} denotes the partial Cross-Entropy [39] loss function, and ℓ_{pDice} represents the partial Dice loss function.

The overall loss calculation is as follows: $\mathcal{L} = \mathcal{L}_{sup} + \eta \cdot \mathcal{L}_{rps}$, where η represents a *sigmoid* ramp-up schedule to gradually introduce the pseudo-supervision.

4. Experiments

4.1. Datasets

BTCV: The BTCV multi-organ segmentation dataset originates from the *MICCAI Multi-Atlas Labeling Beyond Cranial Vault Workshop Challenge* [19]. It comprises 30 subjects with 3779 axial abdominal CT slices, annotated with 13 organ classes and one background class. In preprocessing, we resampled all scans to a voxel spacing of $[1.5 \times 1.5 \times 2.0]\,\text{mm}^3$ according to [40] and normalized them to have zero mean and

Table 1
Quantitative comparison results on the BTCV dataset. The V-Net is trained with full supervision, serving as an upper bound. The results in **bold** are the best. The primary evaluation metric is Dice Similarity Coefficient (DSC) (%). Note: **Avg**: average. **95HD**: 95% Hausdorff Distance (**mm**), **Spl**: Spleen, **R.Kid**: right kidney, **L.Kid**: left kidney, **Gall**: gallbladder, **Eso**: esophagus, **Stom**: Stomach, **IVC**: inferior vena cava, Veins: portal and splenic veins, Pan: Pancreas, **LG/RG**: left/right adrenal glands.

Methods	Spl	R.kid	L.kid	Gall	Eso	Liver	Stom	Aorta	IVC	Veins	Pan	RG	LG	Avg.DSC	Avg.95HD
V-Net	95.59	94.12	94.00	74.05	73.69	96.72	85.87	88.94	86.42	73.95	79.13	67.18	65.75	$82.72_{11.73}$*	$6.91_{20.01}$
pCE [39]	85.91	76.54	82.55	40.05	23.38	87.43	61.91	59.61	50.34	32.37	35.76	17.13	9.25	$50.94_{27.51}$*	$60.62_{39.73}$*
DMPLS [1]	78.67	84.36	80.67	31.97	25.57	85.66	48.80	67.36	56.92	40.05	46.65	14.89	14.08	$51.97_{26.53}$*	$56.37_{32.70}$*
TV [41]	74.83	80.62	76.70	35.34	28.91	88.13	58.88	63.01	59.89	44.93	44.45	11.01	13.67	$52.34_{25.67}$*	$55.02_{30.45}$*
CPS [24]	89.92	86.56	85.42	54.87	51.85	89.82	70.56	67.31	64.00	52.08	55.95	26.42	29.22	$63.38_{22.34}$*	$20.79_{17.00}$*
EM [42]	85.86	84.10	84.07	57.90	58.29	83.16	76.55	70.51	65.35	48.98	58.06	30.74	31.56	$64.24_{19.86}$*	$45.50_{44.57}$*
USTM [16]	91.06	89.59	90.49	58.79	64.41	88.79	61.18	83.95	75.61	60.65	71.69	41.09	43.82	$70.85_{18.37}$*	$36.37_{37.06}$*
EFFDNet [43]	92.77	89.83	89.94	65.54	65.13	86.55	78.53	86.67	74.79	62.68	73.02	46.32	51.26	$74.08_{15.62}$*	$17.11_{26.52}$*
S2L [2]	93.49	87.94	**90.54**	66.67	64.83	**94.06**	80.83	81.44	77.63	61.95	71.11	45.91	51.48	$74.45_{16.82}$*	$11.18_{9.76}$*
DMSPS [44]	93.07	89.87	89.81	66.73	65.26	86.16	81.59	87.15	74.65	63.04	73.25	47.24	52.52	$74.65_{16.08}$*	$15.13_{25.04}$*
Ours	**94.64**	**90.75**	90.21	**70.21**	**68.80**	86.92	**86.91**	**89.01**	**83.19**	**71.24**	**78.53**	**52.60**	**57.96**	$\mathbf{78.54}_{13.96}$	$\mathbf{8.08}_{7.21}$

* $P < 0.05$ via paired **t-test** for Avg.DSC and Avg.95HD.

unit variance. The training and test sets were divided as 24 and 6 cases, respectively, following [40].

WORD: The WORD dataset [20] comprises 150 abdominal CT scans from 150 subjects at a single center, featuring 16 anatomical classes and 1 background class. Each scan contains 159–330 slices of 512 × 512 pixels, with an in-plane resolution of 0.976 mm × 0.976 mm and a slice thickness of 2.5–3.0 mm. Preprocessing of volumes is similar to that of BTCV. The training, validation, and test sets follow [20], consisting of 100, 20, and 30 cases, respectively.

4.2. Implementation details and evaluation metrics

The backbone of all methods is based on V-Net. Our proposed method, alongside other comparative methods, was implemented utilizing Python 3.8, PyTorch 1.12, and the WSL4MIS[2] codebase. Due to the scarcity of methods for 3D scribble-supervised segmentation, we adapted suitable 2D methods to 3D data, excluding those incompatible. Training was conducted on an NVIDIA RTX 3090 GPU. We employ SGD to minimize the combined objective function, with a weight decay of 10^{-4}, momentum of 0.9, a total of 50k iterations, an initial learning rate of 0.01, and online decay based on $lr = 0.9 \times lr$. During training, 96 × 96 × 96 samples are randomly cropped from CT scans to serve as inputs to the network. All experiments were conducted in the same experimental environment, employing Dice Similarity Coefficient (DSC) and 95% Hausdorff Distance (95HD) as evaluation metrics.

Scribble: All scribble annotations were generated from dense labels. Specifically, we randomly sampled points within segmentation targets and connected them using B-spline algorithms to create natural, human-like scribbles.

Inference Protocol: For inference, we strictly follow the deterministic two-pass forward protocol identical to the training pipeline: an initial pass computes reliability cues and pseudo-labels, which are immediately fed back for one refinement pass to produce the final segmentation.

4.3. Comparison with existing methods

We compared our proposed method with 9 classic and advanced scribble-supervised learning methods on two public multi-organ CT datasets. These methods encompass **TV** (Javanmardi et al. 2016) [41], **pCE** (Tang et al. 2018) [39], **EM** (Yu et al. 2019) [42], **S2L** (Lee et al. 2020) [2], **CPS** (Chen et al. 2021) [24], **DMPLS** (Luo et al. 2022) [1], **USTM** (Liu et al. 2022) [16], **DMSPS** (Han et al. 2024) [44], and **EFFDNet** (Liu et al. 2025) [43]. As baselines, pCE sets the lower bound for all methods, and a fully supervised V-Net was also trained to serve as the upper bound reference.

BTCV: On the BTCV dataset (Table 1), our method achieved superior performance, with an average Dice score of 78.54%, clearly outperforming existing scribble-supervised approaches. It showed particular strength in segmenting small and complex structures such as the gallbladder and esophagus, where prior methods struggled. Compared to the fully supervised V-Net, our approach substantially narrowed the performance gap while relying only on sparse annotations. Moreover, the 95% Hausdorff Distance further confirmed its robustness, outperforming DMSPS and S2L. These results highlight the effectiveness of our feedback learning and distribution modeling in mitigating feature ambiguity in under-supervised regions.

As shown in Fig. 3, some methods misjudge organ boundaries, likely due to insufficient discriminative feature learning between classes. Others misidentify uncommon organs (e.g., segmenting the heart as the liver or left adrenal gland), possibly reflecting poor model awareness of reliability limits. In contrast, our method achieves better segmentation for small organs like the gallbladder. **WORD:** On the WORD dataset (Table 2), our method achieved the best average Dice score (78.45%), outperforming all competitors, including DMSPS. The gains were especially evident for challenging organs such as the duodenum and intestines, where scribble supervision typically struggles due to structural complexity. In addition, our method obtained a lower 95HD (7.45 mm), indicating improved boundary accuracy. These results demonstrate the generalizability of our framework and its effectiveness in handling hard regions through reliability-aware feedback.

As seen in Fig. 3, most methods perform poorly on small organs (e.g., adrenal glands), whereas ours achieves superior results. Our method also delineates complex organs (e.g., colon, pancreas) with higher precision. Moreover, thanks to explicit reliability modeling, our approach rarely misidentifies organs.

4.4. Ablation study

4.4.1. Component analysis

To validate the contribution of each key component, we conducted a comprehensive ablation study. Removing the distribution modeling module leads to a notable performance drop (e.g., −2.36% DSC on BTCV), confirming its role in establishing discriminative feature boundaries via Mahalanobis optimization. Disabling the shared decoder causes the most severe degradation (up to −5.62% DSC), underscoring its necessity for capturing feature diversity and enabling reliable pseudo-label fusion. Omitting the feedback learning mechanism also reduces performance (e.g., −3.22% DSC), as it explicitly guides optimization toward under-supervised hard regions. As summarized in Table 3, the full model synergistically integrates these components, achieving an optimal balance for robust scribble-supervised segmentation.

[2] https://github.com/HiLab-git/WSL4MIS.

Table 2
Quantitative results on the **WORD** dataset. Note: **Liv**: liver, **Duod**: duodenum, **Col**: colon, **Inte**: intestine, **Adr.G**: adrenal gland, **Rect**: rectum, **Blad**: bladder, LFH/RFH: left/right femur head. Other abbreviations as in Table 1.

Methods	Liv	Spl	L.kid	R.kid	Stom	Gall	Eso	Pan	Duod	Col	Inte	Adr.G	Rect	Blad	LFH	RFH	Avg.DSC	Avg.95HD
V-Net	96.42	95.07	95.02	95.34	90.58	77.91	74.78	82.99	64.99	84.56	87.60	67.83	78.37	90.38	88.99	89.11	$85.00_{12.77}$*	$7.32_{19.26}$
DMPLS [1]	88.86	83.24	75.01	80.52	73.82	32.72	23.39	58.47	33.72	56.36	55.61	13.92	49.66	72.61	67.90	62.72	$58.03_{23.63}$*	$51.01_{39.24}$*
TV [41]	89.74	86.79	78.44	85.50	75.30	38.71	23.15	62.12	39.00	46.52	57.07	22.57	53.28	72.39	63.56	65.08	$59.95_{22.71}$*	$63.59_{40.43}$*
pCE [39]	88.94	83.00	77.29	79.44	76.23	58.95	26.03	60.51	38.67	54.58	53.78	16.38	51.02	70.34	64.60	60.27	$60.00_{22.40}$*	$50.65_{35.38}$*
USTM [16]	89.75	88.47	83.65	83.62	79.59	35.85	35.19	62.01	45.58	55.76	59.40	25.21	54.58	73.48	77.68	80.91	$64.42_{22.27}$*	$30.75_{36.41}$*
S2L [2]	91.28	92.00	89.49	86.17	82.79	61.37	45.23	67.05	54.74	69.97	21.50	32.33	63.33	37.94	82.01	80.13	$66.08_{23.65}$*	$25.48_{32.52}$*
CPS [24]	92.05	87.89	82.81	87.41	82.17	57.62	32.96	61.29	44.64	62.83	55.84	25.42	56.74	73.67	77.62	79.35	$66.27_{21.72}$*	$15.66_{20.03}$*
EM [42]	90.51	87.78	83.77	85.06	80.99	54.82	34.25	61.17	47.41	59.22	57.84	30.41	53.99	75.59	78.33	82.13	$66.45_{20.92}$*	$17.93_{23.55}$*
EFFDNet [43]	93.65	90.18	87.95	89.84	85.92	58.21	51.35	70.65	55.77	73.56	62.39	42.15	63.80	80.25	83.58	85.24	$73.41_{18.27}$*	$11.68_{14.96}$*
DMSPS [44]	93.58	89.32	88.61	90.32	84.88	60.53	54.75	69.83	53.22	71.86	59.01	49.33	66.02	81.01	84.09	85.39	$73.86_{17.56}$*	$10.56_{14.27}$*
Ours	**93.92**	**92.38**	**90.11**	**91.56**	**88.14**	**68.03**	**59.45**	**76.09**	**62.99**	**80.24**	**80.72**	**50.51**	**67.77**	**82.50**	**85.04**	**85.83**	$\mathbf{78.45}_{15.84}$	$\mathbf{7.45}_{10.93}$

* $P < 0.05$ via paired **t-test**.

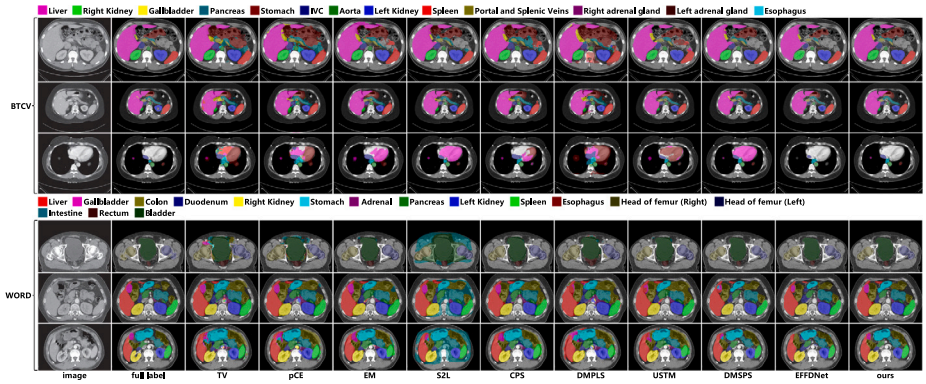

Fig. 3. Visualization of segmentation results by various methods on the BTCV and WORD datasets, with different colors representing each organ. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Ablation study results of key components on BTCV and WORD datasets. **w/o** denotes without.

Module	BTCV		WORD	
	DSC	95HD	DSC	95HD
Full model (Ours)	$\mathbf{78.54}_{13.96}$	$\mathbf{8.08}_{7.21}$	$\mathbf{78.45}_{15.84}$	$\mathbf{7.45}_{10.93}$
w/o Distribution Modeling	$76.18_{14.05}$	$9.45_{9.54}$	$76.02_{16.15}$	$8.73_{12.02}$
w/o Shared Decoder	$73.96_{15.82}$	$12.21_{14.26}$	$72.83_{17.34}$	$11.02_{15.34}$
w/o Feedback Learning	$75.32_{14.62}$	$10.87_{10.13}$	$74.91_{16.82}$	$9.56_{14.18}$

Table 4
Ablation results of pseudo-label generation strategies on the BTCV and WORD datasets.

Pseudo label generation	BTCV		WORD	
	DSC	95HD	DSC	95HD
Ours	$\mathbf{78.54}_{13.96}$	$\mathbf{8.08}_{7.21}$	$\mathbf{78.45}_{15.84}$	$\mathbf{7.45}_{10.93}$
Hard pseudo-labels	$75.29_{14.83}$	$11.29_{10.04}$	$74.18_{16.53}$	$10.83_{12.11}$
Single-decoder pseudo	$73.87_{15.41}$	$13.45_{14.17}$	$72.56_{17.62}$	$12.91_{15.89}$
Fixed scribble labels	$68.21_{18.96}$	$18.76_{19.66}$	$66.94_{20.85}$	$17.32_{23.10}$
Random mask fusion	$72.05_{16.84}$	$14.67_{15.98}$	$71.23_{18.12}$	$13.45_{15.73}$

4.4.2. Analysis of pseudo-label generation

To evaluate the effectiveness of pseudo-label generation strategies, we conducted an ablation study comparing several alternatives (Table 4). The proposed Mahalanobis distance-based reliability weighting achieves optimal performance, substantially outperforming all variants. Direct hard fusion degrades DSC by over 3%, confirming that uncertainty-aware weighting is essential to mitigate error propagation. Relying on static scribble labels or a single decoder further reduces accuracy to 67% DSC, underscoring the necessity of dynamic, complementary predictions. Additionally, random mask fusion underperforms, validating that our distance-based reliability masking outperforms heuristic alternatives. Collectively, these results demonstrate that reliability-weighted pseudo-label fusion is crucial for effectively leveraging sparse scribble annotations.

4.4.3. Analysis of reliable feedback mechanism

The ablation study on reliable feedback mechanism (Fig. 4) demonstrates the critical role of our feedback learning mechanism. Our full method achieves superior performance by adaptively directing the model's focus toward under-learned regions using reliability-aware distance maps. Static feedback shows a significant drop (−3.22% and −3.54%), highlighting the limitation of fixed spatial guidance in handling evolving feature distributions. Random feedback performs worst, confirming that arbitrary spatial cues disrupt feature learning. Notably, completely removing spatial feedback results in worse performance than our method, demonstrating that explicitly targeting challenging regions is essential.

4.4.4. Contribution of shared decoder architecture

The shared-decoder architecture is central to our framework, facilitating diverse feature learning and reliability-aware pseudo-label generation under sparse scribble supervision. As shown in Table 5, a single-decoder setup yields notably lower accuracy (72.15% DSC on BTCV, 70.82% on WORD), while independent dual decoders offer moderate gains. Our shared-parameter design further improves consistency and efficiency by synchronizing feature extraction and preserving diversity through dropout perturbations. This enables robust pseudo-label fusion and precise identification of challenging regions, achieving an optimal trade-off between segmentation accuracy, boundary adherence, and computational cost.

Table 5
Ablation results on the contribution of the shared-decoder structure on BTCV and WORD datasets. "**independent**" denotes independent parameters.

Pseudo label generation	BTCV		WORD		Param
	DSC	95HD	DSC	95HD	
Single decoder	$72.15_{16.74}$	$15.83_{16.32}$	$70.82_{19.08}$	$16.47_{22.14}$	9.45M
Single decoder + Dropout	$73.96_{15.58}$	$14.26_{15.61}$	$72.64_{17.75}$	$15.12_{19.80}$	9.45M
Dual decoders (independent)	$75.42_{14.69}$	$12.91_{13.88}$	$74.35_{16.27}$	$13.28_{15.96}$	11.53M
Shared decoder + Dropout	$\mathbf{78.54}_{13.96}$	$\mathbf{8.08}_{7.21}$	$\mathbf{78.45}_{15.84}$	$\mathbf{7.45}_{10.93}$	9.45M

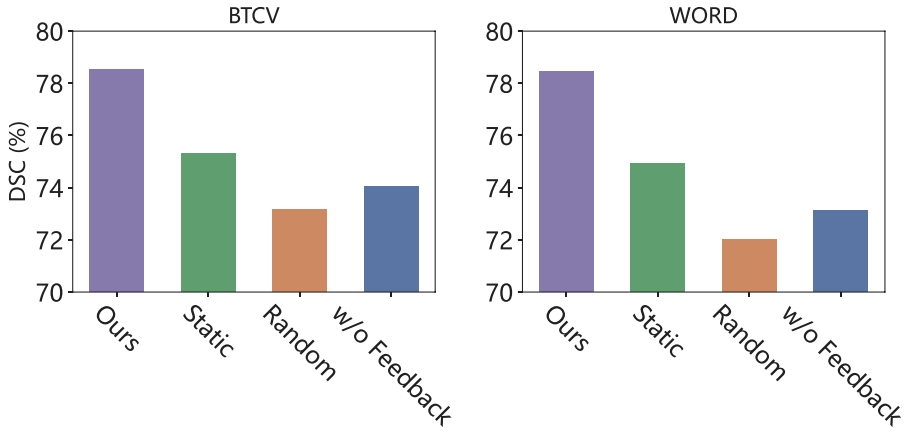

Fig. 4. Ablation results of reliable feedback mechanism on BTCV and WORD datasets.

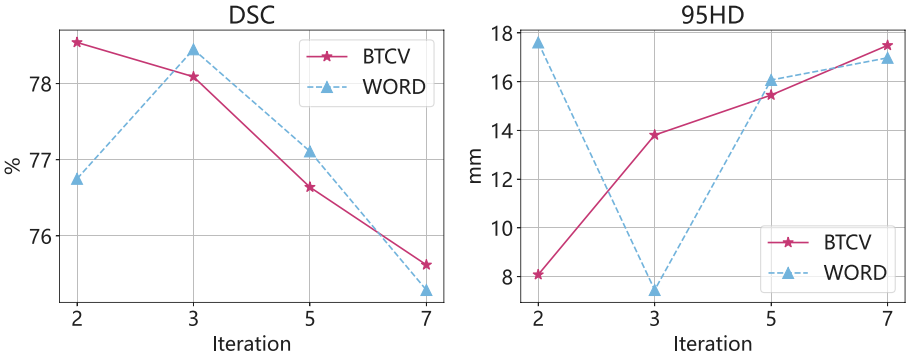

Fig. 5. Comparison of feedback iterations on BTCV and WORD datasets.

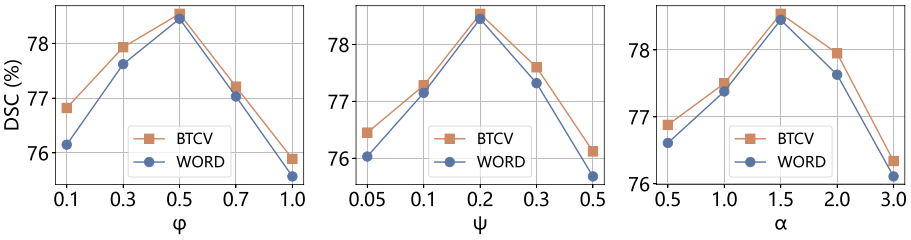

Fig. 6. Ablation study on hyperparameters φ, ψ, and α. Quantitative results (DSC %) of varying hyperparameter values on BTCV and WORD datasets.

4.4.5. *Analysis of feedback iterations during testing*

We conduct an ablation study on iterative reliability feedback (Fig. 5). Performance peaks at the first iteration (78.54% DSC on BTCV) and gradually declines with additional iterations (e.g., 75.62% after 7), consistent with our design: reliability-guided feedback targets under-supervised regions via single-step application during training. The initial gain confirms that Mahalanobis-based reliability maps provide spatially coherent guidance aligned with segmentation difficulty, while subsequent degradation arises from reusing the same signal without representation re-adaptation. Importantly, all iterations outperform the no-feedback baseline (Table 3), validating the efficacy of the reliability signal and justifying our single-step feedback design, which achieves optimal performance without complex stopping criteria.

4.4.6. *Hyperparameter analysis*

We conducted a comprehensive analysis of three key hyperparameters, φ (pseudo-supervision weight), ψ (inter-class separation weight), and α (Mahalanobis distance margin), to evaluate their impact on segmentation performance. The results are shown in Fig. 6. Results show

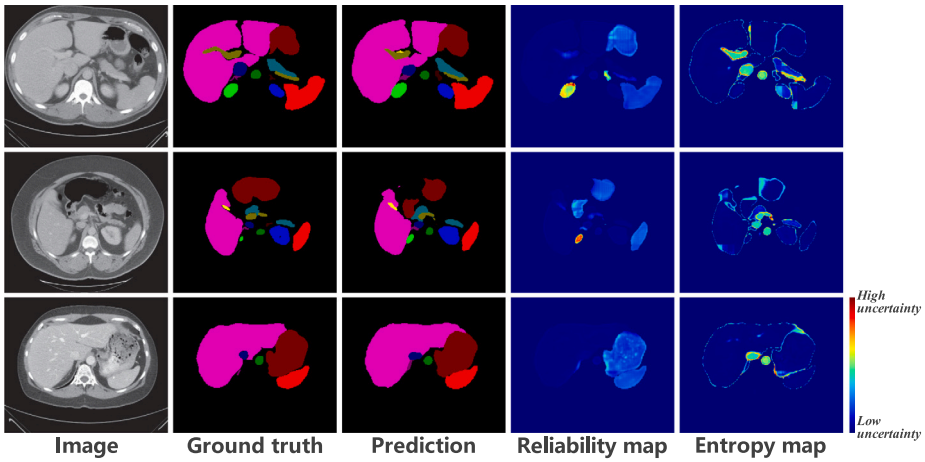

Fig. 7. Visualization of reliability map and entropy map. The reliability map exhibits distinct patterns from the entropy map.

Table 6
Quantitative correlation between reliability maps and segmentation errors across datasets. Higher correlation coefficients and AUC values indicate better alignment between reliability predictions and actual errors. **RR** denotes Reliability Regions.

Methods	Dataset	Pearson Corr.	Spearman Corr.	AUC for Error Detection	Error in Low-RR (%)	Error in High-RR (%)
Ours	BTCV	$0.783_{0.052}$	$0.821_{0.049}$	0.892	$32.45_{5.21}$	$4.87_{1.36}$
MC Dropout		$0.612_{0.074}$	$0.648_{0.065}$	0.763	$41.38_{6.74}$	$8.92_{2.15}$
Entropy map		$0.526_{0.087}$	$0.557_{0.076}$	0.691	$43.27_{7.12}$	$10.64_{3.07}$
Ours	WORD	$0.801_{0.041}$	$0.835_{0.030}$	0.907	$28.76_{4.89}$	$3.95_{1.12}$
MC Dropout		$0.628_{0.062}$	$0.659_{0.058}$	0.778	$39.54_{6.23}$	$9.37_{2.08}$
Entropy map		$0.541_{0.081}$	$0.572_{0.063}$	0.705	$41.81_{7.04}$	$10.82_{2.86}$

optimal settings at $\varphi = 0.5$ (balancing pseudo-label utility and noise suppression), $\psi = 0.2$ (enhancing small-organ segmentation without feature over-dispersion), and $\alpha = 1.5$ (maximizing inter-class separation while preserving convergence). Both datasets exhibited consistent trends, though WORD's anatomical complexity heightened sensitivity to ψ. These findings validate our hyperparameter choices and demonstrate the framework's robustness under varying supervision conditions.

4.5. Reliability map analysis

Fig. 7 compares the reliability map and the entropy map, highlighting their key differences. The reliability map dynamically identifies regions with unreliable predictions caused by insufficient feature representation, especially in sparsely annotated areas (e.g., organ boundaries, small structures, and complex anatomies). In contrast, the entropy map reflects prediction probability dispersion, capturing ambiguous regions but failing to detect overconfident errors (e.g., misclassified small organs). By leveraging Mahalanobis distance, the reliability map detects deviations from learned feature distributions, enabling precise localization of under-supervised regions. Unlike entropy, it distinguishes true ambiguity from model ignorance, providing more actionable guidance for reliability feedback.

To validate its effectiveness, we analyze the correlation between predicted reliability and actual segmentation errors. As shown in Table 6, our method achieves higher Pearson and Spearman correlations than conventional approaches, including Monte Carlo Dropout and entropy maps. This leads to improved error detection, with AUC scores of 0.892 and 0.907 on the BTCV and WORD datasets, respectively. Moreover, low-reliability regions consistently correspond to higher error rates. These results demonstrate that distribution-aware modeling via Mahalanobis distance captures feature representation quality more effectively than variance- or entropy-based methods, enabling more accurate spatial guidance for adaptive learning.

Table 7
Ablation study on different uncertainty feedback signals. Our Mahalanobis-based feature-space reliability consistently outperforms prediction-level uncertainty measures (Entropy, MC Dropout, Evidential Deep Learning) when used as feedback conditioning signals.

Feedback Signal	Space level	BTCV		WORD	
		DSC	95HD	DSC	95HD
Mahalanobis distance	Feature	78.54	8.08	78.45	7.45
Entropy	Prediction	75.43	11.23	74.91	10.04
Monte-carlo dropout	Prediction	76.12	10.34	75.89	9.12
Evidential deep learning	Prediction	75.85	10.87	75.34	9.56
Spatial attention	Spatial	75.98	10.61	75.62	9.34

Furthermore, we evaluated different uncertainty feedback signals to verify the advantage of feature-space modeling (Table 7). Our Mahalanobis-based reliability consistently outperforms prediction-level measures and spatial attention, including entropy, Monte Carlo Dropout, and Evidential Deep Learning, in both segmentation accuracy and boundary adherence across datasets. This demonstrates that leveraging embedding-space distributions provides more robust guidance for adaptive learning than relying solely on output probability variations or spatial weighting.

4.6. Generalizability study

Bidirectional cross-dataset experiments (BTCV ↔ WORD) validate our model's generalizability. Our method achieved 70.30% DSC for BTCV→WORD (Table 8) and 67.56% for the more challenging WORD→BTCV (Table 9), outperforming the second-best DMSPS by over 6 percentage points in the latter. Notably, it maintained robust accuracy on small, complex organs (e.g., esophagus, gallbladder) where most baselines nearly failed. These results confirm that our reliability-aware framework captures intrinsic anatomical patterns rather than dataset-specific artifacts, enabling robust deployment across diverse

Table 8
Generalization performance across datasets (BTCV → WORD).

Methods	Spl	R.kid	L.kid	Gall	Eso	Liver	Stom	Pan	Avg.DSC	Avg.95HD
EM [42]	0.00	4.42	0.66	2.81	0.00	22.49	16.33	4.00	$6.34_{9.96}$*	$105.28_{41.78}$*
USTM [16]	0.00	17.09	0.00	1.97	0.01	32.17	9.95	10.93	$9.01_{12.55}$*	$89.24_{37.26}$*
pCE [39]	21.77	59.46	15.59	0.00	0.00	80.98	47.95	27.31	$31.63_{31.00}$*	$66.44_{35.13}$*
DMPLS [1]	59.46	45.52	57.00	17.00	21.70	65.18	28.81	28.41	$40.38_{26.18}$*	$65.18_{37.36}$*
TV [41]	61.57	69.18	74.25	13.53	18.87	79.33	42.50	29.18	$48.55_{27.95}$*	$62.09_{36.21}$*
CPS [24]	73.18	54.23	61.29	23.06	21.31	84.78	50.96	24.09	$49.11_{30.20}$*	$63.49_{43.36}$*
S2L [2]	78.02	82.99	68.03	35.04	36.60	**90.85**	63.69	45.45	$63.15_{41.24}$*	$25.78_{48.47}$*
EFFDNet [43]	80.35	79.12	80.47	40.28	42.63	86.15	68.84	43.56	$65.18_{24.33}$*	$28.36_{29.84}$*
DMSPS [44]	81.82	79.33	81.67	41.89	41.42	84.42	66.77	48.47	$65.72_{24.04}$*	$28.19_{30.37}$*
Ours	**85.92**	**86.20**	**88.40**	**44.11**	**43.56**	86.37	**72.19**	**55.69**	$\mathbf{70.30}_{25.47}$	$\mathbf{17.02}_{16.44}$

* $P < 0.05$ via paired **t-test**.

Table 9
Generalization performance across datasets (WORD → BTCV).

Methods	Spl	R.kid	L.kid	Gall	Eso	Liver	Stom	Pan	Avg.DSC	Avg.95HD
USTM [16]	0.00	1.85	3.56	0.0	0.0	37.74	6.88	0.64	$6.33_{13.26}$*	$97.62_{29.69}$*
EM [42]	0.09	0.0	83.77	0.0	1.14	0.32	0.09	0.36	$10.72_{27.65}$*	$64.32_{24.06}$*
DMPLS [1]	19.71	13.15	49.24	1.47	12.54	46.24	43.66	24.45	$26.31_{21.59}$*	$85.43_{47.24}$*
TV [41]	24.97	74.01	62.07	4.99	26.78	66.14	43.88	22.74	$40.70_{29.16}$*	$81.42_{41.59}$*
pCE [39]	37.45	69.0	66.61	21.86	23.24	79.99	46.5	25.78	$46.30_{26.88}$*	$61.12_{32.42}$*
S2L [2]	57.44	68.74	75.47	11.52	35.86	75.69	52.06	32.18	$51.12_{27.36}$*	$64.55_{37.22}$*
CPS [24]	67.78	75.56	72.57	20.54	29.74	82.13	46.86	24.64	$52.48_{29.78}$*	$54.77_{34.41}$*
EFFDNet [43]	73.61	81.27	78.94	19.83	28.45	88.32	51.36	37.35	$57.39_{26.85}$*	$50.21_{28.52}$*
DMSPS [44]	74.52	**84.06**	74.88	33.56	42.39	**88.41**	53.52	38.28	$61.20_{26.66}$*	$47.18_{42.36}$*
Ours	**84.27**	83.91	**85.54**	**42.20**	**60.11**	86.00	**53.44**	**44.97**	$\mathbf{67.56}_{25.16}$	$\mathbf{38.52}_{38.18}$

* $P < 0.05$ via paired **t-test**.

Table 10
Quantitative comparison under different scribble sparsity levels on BTCV and WORD datasets.

Sparsity	Methods	BTCV							WORD					
		Gall	Eso	Pan	RG	LG	Avg.DSC	Avg.95HD	Gall	Eso	Pan	Adr.G	Avg.DSC	Avg.95HD
5%	pCE	32.15	15.28	28.76	10.23	8.15	$42.18_{26.51}$	$68.73_{38.42}$	45.28	18.37	52.41	9.28	$52.18_{24.40}$	$72.34_{42.18}$
	USTM	45.29	48.17	58.69	32.18	35.42	$62.18_{20.37}$	$45.73_{38.06}$	25.18	26.37	53.58	18.21	$56.42_{24.27}$	$98.24_{42.36}$
	DMSPS	52.64	50.28	60.35	38.42	42.57	$66.18_{18.04}$	$25.73_{28.14}$	48.42	42.57	60.83	38.42	$66.18_{20.04}$	$38.19_{32.14}$
	Ours	58.42	54.73	65.78	42.59	46.83	$70.18_{16.45}$	$15.73_{18.42}$	55.28	45.37	65.09	40.28	$70.18_{18.45}$	$15.73_{20.42}$
10%	pCE	40.05	23.38	35.76	17.13	9.25	$50.94_{27.51}$	$60.62_{39.73}$	58.95	26.03	60.51	16.38	$60.00_{22.40}$	$66.44_{35.13}$
	USTM	58.79	64.41	71.69	41.09	43.82	$70.85_{18.37}$	$36.37_{37.06}$	35.85	35.19	62.01	25.21	$64.42_{22.27}$	$89.24_{37.26}$
	DMSPS	66.73	65.26	73.25	47.24	52.52	$74.65_{16.08}$	$15.13_{25.04}$	60.53	54.75	69.83	49.33	$73.86_{17.56}$	$28.19_{30.37}$
	Ours	70.21	68.80	78.53	52.60	57.96	$78.54_{13.96}$	$8.08_{7.21}$	68.03	59.45	76.09	50.51	$78.45_{15.84}$	$7.45_{10.93}$
20%	pCE	48.37	35.42	45.21	25.67	18.34	$58.73_{25.89}$	$52.18_{36.45}$	65.37	38.42	68.73	25.67	$67.18_{20.37}$	$58.73_{32.45}$
	USTM	65.32	68.73	75.84	48.76	52.91	$75.12_{16.28}$	$28.45_{32.17}$	48.73	45.28	70.35	38.42	$71.18_{20.18}$	$75.18_{35.42}$
	DMPLS	72.15	70.83	78.96	55.37	58.62	$78.92_{14.37}$	$12.45_{20.18}$	68.42	62.18	76.45	55.28	$78.45_{16.28}$	$20.18_{25.73}$
	Ours	75.63	72.45	82.17	58.94	62.38	$82.37_{12.18}$	$6.45_{5.73}$	75.18	68.42	82.37	58.73	$83.18_{14.28}$	$6.18_{8.45}$

protocols, patient populations, and scanner types without retraining or additional annotations.

4.7. Robustness analysis

4.7.1. Scribble sparsity analysis

We validate robustness under varying annotation densities (5%, 10%, 20% scribble sparsity) in Table 10. Our framework maintains superior performance across all conditions, especially for small organs (e.g., gallbladder, adrenal glands) where baselines degrade significantly. At minimal 5% density, our method achieves 70.18% DSC on BTCV, outperforming DMSPS by 4% and USTM by 8%, with 95HD reduced by 10 mm. Performance scales with supervision (78.54% DSC at 10%, 82.37% at 20%), while the advantage over baselines remains consistent across datasets (WORD: +7% DSC over DMSPS at 10%). Stable boundary adherence (95HD < 8 mm at all levels) further confirms effective mitigation of sparse-supervision challenges.

4.7.2. Scribble distribution analysis

To verify the robustness of the model under different annotation habits, we systematically compared three scribble distribution patterns: (1) center-biased scribbles, sampled from the inner 30% region away from the organ boundary; (2) boundary-biased scribbles, sampled from the 10% region near the organ boundary; and (3) random scribbles, generated using the B-spline sampling method in the original paper. We kept the annotation density fixed and only changed the spatial distribution. As shown in Table 11, our method remains stable across all patterns. On BTCV and WORD, center/boundary-biased scribbles incur only marginal performance drops (0.42%–0.76% DSC) relative to random sampling, with minor 95HD variations. These results confirm that our reliability-aware feedback mechanism effectively adapts to diverse annotation patterns, demonstrating robustness without relying on specific spatial assumptions.

Table 11
Comparison of segmentation performance under different scribble spatial distribution modes.

Pseudo label generation	BTCV		WORD	
	DSC	95HD	DSC	95HD
Random scribbles (Ours)	$\mathbf{78.54}_{13.96}$	$\mathbf{8.08}_{7.21}$	$\mathbf{78.45}_{15.84}$	$\mathbf{7.45}_{10.93}$
Center-biased scribbles	$78.12_{14.23}$	$8.35_{7.45}$	$78.03_{16.12}$	$7.68_{11.27}$
Boundary-biased scribbles	$77.86_{14.58}$	$8.62_{7.82}$	$77.69_{16.45}$	$7.94_{11.68}$

Table 12
Performance comparison under different noise types and levels, showing the relative performance drop of pCE, DMSPS and Ours methods on the BTCV and WORD dataset.

Noise Type	Level	BTCV (ΔDSC)			WORD (ΔDSC)		
		pCE	DMSPS	Ours	pCE	DMSPS	Ours
Gaussian	$\sigma = 0.05$	−12.3%	−5.2%	−3.1%	−14.1%	−6.7%	−4.2%
	$\sigma = 0.1$	−18.6%	−9.8%	−6.4%	−20.3%	−11.5%	−7.9%
Impulse	5%	−9.8%	−4.1%	−2.3%	−11.5%	−5.6%	−3.4%
	10%	−15.2%	−7.9%	−4.7%	−17.8%	−9.1%	−5.8%
Motion Blur	$k = 5$	−9.8%	−4.1%	−2.3%	−9.2%	−4.5%	−2.6%
	$k = 9$	−13.7%	−6.2%	−3.8%	−15.9%	−7.8%	−4.9%

4.7.3. Noise robustness analysis

Our framework demonstrates superior resilience against various noise corruptions compared to baseline methods, as evidenced by Table 12. Under Gaussian noise ($\sigma = 0.1$), our method incurs only 6.4% DSC degradation on BTCV versus pCE's 18.6% drop. For impulse noise (10% corruption), performance decline remains <5%, outperforming DMSPS by 3.2%, while motion blur ($k = 9$) causes merely 3.8% reduction compared to competitors' >13% drops. This robustness stems from our shared-decoder architecture and reliability feedback mechanism, which synergistically filter high-frequency artifacts via distribution-aware feature modeling while preserving anatomical structures.

4.7.4. Performance stability and computational efficiency analysis

To assess practical deployment feasibility, we evaluated performance stability and computational efficiency against existing methods. Across five independent runs with 80% random test-set sampling, we analyzed variations in DSC and 95HD. As shown in Table 13, our method achieves minimal standard deviations and consistently outperforms baselines in both accuracy and consistency. This robustness to sampling variations underscores the clinical readiness of our reliability-aware framework, where stable performance is critical.

Regarding computational efficiency, we evaluated training time (1k iterations) and GPU memory on an NVIDIA RTX 3090 (patch size 96^3, batch size 2). As shown in Table 13, our framework requires 401.1 s and 13.5 GB, exceeding pCE (243.9 s, 9.6 GB) and fully supervised V-Net (258.4 s, 9.8 GB), primarily due to distribution modeling, Mahalanobis distance computation, and the feedback mechanism. Nevertheless, inference remains efficient at 0.42 s/volume, comparable to other 3D segmentation methods. We deem this trade-off acceptable given the notable performance gains, especially for small and clinically relevant anatomical structures. Future work will target reducing computational overhead while preserving segmentation quality.

4.8. Feature distribution analysis

We quantitatively assess representation quality through statistical testing, t-SNE visualization, and comparative modeling. The Henze–Zirkler test (Table 14) shows our method yields significantly lower HZ statistics and higher p-values than baselines, indicating stronger alignment with multivariate Gaussian distributions. Rather than assuming strict unimodality, this Gaussian formulation serves as a tractable, covariance-aware proxy for feature compactness and spatial reliability. This is visually confirmed in Fig. 8, where our approach produces compact intra-class clusters with clear 95% confidence ellipses and enhanced inter-class margins, contrasting sharply with the dispersed features of pCE and the baseline. Comparisons with flexible alternatives like GMM and KDE (Table 15) reveal that GMM offers only marginal DSC gains ($\leq$0.25%, $p > 0.05$) at substantial computational cost. Consequently, the diagonal Gaussian approximation optimally balances distributional expressiveness and training efficiency, proving highly effective for reliability estimation under sparse scribble supervision.

5. Discussion

This work presents a reliability-aware learning framework for scribble-supervised 3D multi-organ segmentation, aiming to address the fundamental challenge of representation ambiguity induced by extremely sparse supervision. Rather than relying on heuristic uncertainty measures or purely prediction-level confidence, the proposed framework explicitly models feature reliability in the embedding space and leverages it to guide adaptive representation refinement.

From a broader perspective, this work differs fundamentally from existing uncertainty-aware or consistency-based approaches. While individual components such as distribution modeling and pseudo-label consistency have been explored separately, our framework does not rely on heuristic stacking. Instead, it establishes a loop reliability feedback mechanism that tightly couples feature-space estimation with adaptive training dynamics. First, the reliability modeling is grounded in embedding-space distribution characteristics rather than output probability dispersion, enabling more precise identification of representation failures at an earlier stage of the learning process. By operating directly on embeddings, the framework can capture subtle semantic inconsistencies that are often obscured at the prediction level. Second, reliability is not merely used as an auxiliary diagnostic signal or post hoc weighting factor, but is actively fed back as a conditioning input to modulate the optimization trajectory. This design allows the model to progressively reallocate learning emphasis toward unreliable or ambiguous regions as training evolves, aligning optimization behavior with representational competence. This synergistic integration is empirically validated by the strong reliability-error correlation in unlabeled regions (Table 6). This demonstrates that the coupled formulation yields a robust quality proxy, a capability unattainable by applying these components independently or in standard semi-supervised pipelines. As a result, the framework promotes more stable convergence and contributes to the observed improvements in generalization performance and robustness under extremely sparse supervision. To clearly position our framework relative to existing uncertainty-aware methods, we summarize the key distinctions between our approach and existing uncertainty modeling methods in Table 16. As shown, while conventional methods estimate uncertainty from prediction outputs (e.g., variance, entropy, or evidential distributions), they not only fail to effectively distinguish between inherent anatomical ambiguity and representation failure due to under-supervision but also predominantly treat uncertainty as an auxiliary post-hoc signal rather than integrating it into training dynamics.

Despite these advantages, several limitations remain. (1) Although we approximate covariance matrices as diagonal to circumvent the complexity of full Mahalanobis inversion, computing class-wise statistics and distances still introduces non-negligible training overhead, which may affect scalability to very large datasets or high-resolution volumes. (2) Our reliability modeling explicitly focuses on foreground classes and excludes the background, as heterogeneous anatomical contexts and acquisition artifacts are poorly characterized by unimodal distributions. While the background remains supervised through the standard segmentation loss, explicit background reliability estimation could further enhance robustness in artifact-prone regions. (3) The feedback mechanism's sensitivity to hyperparameter selection (e.g., ϕ, ψ) requires careful tuning across different annotation densities, and its extension to other weak supervision settings (e.g., point- or box-based annotations) warrants further investigation.

6. Conclusion

This work proposes a reliability-aware learning framework to address the challenge of insufficient supervision in scribble-supervised 3D multi-organ segmentation. By explicitly modeling class-wise feature distributions in the embedding space and leveraging Mahalanobis distance to estimate feature reliability, the proposed framework enables adaptive emphasis on under-supervised and ambiguous regions during

Table 13
Performance stability (DSC and 95HD) and computational efficiency comparison across methods on BTCV and WORD. Reported values are mean with standard deviation (subscript), where DSC and 95HD statistics are computed over five runs with random 80% test set sampling. Training time and GPU memory are averaged over five independent runs.

Methods	BTCV		WORD		Training time	GPU Memory
	DSC	95HD	DSC	95HD	(s/1k-iter)	(GB)
V-Net	$82.82_{0.40}$	$6.77_{1.18}$	$85.14_{0.62}$	$7.30_{0.62}$	$258.4_{9.2}$	9.8
pCE [39]	$50.71_{0.47}$	$61.17_{2.85}$	$59.92_{0.40}$	$50.89_{0.49}$	$243.9_{10.6}$	9.6
USTM [16]	$70.82_{0.48}$	$37.41_{1.70}$	$64.85_{0.36}$	$30.54_{0.52}$	$301.8_{10.4}$	10.5
DMSPS [44]	$74.69_{0.47}$	$15.85_{0.44}$	$74.26_{0.31}$	$10.46_{0.44}$	$326.6_{11.7}$	11.3
Ours	$78.62_{0.39}$	$7.94_{0.33}$	$78.68_{0.27}$	$7.42_{0.31}$	$401.1_{15.8}$	13.5

Table 14
Henze–Zirkler multivariate normality test results for feature distributions on BTCV and WORD datasets. Lower HZ statistic and higher p-value indicate closer approximation to multivariate Gaussian distribution. Results are averaged over 5 independent runs. HZ: Henze–Zirkler statistic. p-value: significance level for multivariate normality test.

Dataset	Method	Statistic	Organ Classes						$\text{Mean}_{\pm\text{std}}$
			Liver	Spleen	L.Kid	Gall	Eso	Pan	
BTCV	Baseline	HZ	2.81	3.14	2.93	4.23	5.08	4.76	$3.82_{0.85}$
		p-value	0.121	0.093	0.108	0.042	0.021	0.034	
	Ours	HZ	1.23	1.41	1.32	1.89	2.18	2.03	$1.68_{0.38}$
		p-value	0.318	0.284	0.301	0.048	0.039	0.062	
WORD	Baseline	HZ	2.63	2.91	2.74	3.87	4.79	4.52	$3.58_{0.81}$
		p-value	0.138	0.104	0.119	0.051	0.028	0.039	
	Ours	HZ	1.14	1.32	1.21	1.73	2.01	1.86	$1.54_{0.33}$
		p-value	0.347	0.312	0.329	0.053	0.044	0.071	

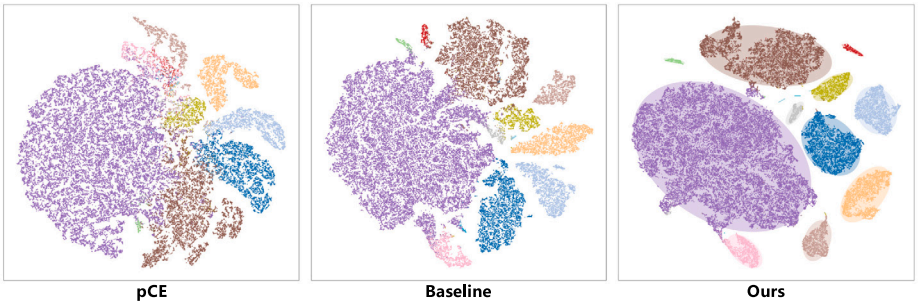

Fig. 8. t-SNE visualization of feature embeddings learned by different methods on the BTCV dataset. From left to right: pCE, our baseline (w/o reliability-aware components), and our full framework. Class-conditional 95% confidence ellipses are overlaid for our method to illustrate the estimated feature distributions. Compared with pCE and the baseline, our approach yields significantly more compact intra-class clusters and better-separated inter-class boundaries.

Table 15
Ablation study on feature distribution modeling strategies. We compare our Gaussian model with more flexible models (GMM, KDE). Results are reported as $\text{mean}_{\pm\text{std}}$ over 5 runs.

Distribution model	BTCV		WORD		Train time (×)
	DSC	95HD	DSC	95HD	
Ours	$78.54_{0.39}$	$8.08_{0.33}$	$78.45_{0.27}$	$7.45_{0.31}$	1.0 ×
GMM ($K = 2$)	$78.71_{0.42}$*	$8.01_{0.38}$*	$78.63_{0.31}$*	$7.39_{0.35}$*	1.8 ×
GMM ($K = 3$)	$78.79_{0.45}$*	$7.96_{0.41}$*	$78.72_{0.34}$*	$7.33_{0.39}$*	2.5 ×
KDE (bandwidth = 0.5)	$78.32_{0.67}$*	$8.24_{0.52}$*	$78.21_{0.58}$*	$7.68_{0.49}$*	3.2 ×

* No significant difference from our method (paired t-test, $p > 0.05$).

Table 16
Comparison of uncertainty modeling approaches. (✓: fully supported; △: partially supported/conditionally effective; ✗: not supported).

Dimension	MC Dropout or Ensemble	Entropy-based	EDL	Ours
Uncertainty Source	Prediction variance	Probability distribution	Higher-order distribution	Embedding feature distribution
Distinguish ambiguity vs. failure	✗	✗	△	✓
Integration in training dynamics	✗	✗	△	✓
Feature-space modeling	✗	✗	△	✓
Suitability for sparse supervision	△	✗	△	✓

training. This design has improved feature compactness and inter-class separability, leading to consistent performance gains, particularly for small and structurally complex organs, while maintaining strong robustness across datasets and noise conditions.

CRediT authorship contribution statement

Xiaoxiang Han: Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Yiman Liu:** Writing – original draft, Validation, Investigation, Data curation, Conceptualization. **Jiang Shang:** Validation, Software, Methodology, Investigation, Formal analysis. **Yunhao Bai:** Visualization, Validation, Methodology, Formal analysis. **Yan Wang:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization. **Qi Zhang:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [No. 62571309], the Natural Science Foundation of Shanghai Municipality [No. 25ZR1401135], the Fundamental Research Funds for the Central Universities [No. YG2025QNA07], the Pudong New Area Science and Technology Development Fund [No. PKJ2025Y04], and Shanghai Technical Service Center of Science and Engineering Computing, Shanghai University.

Data availability

Data will be made available on request.

References

[1] X. Luo, M. Hu, W. Liao, S. Zhai, T. Song, G. Wang, S. Zhang, Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels supervision, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 528–538.

[2] H. Lee, W.-K. Jeong, Scribble2label: Scribble-supervised cell segmentation via self-generating pseudo-labels with consistency, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23, Springer, 2020, pp. 14–23.

[3] Y. Li, L. Luo, H. Lin, H. Chen, P.-A. Heng, Dual-consistency semi-supervised learning with uncertainty quantification for covid-19 lesion segmentation from ct images, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24, Springer, 2021, pp. 199–209.

[4] W. Lu, D. Gong, K. Fu, X. Sun, W. Diao, L. Liu, Boundarymix: Generating pseudo-training images for improving segmentation with scribble annotations, Pattern Recognit. 117 (2021) 107924.

[5] M. Zhou, Z. Xu, K. Zhou, R.K. y. Tong, Weakly supervised medical image segmentation via superpixel-guided scribble walking and class-wise contrastive regularization, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 137–147.

[6] A. Obukhov, S. Georgoulis, D. Dai, L. Van Gool, Gated crf loss for weakly supervised semantic image segmentation, 2019, arXiv preprint arXiv:1906.04651.

[7] M. Tang, F. Perazzi, A. Djelouah, I. Ben Ayed, C. Schroers, Y. Boykov, On regularized losses for weakly-supervised cnn segmentation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 507–522.

[8] B. Kim, J.C. Ye, Mumford–shah loss functional for image segmentation with deep learning, IEEE Trans. Image Process. 29 (2019) 1856–1866.

[9] F. Gao, M. Hu, M.-E. Zhong, S. Feng, X. Tian, X. Meng, Z. Huang, M. Lv, T. Song, X. Zhang, et al., Segmentation only uses sparse annotations: Unified weakly and semi-supervised learning in medical images, Med. Image Anal. 80 (2022) 102515.

[10] Q. Chen, Y. Hong, Scribble2d5: Weakly-supervised volumetric image segmentation via scribble annotations, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 234–243.

[11] K. Zhang, X. Zhuang, Cyclemix: A holistic strategy for medical image segmentation from scribble supervision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11656–11665.

[12] D. Chen, Y. Bai, W. Shen, Q. Li, L. Yu, Y. Wang, Magicnet: Semi-supervised multi-organ segmentation via magic-cube partition and recovery, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 23869–23878.

[13] Z. Li, Y. Zheng, D. Shan, S. Yang, Q. Li, B. Wang, Y. Zhang, Q. Hong, D. Shen, Scribformer: Transformer makes cnn work better for scribble-based medical image segmentation, IEEE Trans. Med. Imaging (2024).

[14] B. Zhang, J. Xiao, Y. Zhao, Dynamic feature regularized loss for weakly supervised semantic segmentation, Pattern Recognit. 164 (2025) 111540.

[15] X. Han, X. Li, J. Shang, Y. Liu, K. Chen, S. Xu, Q. Liu, Q. Zhang, Mambaeviscrib: Mamba and evidence-guided consistency enhance cnn robustness for scribble-based weakly supervised ultrasound image segmentation, Inf. Fusion 126 (2026) 103590.

[16] X. Liu, Q. Yuan, Y. Gao, K. He, S. Wang, X. Tang, J. Tang, D. Shen, Weakly supervised segmentation of covid19 infection with scribble annotation on ct images, Pattern Recognit. 122 (2022) 108341.

[17] Y. Huang, S. Li, Z. Guo, Q. Mei, Z. Han, X. Wang, H. Wang, Boundary feature alignment for semi-supervised medical image segmentation, Pattern Recognit. (2025) 111946.

[18] N. Tishby, N. Zaslavsky, Deep learning and the information bottleneck principle, in: 2015 Ieee Information Theory Workshop (Itw), Ieee, 2015, pp. 1–5.

[19] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, A. Klein, Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge, in: Proc. MICCAI Multi-Atlas Labeling beyond Cranial Vault—Workshop Challenge, Vol. 5, 2015, p. 12.

[20] X. Luo, W. Liao, J. Xiao, J. Chen, T. Song, X. Zhang, K. Li, D.N. Metaxas, G. Wang, S. Zhang, Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image, Med. Image Anal. 82 (2022) 102642.

[21] L. Grady, Random walks for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 28 (11) (2006) 1768–1783.

[22] V. Vezhnevets, V. Konouchine, Growcut: Interactive multi-label nd image segmentation by cellular automata, in: Proc. of Graphicon, Vol. 1, Citeseer, 2005, pp. 150–156.

[23] X. Han, Y. Liu, J. Shang, H. Chen, X. Liu, Z. Qiu, Y. Wang, Q. Zhang, Scribble-supervised multi-organ segmentation via epistemic-driven hardness-adaptive focusing, IEEE Trans. Med. Imaging (2026).

[24] X. Chen, Y. Yuan, G. Zeng, J. Wang, Semi-supervised semantic segmentation with cross pseudo supervision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2613–2622.

[25] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: International Conference on Machine Learning, PMLR, 2016, pp. 1050–1059.

[26] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, Adv. Neural Inf. Process. Syst. 30 (2017).

[27] K. Lee, K. Lee, H. Lee, J. Shin, A simple unified framework for detecting out-of-distribution samples and adversarial attacks, Adv. Neural Inf. Process. Syst. 31 (2018).

[28] C. González, K. Gotkowski, M. Fuchs, A. Bucher, A. Dadras, R. Fischbach, I.J. Kaltenborn, A. Mukhopadhyay, Distance-based detection of out-of-distribution silent failures for covid-19 lung lesion segmentation, Med. Image Anal. 82 (2022) 102596.

[29] M. Sensoy, L. Kaplan, M. Kandemir, Evidential deep learning to quantify classification uncertainty, Adv. Neural Inf. Process. Syst. 31 (2018).

[30] D. Duvenaud, J. Wang, J. Jacobsen, K. Swersky, M. Norouzi, W. Grathwohl, Your classifier is secretly an energy based model and you should treat it like one, in: ICLR 2020, 2020.

[31] A.v.d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, 2018, arXiv preprint arXiv:1807.03748.

[32] A.H. Abdelaziz, S. Watanabe, J.R. Hershey, E. Vincent, D. Kolossa, Uncertainty propagation through deep neural networks, in: Interspeech 2015, 2015.

[33] Y. Lin, D. Nie, Y. Liu, M. Yang, D. Zhang, X. Wen, Multi-target domain adaptation with prompt learning for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 717–727.

[34] A. Chang, X. Tao, X. Yang, Y. Huang, X. Zhou, J. Zeng, R. Huang, D. Ni, Pe-med: Prompt enhancement for interactive medical image segmentation, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2023, pp. 257–266.

[35] J. Ma, Y. He, F. Li, L. Han, C. You, B. Wang, Segment anything in medical images, Nat. Commun. 15 (1) (2024) 654.

[36] Q. Zeng, Y. Xie, Z. Lu, M. Lu, J. Zhang, Y. Zhou, Y. Xia, Consistency-guided differential decoding for enhancing semi-supervised medical image segmentation, IEEE Trans. Med. Imaging (2024).

[37] M. Gadermayr, L. Gupta, V. Appel, P. Boor, B.M. Klinkhammer, D. Merhof, Generative adversarial networks for facilitating stain-independent supervised and unsupervised segmentation: a study on kidney histology, IEEE Trans. Med. Imaging 38 (10) (2019) 2293–2302.

[38] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision (3DV), Ieee, 2016, pp. 565–571.

[39] M. Tang, A. Djelouah, F. Perazzi, Y. Boykov, C. Schroers, Normalized cut loss for weakly-supervised cnn segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1818–1827.

[40] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H.R. Roth, D. Xu, Unetr: Transformers for 3d medical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 574–584.

[41] M. Javanmardi, M. Sajjadi, T. Liu, T. Tasdizen, Unsupervised total variation loss for semi-supervised deep learning of semantic segmentation, 2016, arXiv preprint arXiv:1605.01368.

[42] S. Yu, B. Zhang, J. Xiao, E.G. Lim, Structure-consistent weakly supervised salient object detection with local saliency coherence, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 3234–3242.

[43] J. Liu, S.Y. Tan, X. Yang, Y. Xu, S.Y. Yeo, Effdnet: A scribble-supervised medical image segmentation method with enhanced foreground feature discrimination, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2025, pp. 194–204.

[44] M. Han, X. Luo, X. Xie, W. Liao, S. Zhang, T. Song, G. Wang, S. Zhang, Dmsps: Dynamically mixed soft pseudo-label supervision for scribble-supervised medical image segmentation, Med. Image Anal. 97 (2024) 103274.