Scribble-Supervised Multi-Organ Segmentation via Epistemic-Driven Hardness-Adaptive Focusing

Xiaoxiang Han, Yiman Liu, Jiang Shang, Haobo Chen, Xiaohong Liu, Zhen Qiu, Yan Wang, and Qi Zhang

***Abstract*— Scribble supervision reduces annotation costs in multi-organ segmentation. However, its sparsity results in insufficient supervision for most regions and inadequate feature learning in hard areas (e.g., organ boundaries). These hard areas cause model confirmation bias and high epistemic uncertainty, which existing methods fail to address. To overcome these core challenges, we propose an epistemic-driven hardness-adaptive focusing framework. This framework establishes a self-improving loop: quantified epistemic uncertainty guides hard sample generation, while hard sample learning and feature alignment jointly reduce epistemic uncertainty. Specifically, we first propose a phase-adaptive hardness-aware loss function to quantify epistemic uncertainty and generate dynamic hardness maps during training. Based on these maps, we employ a distribution-divergence-aware copy-paste operation to create hard samples, which are progressively incorporated into learning to reduce epistemic uncertainty. Furthermore, we introduce feature distribution alignment to mitigate bias and epistemic uncertainty by aligning organ-specific hard regions with global features. Extensive experiments on multi-organ CT and ultrasound datasets demonstrate the competitiveness and effectiveness of our method. The framework's generalizability and robustness are further validated under cross-dataset and noise-corrupted scenarios. This work offers a practical solution for clinical applications where annotation efficiency is critical.**

***Index Terms*— Medical image segmentation, weakly-supervised learning, scribble annotation, epistemic uncertainty, data augmentation.**

Received 4 January 2026; revised 3 February 2026; accepted 13 February 2026. Date of publication 17 February 2026; date of current version 3 June 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62571309; in part by the Natural Science Foundation of Shanghai Municipality under Grant 25ZR1401135; in part by the Fundamental Research Funds for the Central Universities under Grant YG2025QNA07; in part by the Pudong New Area Science and Technology Development Fund under Grant PKJ2025-Y04; and in part by Shanghai Technical Service Center of Science and Engineering Computing, Shanghai University. Recommended by Associate Editor H. Rivaz. *(Xiaoxiang Han and Yiman Liu are co-first authors.) (Corresponding authors: Qi Zhang; Yan Wang.)*

Xiaoxiang Han, Jiang Shang, Haobo Chen, and Qi Zhang are with the Smart Medicine and AI-Based Radiology Technology (SMART) Laboratory, School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: hanxx@shu.edu.cn; jiangshang@shu.edu.cn; haobochen@shu.edu.cn; zhangq@t.shu.edu.cn).

Yiman Liu is with the Department of Pediatric Cardiology, Shanghai Children's Medical Center, Shanghai Jiao Tong University School of Medicine, Shanghai 200127, China (e-mail: liuyiman@scmc.com.cn).

Xiaohong Liu is with the John Hopcroft Center, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: xiaohongliu@sjtu.edu.cn).

Zhen Qiu is with the Department of Biomedical Engineering, University of Strathclyde, G4 0NW Glasgow, U.K. (e-mail: z.qiu@strath.ac.uk).

Yan Wang is with Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai 200241, China (e-mail: ywang@cee.ecnu.edu.cn).

Code is available on-line at https://github.com/GtLinyer/EDHAF

Digital Object Identifier 10.1109/TMI.2026.3665556

I. Introduction

MULTI-ORGAN imaging (e.g., CT and ultrasound) is essential for clinical decision-making, enabling non-invasive structural and pathological evaluation across anatomical systems. Consequently, multi-organ segmentation serves as a basis for computer-assisted interventions [1]. Deep learning techniques have greatly advanced medical image segmentation. However, traditional supervised learning methods often require extensive pixel-wise dense annotations to train an accurate multi-organ segmentation model. Since some medical images (e.g., CT) are three-dimensional, pixel-wise annotation is extremely time-consuming and costly [2]. Scribble annotation, due to its convenience and flexibility, is a favored efficient annotation method. Consequently, scribble-supervised learning has emerged as a popular weakly supervised learning approach that does not rely on dense annotation data [3], [4].

Given that the scribble region only represents a minute subset of the region of interest, leading to insufficient supervisory signals and posing challenges to model training, as shown in Fig. 1. Thus, how to utilize the information from unlabeled pixels has become the focus of scribble-supervised learning methods. Existing scribble-supervised methods can be broadly categorized into three paradigms [Fig. 2 (1–3)]: (1) Traditional optimization-based methods (e.g., Random Walker [5], GrowCut [6]) leverage graph structures to propagate scribble labels but struggle with capturing high-dimensional features and often yield suboptimal performance on complex anatomies. (2) Pseudo-label generation methods [3], [7] employ iterative

1558-254X © 2026 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on June 20,2026 at 13:01:41 UTC from IEEE Xplore. Restrictions apply.

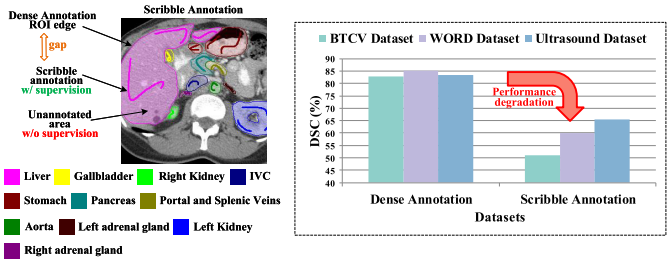

Fig. 1. Challenges of scribble annotation. Compared to dense annotation, scribbles cover only a small fraction of the region of interest (ROI), posing significant difficulties for model training and leading to a substantial performance degradation.

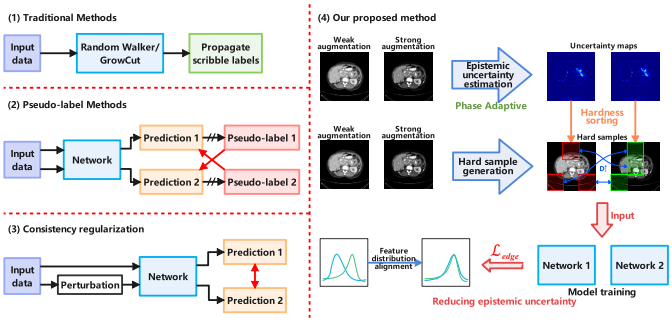

Fig. 2. Comparison between our method and existing approaches. Illustration of (1) pseudo-labeling methods, (2) consistency regularization methods, (3) scribble annotations covering only part of the region of interest, and (4) our proposed method, which forms a self-improving loop where quantified epistemic uncertainty guides hard-sample generation, while hard-sample learning and feature alignment jointly reduce uncertainty.

self-training to expand supervision; however, they suffer from confirmation bias due to noisy pseudo-labels in unannotated regions, especially in early training phases. (3) Consistency regularization is a popular and effective strategy. Based on the smoothness assumption, this strategy generates a sample that differs slightly from the input sample but is semantically close through perturbation, thereby avoiding the challenging task of directly calculating distances between different samples. These methods [8], [9] enforce prediction invariance under perturbations (e.g., input transformations [10], feature [11] and network [12] perturbations). However, these data augmentation-based approaches often overlook epistemic uncertainty, failing to effectively leverage it to reduce model bias.

Discriminative features of challenging anatomical regions (e.g., small organs or edges) often remain under-captured due to the model's incomplete understanding of data distributions, reflecting the model's epistemic uncertainty. This differs from data's aleatoric uncertainty and can be mitigated through enhanced learning and diversified data distribution, which current methods overlook. This raises two fundamental questions: (1) How to effectively estimate epistemic uncertainty during training? (2) How to construct diverse hard samples based on uncertainty maps to guide the model's focus on learning challenging regions? Furthermore, challenging regions are more vulnerable to shifts in feature distributions under strong disturbances, leading to learning biases. This raises another key question: (3) How can the bias between local hard feature distributions and the global distribution be reduced?

To address these core challenges, we propose an epistemic-driven hardness-adaptive focusing framework. This framework explicitly targets the estimation, exploitation, and reduction of epistemic uncertainty through three synergistic innovations operating in a coarse-to-fine manner [Fig. 2 (4)]: (1) Phase-adaptive hardness quantification: We develop a novel loss function that explicitly quantifies epistemic uncertainty using Dempster-Shafer Theory, dynamically generating hardness maps tailored to the evolving needs of different training phases. (2) Distributionally divergent hard sample augmentation: Guided by the hardness maps, we strategically construct challenging training samples. By partitioning volumes into small cubes and leveraging Copy-Paste augmentation, we explicitly overlay high-hardness regions from one volume onto low-hardness regions exhibiting high distributional divergence in another volume, maximizing exposure to diverse yet challenging features. (3) Feature distribution alignment: To refine ambiguous boundaries where sparse scribbles offer little guidance, we leverage global organ feature distributions as stable anchors. Specifically, we align the feature distributions of identified hard regions (e.g., organ boundaries) with these global priors, mitigating local feature bias and enhancing anatomical consistency. These components form a synergistic loop: quantified uncertainty guides the generation of augmented hard samples, which, together with the feature alignment mechanism, collaboratively work to reduce epistemic uncertainty throughout the training process. This end-to-end framework effectively bridges the performance gap inherent in learning from sparse scribble annotations.

Extensive experiments were conducted on two 3D multi-organ CT datasets and three 2D multi-organ ultrasound datasets. **Note** that our model employs V-Net [13] (3D)/U-Net [14] (2D) without any modifications.

The main contributions of this paper are as follows:

1) We introduce an epistemic-driven hardness-adaptive focusing framework that quantifies model uncertainty via Dempster-Shafer theory to dynamically generate hardness maps and guide targeted learning on hard regions.
2) We propose a distribution-divergence-aware copy-paste augmentation that strategically overlays high-uncertainty cubes onto low-uncertainty, high-divergence locations, creating diverse hard samples.
3) We present a feature-distribution alignment loss that aligns hard boundary features with global organ priors and enforces inter-organ separation, reducing local bias and epistemic uncertainty.

II. Related Work

A. Scribble-Supervised Segmentation

Scribble-supervised learning is a form of weakly supervised learning, in which scribbles serve as a sparse annotation technique by casually drawing curves over regions of interest. In earlier years, researchers attempted to address the challenges of scribble-supervised segmentation using traditional optimization-based methods, such as Random Walker [5] and GrowCut [6]. In the era of deep learning, various approaches

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on June 20,2026 at 13:01:41 UTC from IEEE Xplore. Restrictions apply.

have been proposed, including graphical-based methods [4], conditional random field (CRF)-based methods [15], level set-based methods [16], etc. Some methods utilize auxiliary tasks for learning, such as Geodesic distance map [17] and super-pixel generation [18]. Some methods generate pseudo-labels, for instance, Zhou et al. [19] propose a warm-up strategy and label filtering to produce higher-quality pseudo-labels. Consistency regularization is another popular approach. For instance, Liu et al. [8] introduced transformation consistency and uncertainty awareness. Han et al. [20] generated feature-level perturbations between different networks by utilizing varying dilation rates in convolutions, and ensured consistency in their predictions. However, existing methods have neither considered nor leveraged hard samples effectively. Additionally, unlike interactive methods such as ScribblePrompt [21] that use scribbles as real-time prompts, our approach employs sparse scribbles solely as weak supervision during training to reduce annotation costs for clinical deployment, without human-in-the-loop interaction.

B. Data Augmentation

Input perturbations in consistency regularization paradigms are generally based on data augmentation. Common data augmentation techniques, such as rotation, flipping, transformation, and noise addition [8], maintain the same image label after transformation. The interpolation-based methods, however, are different as they generate new and diverse samples, such as Mixup [22], CutMix, CutOut [23], ClassMix [24], PuzzleMix [25], etc. CycleMix combines CutMix and CutOut to perform increments and decrements of scribbles [10]. MagicNet treats 3D multi-organ CT volumes as magic cubes and proposes a data augmentation strategy akin to solving a magic cube [26]. However, most existing data augmentation techniques are random, failing to consider the ease of learning of the samples and thus unable to further enhance sample diversity.

C. Epistemic Uncertainty

Epistemic uncertainty estimation provides insights into the model's mastery of sample features. In deep learning, Bayesian neural networks are used to compute the posterior distribution over parameters on training samples to estimate uncertainty. To simplify the problem, several approximate methods have been proposed. For instance, Monte Carlo Dropout [27] employs dropout randomly during inference to approximate Bayesian posterior sampling, while conditional variational autoencoders [28] encode data variability into the latent space using a Bayesian paradigm, providing a structured approach for uncertainty estimation. Additionally, ensemble-based methods [29] derive uncertainty estimates from the collective outputs of multiple models, but they are computationally expensive. There are also methods based on data augmentation [30]. However, these aforementioned methods typically incur high computational costs. Evidence deep learning (EDL) based on Dempster-Shafer theory has gained popularity recently due to its low computational cost, easy training, and plug-and-play capability [31], [32]. Yet, uncertainty estimation methods specifically designed for scribble annotations remain lacking.

III. Methods

A. Preliminary

1) Evidential Deep Learning: To estimate sample hardness for constructing hard samples, we explicitly model the model's epistemic uncertainty, whereas traditional *softmax* only captures aleatoric uncertainty [30]. Thereby, we extend EDL [31], grounded in Dempster-Shafer Theory and Subjective Logic. EDL reformulates classification as evidence modeling by parameterizing Dirichlet distributions through neural network outputs. Given logits z_k for class k, evidence e_k is computed as $e_k = softplus(z_k)$. The total evidence strength $S = \sum_{i=1}^{K} e_i + 1$ determines belief masses $b_k = \frac{e_k}{S}$ and uncertainty $u = \frac{K}{S}$. The Dirichlet distribution $\mathrm{Dir}(\mathbf{p}|\alpha)$ is parameterized by $\alpha_k = e_k + 1$, with expected class probability $\mathbb{E}[p_k] = \frac{\alpha_k}{S}$. The classification loss minimizes cross-entropy between expected probabilities and ground-truth labels:

$$\mathcal{L}_{cls} = \mathbb{E}_{p\sim \mathrm{Dir}(\alpha)}\left[-\sum_{k=1}^{K} y_k \log p_k\right] = \sum_{k=1}^{K} y_k[\psi(S) - \psi(\alpha_k)], \tag{1}$$

where $\psi(\cdot)$ is *digamma* function.

2) Mahalanobis Distance: The Mahalanobis Distance is a statistical measure of similarity between a point and a distribution, addressing the limitations of Euclidean distance in correlated features by incorporating covariance structure. Given a feature vector $z \in \mathbb{R}^d$ and a dataset following a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ (where μ is the mean vector and Σ is the covariance matrix), the Mahalanobis Distance is computed as:

$$\mathrm{MD}\{z, \mathcal{N}(\mu, \Sigma)\} = \sqrt{(z-\mu)^T \Sigma^{-1} (z-\mu)}, \tag{2}$$

B. Overview

The 3D volume of a CT scan can be defined as $\mathbf{X} \in \mathbb{R}^{L\times H\times W}$, where three dimensions are represented by W, H, and L for width, height, and length, respectively. The scribble label for $\mathbf{X}$ can be defined as $\mathbf{Y} \in \{0, 1, \ldots, K\}^{L\times H\times W}$, where 0 represents the background class, 1 to $K-1$ indicate anatomical structures, and K denotes unlabeled pixels. They constitute a pair of samples, and the dataset $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{N}$ comprises N such pairs. First, the input sample $\mathbf{X}$ is processed through weak and strong augmentations to produce $\mathbf{X}^w$ and $\mathbf{X}^s$, where hereafter w and s denote the weak- and strong-augmentation branches. Weak augmentation employs basic transformations such as rotation and flipping, while strong augmentation includes intensity/contrast adjustments and our epistemic-adaptive unadversarial perturbation (EAUP) (see Sec. III-F). $\mathbf{X}^w$ and $\mathbf{X}^s$ are fed into two separately initialized VNets $\mathcal{F}_w(\cdot; \mathbf{\Theta}_w)$ and $\mathcal{F}_s(\cdot; \mathbf{\Theta}_s)$, generating logits $\mathbf{Z}^w$, $\mathbf{Z}^s \in \mathbb{R}^{C\times L\times H\times W}$, along with penultimate layer feature vectors $\mathbf{H}^w$, $\mathbf{H}^s \in \mathbb{R}^{T\times L\times H\times W}$. Here, $\mathbf{\Theta}_w$ and $\mathbf{\Theta}_s$ represent the learnable parameters of the networks, $C = K-1$ represents the number of classes, and T denotes the length of the feature vector.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on June 20,2026 at 13:01:41 UTC from IEEE Xplore. Restrictions apply.

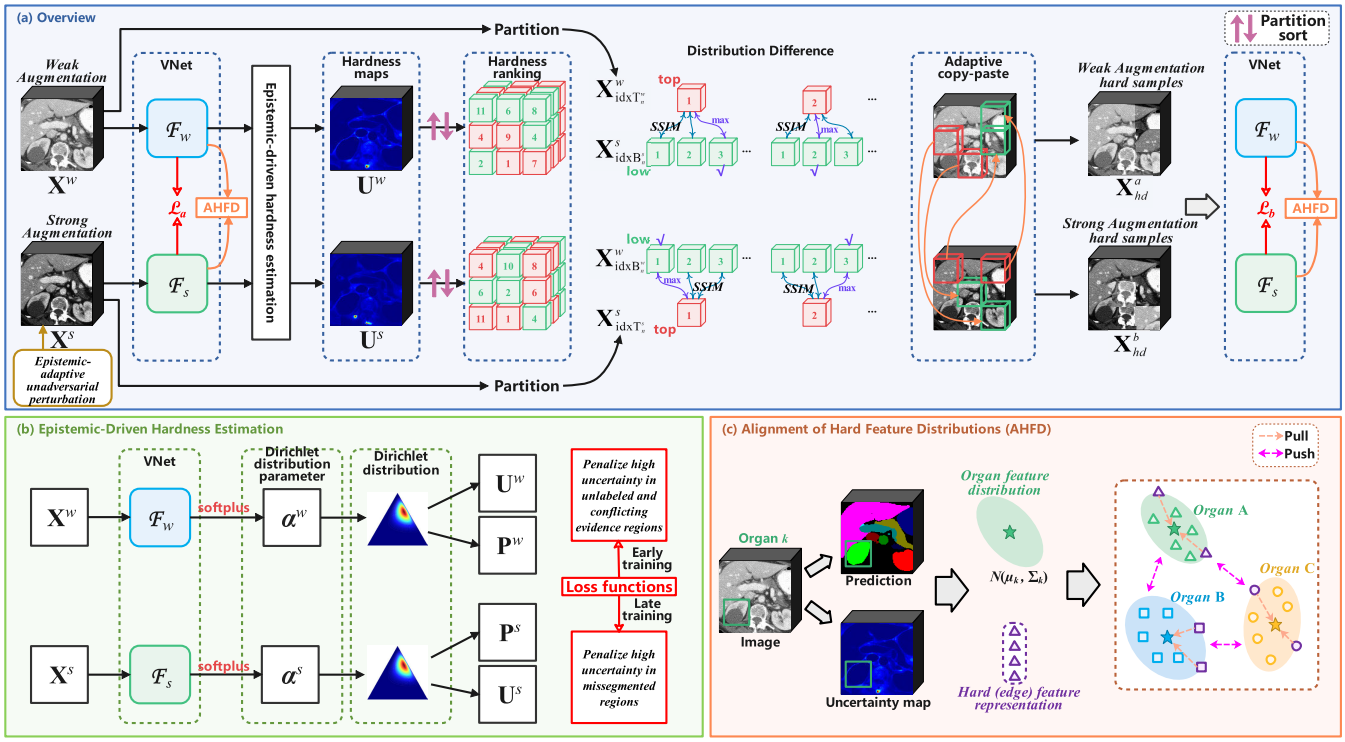

Fig. 3. Illustration of the proposed framework. (a) Overall pipeline with diverse hard-sample construction; (b) Epistemic-driven hardness estimation; (c) Alignment of hard feature distributions.

The outputs are processed by EDL in Sec. III-C to obtain prediction probabilities $\mathbf{P}^w$, $\mathbf{P}^s \in \mathbb{R}^{C\times L\times H\times W}$, along with hardness (epistemic uncertainty) maps $\mathbf{U}^w$, $\mathbf{U}^s \in \mathbb{R}^{L\times H\times W}$. Our proposed framework is shown in Fig. 3: (a) illustrates the pipeline, including construction of diverse hard samples (Sec. III-D); (b) presents epistemic-driven hardness estimation (Sec. III-C); and (c) depicts alignment of hard edge feature distributions (Sec. III-E).

C. Epistemic-Driven Hardness Estimation

Our Epistemic-Driven Hardness Estimation (EDHE) is grounded in EDL and introduces key extensions specifically tailored for scribble-supervised segmentation. Conventional EDL frameworks assume either fully-labeled data or clean unlabeled data with consistent class distributions. In contrast, scribble annotation creates a unique challenge where most pixels lack direct supervision and organ boundaries contain conflicting evidence. Our EDHE addresses this through three key innovations: (1) a conflict-minimization mechanism that distinguishes between genuine epistemic uncertainty and label inconsistency noise; (2) a phase-adaptive uncertainty quantification strategy that evolves training objectives based on learning progress; and (3) a dual-branch consensus mechanism that leverages prediction disagreement to identify regions requiring focused learning. These extensions transform EDL from a passive uncertainty estimator into an active guide for targeted learning under extreme annotation sparsity.

1) Early Training Phase: Loss computation is restricted to scribble-annotated regions ($\mathcal{M}_i = 1$):

$$\mathcal{L}_{p-seg} = \frac{1}{\sum_i \mathcal{M}^i} \sum_i \mathcal{M}^i \mathcal{L}^i_{cls}, \tag{3}$$

where i denotes the pixel index. Thus, the supervised loss is:

$$\mathcal{L}_{sup} = \mathcal{L}^w_{p-seg} + \mathcal{L}^s_{p-seg}. \tag{4}$$

This implies scribble-free regions lack precise ground-truth constraints. To direct hard samples (generated as in Sec. III-D) focus more on these unannotated areas, we encourage high uncertainty in unannotated conflicting regions. Evidence conflict between branches is quantified using Dempster's combination rule, minimized alongside a constraint pushing the Dirichlet distribution toward uniformity (zero evidence):

$$\mathcal{L}_{dis} = \mathcal{M}^u \odot \{\text{Conflict}(\mathbf{P}^w, \mathbf{P}^s) + \eta \cdot \mathcal{M}^{conf} \odot (\mathcal{L}^w_{cons} + \mathcal{L}^s_{cons})\}, \tag{5}$$

where $\mathcal{M}^u$ is the unlabeled region mask, $\text{Conflict} = 1 - \sum_{k=1}^{K}(b^w_k \cdot b^s_k)$ represents the conflict measure in Dempster's combination rule,

$$\mathcal{M}^{conf} = \mathbb{I}[\text{Conflict}(\mathbf{P}^w, \mathbf{P}^s) > \tau], \tag{6}$$

$$\text{and } \mathcal{L}_{cons} = \text{KL}[\text{Dir}(\mathbf{p}|\alpha)\|\text{Dir}(\mathbf{p}|1)]. \tag{7}$$

Additionally, to mitigate confirmation bias in unannotated regions, we align Dirichlet distribution between annotated (Ω^{sc}_k) and unannotated areas within each organ class k. For class k, the mean evidence $\bar{\alpha}^{sc}_k = \frac{1}{|\Omega^{sc}_k|}\sum_{x\in\Omega^{sc}_k} \alpha_k(x)$ is computed over its scribble-annotated area, followed by KL-divergence minimization:

$$\mathcal{L}_{align} = \sum_k \text{KL}[\text{Dir}(\mathbf{p}|\alpha^{usc}_k)\|\text{Dir}(\mathbf{p}|\bar{\alpha}^{sc}_k)]. \tag{8}$$

Thus, the overall alignment loss is:

$$\mathcal{L}^{all}_{align} = \mathcal{L}^w_{align} + \mathcal{L}^s_{align}. \tag{9}$$

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on June 20,2026 at 13:01:41 UTC from IEEE Xplore. Restrictions apply.

2) Late Training Phase: As training progresses, pseudo-label quality improves. We dynamically fuse predictions from both branches, weighted by uncertainty:

$$\widehat{\mathbf{P}} = \varpi^w \cdot \mathbf{P}^w + \varpi^s \cdot \mathbf{P}^s, \tag{10}$$

where $\varpi^w = \frac{\sum \mathbf{U}^s}{\sum \mathbf{U}^w + \sum \mathbf{U}^s}$, $\varpi^s = \frac{\sum \mathbf{U}^w}{\sum \mathbf{U}^w + \sum \mathbf{U}^s}$, and $\varpi^w + \varpi^s = 1.0$. Then, we can obtain pseudo-label $\widehat{\mathbf{Y}} = \text{argmax}(\widehat{\mathbf{P}})$. The pseudo-supervised loss, based on Eq. (1), incorporates uncertainty weighting to prioritize high-hardness regions:

$$\mathcal{L}_{ps} = (\mathbf{W}_u)^\beta \odot [\mathcal{L}_{seg}(\mathbf{P}^w, \widehat{\mathbf{Y}}) + \mathcal{L}_{seg}(\mathbf{P}^s, \widehat{\mathbf{Y}})], \tag{11}$$

where $\beta \geq 1$ is the control factor, $\mathbf{W}_u = \text{Norm}(\mathbf{U}^w + \mathbf{U}^s)$, and $\mathcal{L}_{seg} = \frac{1}{N}\sum_i \mathcal{L}^i_{cls}$. To prevent the model from generating excessive evidence for wrong categories, the Dirichlet distribution is constrained to approach a uniform distribution:

$$\mathcal{L}_{reg} = \lambda_t \cdot \text{KL}[\text{Dir}(\mathbf{p}|\widetilde{\alpha})\|\text{Dir}(\mathbf{p}|1)], \tag{12}$$

where $\lambda_t = \lambda \cdot \exp(-\gamma \overline{S})$ is an adaptive weight adjustment method based on batch uncertainty. Here, $\overline{S} = \frac{1}{B}\sum_{i=1}^{B} S^{(i)}$. It aims to avoid excessive regularization that may disrupt learning. Thus, the overall regularization term is:

$$\mathcal{L}^{all}_{reg} = \frac{1}{2}\left(\mathcal{L}^w_{reg} + \mathcal{L}^s_{reg}\right). \tag{13}$$

3) Geometric Interpretation of KL Term Gradients: Expanding the KL divergence term yields:

$$\log\left(\frac{\Gamma\left(\sum_k \alpha_k\right)}{\Gamma(K)\prod_k \Gamma(\alpha_k)}\right) + \sum_{k=1}^{K}(\alpha_k - 1)\left(\psi(\alpha_k) - \psi\left(\sum_j \alpha_j\right)\right), \tag{14}$$

where $\Gamma(\cdot)$ denotes the *Gamma* function and $\psi(\cdot)$ represents the *digamma* function. When $\alpha_k \approx 1$, using $\psi(1) = -\mathcal{E}$ (where $\mathcal{E}$ is Euler's constant) and $\psi(K) \approx \log K$, the gradient approximates to:

$$\frac{\partial \text{KL}}{\partial \alpha_k} \approx -\mathcal{E} - \log K + \psi(\alpha_k) - \psi\left(\sum_j \alpha_j\right). \tag{15}$$

Since $\psi(\alpha_k)$ is monotonically increasing for $\alpha_k \in (1, +\infty)$, the gradient remains negative, driving α_k towards 1.

D. Construction of Diverse Hard Samples

Unlike conventional data augmentation approaches (e.g., MagicNet [26], CycleMix [10]) that perform random region mixing or cube partitioning, our distribution-divergence-aware copy-paste augmentation represents a paradigm shift from random enhancement to strategic hard sample construction. Our approach deliberately maximizes learning efficiency by overlaying high-uncertainty regions onto positions with maximal distribution divergence. This is quantified using SSIM as a proxy for feature distribution difference, ensuring that augmented samples not only increase the proportion of challenging regions but also expose the model to diverse contextual environments. The resulting hard samples are not merely difficult patches but information-rich learning opportunities that specifically target the model's current weaknesses while maintaining anatomical plausibility. The process consists of three key steps:

1) Volume Partitioning and Hardness Quantification: Each 3D volume $\mathbf{X}^w$, $\mathbf{X}^s$ is partitionedinto N^3 non-overlapping cubic sub-regions $\{\mathbf{X}^w_j\}_{j=1}^{N^3}$, $\{\mathbf{X}^s_j\}_{j=1}^{N^3}$, where $\mathbf{X}^w_j$, $\mathbf{X}^s_j \in \mathbb{R}^{\frac{W}{N}\times\frac{H}{N}\times\frac{L}{N}}$. Corresponding hardness maps $\mathbf{U}^w_j$, $\mathbf{U}^s_j \in \mathbb{R}^{\frac{W}{N}\times\frac{H}{N}\times\frac{L}{N}}$ are derived for each cube. The aggregate hardness score $u^w_j \in \mathbb{R}$, for the j-th small cube is computed as: $u^w_j = \frac{N^3}{W\cdot H\cdot L}\cdot\sum \mathbf{U}^w_j$, with u^s_j calculated analogously. These scores form hardness sets $\mathbf{U}^w_{set}$, $\mathbf{U}^s_{set} \in \mathbb{R}^{N^3}$.

2) Hard Sample Identification and Pairing: We identify the indices of the top-M high-hardness cubes ($\mathbf{I}^w_{top}$, $\mathbf{I}^s_{top}$) and the bottom-$\left\lfloor\frac{N^3}{2}\right\rfloor$ low-hardness cubes ($\mathbf{I}^w_{min}$, $\mathbf{I}^s_{min}$) via:

$$\mathbf{I}^w_{top} = \left\{\text{idxT}^w_m | u_{\text{idxT}^w_m} \in \text{Top}_M(\mathbf{U}^w_{set})\right\}, \tag{16}$$

$$\mathbf{I}^w_{min} = \left\{\text{idxB}^w_n | u_{\text{idxB}^w_n} \in \text{Bottom}_{\left\lfloor\frac{N^3}{2}\right\rfloor}(\mathbf{U}^w_{set})\right\}, \tag{17}$$

where $M \in \left\{k \in \mathbb{Z}^+ | k < \frac{N^3}{2}\right\}$. For each high-hardness cube $\mathbf{X}^w_{\text{idxT}^w_m}$, we select the most distributionally divergent low-hardness cube $\mathbf{X}^s_{\text{idxB}^s_n}$ from the paired volume using the Structural Similarity Index (SSIM) [33]:

$$\text{idxM}^s_m = \underset{i\in n}{\text{argmin}}\left[\text{SSIM}\left(\mathbf{X}^s_{\text{idxB}^s_i}, \mathbf{X}^w_{\text{idxT}^w_m}\right)\right], \tag{18}$$

where SSIM measures luminance, contrast, and structural dissimilarity, which can reflect the distribution differences between two images.

3) Epistemic-Driven Copy-Paste Augmentation: High-hardness cubes from $\mathbf{X}^w$ are pasted into $\mathbf{X}^s$ at positions of low uncertainty and high distributional divergence (and vice versa), generating augmented volumes $\mathbf{X}^a_{hd}$ and $\mathbf{X}^b_{hd}$. This strategy explicitly increases the proportion of hard samples while maximizing feature diversity. Corresponding labels $\mathbf{Y}^a_{hd}$ and $\mathbf{Y}^b_{hd}$ are created via identical operations.

The augmented volumes $\mathbf{X}^a_{hd}$ and $\mathbf{X}^b_{hd}$ are fed into networks $\mathcal{F}_w(\cdot;\mathbf{\Theta}_w)$ and $\mathcal{F}_s(\cdot;\mathbf{\Theta}_s)$, respectively, with losses computed as described in Sec. III-C. This approach ensures targeted learning from challenging regions while mitigating confirmation bias through diversified feature exposure.

E. Alignment of Hard Feature Distributions

The Alignment of Hard Feature Distributions (AHFD) mechanism is designed to mitigate a key limitation in scribble-supervised learning: sparse annotations cause feature distribution shifts particularly at organ boundaries, leading to confirmation bias. Unlike global feature alignment methods that ignore regional hardness variations, our approach specifically targets high-uncertainty boundary regions identified through epistemic uncertainty maps. By modeling each organ's global feature distribution as a Gaussian and applying Mahalanobis distance-based pull operations, we establish an anatomical consistency prior that guides boundary refinement even without direct supervision. The simultaneous push operations enforce inter-organ separability, which is crucial for distinguishing adjacent organs with similar appearance characteristics. This targeted alignment differs fundamentally from previous approaches by creating a direct connection between uncertainty quantification and feature space regularization.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on June 20,2026 at 13:01:41 UTC from IEEE Xplore. Restrictions apply.

Specifically, we propose a feature alignment strategy that leverages global organ distributions as anchors to refine boundary predictions. We first compute the normalized uncertainty map $\mathbf{U} = \text{Norm}(\mathbf{U}^w + \mathbf{U}^s)$ by aggregating outputs from both network branches. For each organ class k, high-hardness voxels are identified as:

$$\Omega_k^{hd} = \{x | \mathbf{U}(x) > \varrho, \mathbf{S}(x) = k\}, \tag{19}$$

where ϱ denotes the threshold, x represents the voxel coordinates, and $\mathbf{S} = \text{argmax}(\mathbf{P})$ indicates the segmentation prediction. Next, we model the global feature distribution of each organ using a Gaussian distribution: $\mathcal{N}(\mu_k, \Sigma_k)$, where $\mu_k = \mathbb{E}_{x\sim\Omega_k}[\mathbf{H}(x)]$, and $\Sigma_k = \text{Cov}_{x\sim\Omega_k}[\mathbf{H}(x)]$. Hard boundary regions are characterized by $\mathcal{N}(\mu_k^{hd}, \Sigma_k^{hd})$.

Next, we employ Mahalanobis distance to minimize each hard feature vector against its corresponding organ's overall distribution:

$$\mathcal{L}_{pull} = \sum_{k}^{K} \sum_{x\in\Omega_k^{hd}} \text{MD}\{\mathbf{H}(x), \mathcal{N}(\mu_k, \Sigma_k)\}. \tag{20}$$

Concurrently, to prevent feature confusion between organs, we enforce separation between each organ's features and other organs' global distributions:

$$\mathcal{L}_{push} = \sum_{j\neq k}^{K} \sum_{x\in\Omega_k} \max[0, \delta - \text{MD}\{\mathbf{H}(x), \mathcal{N}(\mu_j, \Sigma_j)\}], \tag{21}$$

where $\delta = 3.0$ is a margin ensuring that class means are separated by at least δ. The combined boundary refinement loss is:

$$\mathcal{L}_{edge} = \mathcal{L}_{pull}^w + \mathcal{L}_{pull}^s + \varsigma \cdot (\mathcal{L}_{push}^w + \mathcal{L}_{push}^s). \tag{22}$$

The approach effectively mitigates confirmation bias in scribble-supervised learning while preserving anatomical plausibility.

Convergence Analysis: $\mathcal{L}_{pull}$ is a quadratic form with a Hessian matrix of $2\Sigma_k^{-1}$. Since the covariance matrix Σ_k is positive definite, Σ_k^{-1} is also positive definite. Thus, $\mathcal{L}_{pull}$ is strictly convex. The gradient $\nabla_H \mathcal{L}_{pull} = 2\Sigma_k^{-1}(H(x) - \mu_k)$ of $\mathcal{L}_{pull}$ is linear and approaches zero as $H(x) \to \mu_k$. $\mathcal{L}_{push}$ is a hinge loss-based function with piecewise linearity. Its gradient is zero when $\text{MD}\{\mathbf{H}(x), \mathcal{N}(\mu_j, \Sigma_j)\} \geq \delta$; otherwise, the gradient is $\nabla_H \mathcal{L}_{push} = -2\Sigma_i^{-1}(H(x) - \mu_j)$. Since the hinge loss is convex (but not strictly convex), $\mathcal{L}_{push}$ is also convex. When $MD < \delta$, the gradient direction points away from μ_j.

F. Perturbation and Training

For the strong perturbation branch, we implement an epistemic-adaptive unadversarial perturbation (EAUP) strategy rather than conventional random noise injection. Unlike adversarial perturbations that deliberately mislead models, our approach generates benign perturbations designed to enhance robust feature representations. Crucially, we recognize that applying strong perturbations to high-hardness regions (characterized by elevated epistemic uncertainty) may exacerbate model instability in these already challenging areas. To address this, we employ an adaptive perturbation scheme where perturbation magnitude is inversely proportional to hardness:

$$\mathbf{X}^s = \mathbf{X} - (1 - \mathbf{W}_u) \odot \epsilon \cdot \text{sign}[\nabla_{\mathbf{X}} \mathcal{L}_{ps}(\mathcal{F}_w(\mathbf{X})), \widehat{\mathbf{Y}})], \tag{23}$$

where gradient descent is used to iteratively modify input images by maximizing agreement between predictions and pseudo-label masks.

The complete training objective combines multiple loss components through a phased optimization strategy. For predictions of both weakly ($\mathbf{X}^w$) and strongly ($\mathbf{X}^s$) perturbed samples, we define:

$$\mathcal{L}_a = \mathcal{L}_{sup} + (1 - \phi) \cdot (\mathcal{L}_{dis} + \mathcal{L}_{align}^{all}) + \phi \cdot (\mathcal{L}_{ps} + \mathcal{L}_{reg}^{all} + \varphi \cdot \mathcal{L}_{edge}), \tag{24}$$

where ϕ represents a *sigmoid* ramp-up function that adaptively adjusts with the training iteration count t according to $\phi(t) = \frac{1}{(1+\exp -\kappa\cdot(t-t_0))}$. Here, t_0 is set to 40% of the total number of iterations, and $\kappa = 5$. During the first 30% of iterations, the training primarily relies on supervised and uncertainty-guided losses. Between 30% and 60% of iterations, a hybrid approach combining supervised and pseudo-supervised signals is employed. Beyond 60% of iterations, training is dominated by pseudo-supervised signals weighted by uncertainty.

The corresponding loss for hard augmented samples ($\mathbf{X}_{hd}^w$ and $\mathbf{X}_{hd}^s$) is denoted as $\mathcal{L}_b$, yielding the total optimization objective:

$$\mathcal{L} = \mathcal{L}_a + \mathcal{L}_b. \tag{25}$$

IV. Experiments

A. Datasets

1) 3D CT Datasets: The **BTCV** Dataset [40] contains 30 abdominal CT scans (3779 slices) with 13 organ classes and background. Following [41], all scans were resampled to $[1.5 \times 1.5 \times 2.0]\,\text{mm}^3$, normalized to zero mean and unit variance, and split into 24 training and 6 test cases. The **WORD** dataset [1] includes 150 scans from 150 subjects with 16 organ classes and background, each comprising 159–330 slices of 512×512 pixels at $0.976 \times 0.976\,\text{mm}^2$ in-plane resolution and 2.5–3.0 mm slice thickness. Preprocessing followed the same procedure as BTCV, with dataset partitions of 100/20/30 for training, validation, and test, respectively.

Note that these partitioning strategies follow established benchmarks for within-dataset evaluation. For BTCV, all methods use the model weights saved at the final training iteration and are evaluated on the test set. For WORD, all methods select the checkpoint yielding the best validation performance and are subsequently evaluated on the test set. For cross-dataset generalization experiments (Section IV-F1), we use a different protocol where models are trained exclusively on the BTCV training set and evaluated on the WORD test set (considering only shared organ classes), without utilizing WORD's training or validation sets.

2) 2D Ultrasound Datasets: **EchoNet-Dynamic** [42] comprises 10,030 apical four-chamber echocardiography videos from Stanford Hospital, from which 20,046 annotated frames at end-systole and end-diastole were extracted and used as segmentation targets. **BUSI** [43] contains 780 breast ultrasound

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on June 20,2026 at 13:01:41 UTC from IEEE Xplore. Restrictions apply.

TABLE I

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. NOTE: DSC: DICE SIMILARITY COEFFICIENT, 95HD: 95% HAUSDORFF DISTANCE, LG/RG: LEFT/RIGHT ADRENAL GLANDS, *: THE RESULT IS SIGNIFICANTLY DIFFERENT FROM OURS WITH $P < 0.05$ VIA PAIRED T-TEST. NUMBERS DENOTED AS SUBSCRIPTS ARE STANDARD DEVIATIONS

Methods	Spleen	R. kidney	L. kidney	Gallbladder	Esophagus	Liver	Stomach	Aorta	IVC	Veins	Pancreas	RG	LG	Average DSC	Average 95HD
V-Net [13]	95.59	94.12	94.0	74.05	73.69	96.72	85.87	88.94	86.42	73.95	79.13	67.18	65.75	$82.72_{11.73}$*	$6.91_{20.01}$*
TV [34]	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	$52.34_{25.67}$*	$55.02_{30.45}$*
pCE [35]	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	$50.94_{27.51}$*	$60.62_{39.73}$*
EM [36]	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	$64.24_{19.86}$*	$45.50_{44.57}$*
S2L [37]	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	$74.45_{16.82}$*	$17.18_{9.76}$*
CPS [38]	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	$63.38_{22.34}$*	$20.79_{17.00}$*
DMPLS [3]	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	$51.97_{26.53}$*	$56.37_{32.70}$*
USTM [8]	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	$70.85_{18.37}$*	$36.37_{37.06}$*
DMSPS [7]	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	$74.65_{15.13}$*	$16.08_{25.04}$*
EFFDNet [39]	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	$74.08_{15.62}$*	$17.11_{26.52}$*
Ours	**93.85**	**90.71**	90.01	**71.68**	**71.64**	86.33	**85.33**	**88.58**	**80.29**	**69.16**	**76.47**	**51.58**	**54.14**	$\mathbf{77.67}_{14.37}$	$\mathbf{12.91}_{20.16}$

images from 600 patients (25–75 years), covering normal, benign, and malignant cases with corresponding masks. **DDTI** [44] consists of 637 thyroid ultrasound images from 299 cases, including thyroiditis, cystic nodules, adenomas, and cancer, using a preprocessed version released by the MICCAI 2020 TN-SCUI Challenge winners. All experiments are based on five-fold cross-validation.

3) Scribble Annotation: All scribble annotations for the samples were generated from dense annotations. To ensure that the scribbles resemble those produced by human annotators, we adapted and improved upon prevailing automated scribble generation methods [1], [3], [45]. Specifically, we randomly selected 3–5 points within each segmented object and connected them using B-spline interpolation to produce smooth, natural-looking strokes. These strokes were strictly constrained to lie within the region of interest (ROI) corresponding to the target class. Additionally, 1–2 scribbles were generated per ROI, with the number adjusted according to the ROI's size. For the CT slices in the BTCV and WORD datasets, intermittent annotation was employed to reduce annotation costs.

B. Implementation Details and Evaluation Metrics

All experiments were conducted on an Nvidia RTX 3090 GPU with V-Net as the backbone, implemented in Python 3.8, PyTorch 1.12, and the WSL4MIS[1] codebase. SGD was used with a weight decay of 10^{-4}, momentum 0.9, 50k iterations, an initial learning rate of 0.01, and polynomial decay $lr = lr_{base} \times \left(1 - \frac{iterations}{iterations_{max}}\right)^{0.9}$. Training samples ($96 \times 96 \times 96$) were randomly cropped from CT volumes and divided into N^3 cubes ($N = 3$). For hyperparameter details, see Sec. IV-D7. Ultrasound images were resized to 256×256. V-Net was used exclusively for 3D CT datasets. U-Net was used exclusively for 2D ultrasound datasets. All ultrasound datasets (heart, breast, thyroid) were merged into a single multi-class segmentation task rather than being trained as independent binary problems, ensuring consistent organ-level semantic labeling across datasets.

We carefully designed data augmentation strategies. For weak augmentation, we employed: (1) random rotation with angle $\in [-10°, 10°]$; (2) horizontal and vertical flipping (50% probability each); (3) random translation within ± 3 voxels/pixels in each dimension; and (4) uniform scaling with factor $\in [0.95, 1.05]$. For strong augmentation, we applied: (1) intensity adjustment with multiplicative factor $\in [0.8, 1.2]$; (2) contrast adjustment using gamma correction with $\gamma \in [0.7, 1.5]$; (3) Gaussian noise with $\sigma \in [0.01, 0.05]$ of the image intensity range; and (4) our EAUP with maximum perturbation magnitude $\epsilon = 0.1$. For 3D CT datasets (BTCV and WORD), augmentations were implemented in 3D space, while for 2D ultrasound datasets, they were applied in 2D.

In our experiments, we used the standard SSIM implementation with a window size of 11×11 pixels, a Gaussian filter sigma of 1.5, and default values for the stabilization constants ($K_1 = 0.01$, $K_2 = 0.03$).

All methods were evaluated under identical conditions using DSC (%, Dice-Sørensen Coefficient) and 95HD (pixel, 95% Hausdorff Distance). Our generalization experiment was conducted in a cross-dataset setting: trained on BTCV and tested on WORD (shared organ classes).

C. Comparison With Existing Methods

1) 3D CT Datasets: We compared our method with 9 advanced scribble-supervised learning approaches on two 3D multi-organ CT datasets: BTCV and WORD. These methods include **TV** [34], **pCE** [35], **EM** [36], **S2L** [37], **CPS** [38], **DMPLS** [3], **USTM** [8], **DMSPS** [7], and **EFFDNet** [39]. Our approach achieved superior results on the BTCV dataset (Table I), with an average DSC of 77.67% and 95HD of 12.91, outperforming all comparative methods ($p<0.05$), and significantly improving segmentation of challenging structures, such as the gall bladder, esophagus, and portal/splenic veins. The 95HD reduction of 19.2% over DMSPS validates the effectiveness of our edge feature alignment strategy. On the WORD dataset (Table II), our method achieved 77.11% DSC and 7.26 95HD, showing a 3.25% DSC improvement over DMSPS and a 31.3% reduction in boundary errors. Our method notably excelled in segmenting small or complex structures, with improvements of 7.72%, 20.19%, and 7.57% for the duodenum, intestines, and colon, respectively. The results demonstrate the robustness and effectiveness of

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

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on June 20,2026 at 13:01:41 UTC from IEEE Xplore. Restrictions apply.

TABLE II
QUANTITATIVE COMPARISON RESULTS ON THE WORD DATASET. NOTE: LFH/RFH: LEFT/RIGHT FEMUR HEAD, *: THE RESULT IS SIGNIFICANTLY DIFFERENT FROM OURS WITH $P < 0.05$ VIA PAIRED T-TEST

Methods	Liver	Spleen	L. kidney	R. kidney	Stomach	Gallbladder	Esophagus	Pancreas	Duodenum	Colon	Intestine	Adrenal	Rectum	Bladder	LFH	RFH	Average DSC	Average 95HD
V-Net [13]	96.42	95.07	95.02	95.34	90.58	77.91	74.78	82.99	64.99	84.56	87.6	67.83	78.37	90.38	88.99	89.11	$85.00_{12.77}$*	$7.32_{19.26}$
TV [34]	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 [35]	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	$60.00_{22.40}$*	$50.65_{35.38}$*
EM [36]	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	$66.45_{20.92}$*	$17.93_{23.55}$*
S2L [37]	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 [38]	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	$66.27_{21.72}$*	$15.66_{20.03}$*
DMPLS [3]	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	$58.03_{23.63}$*	$51.01_{39.24}$*
USTM [8]	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	$64.42_{22.27}$*	$30.75_{36.41}$*
DMSPS [7]	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	$73.86_{17.56}$*	$10.56_{14.27}$*
EFFDNet [39]	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	$73.41_{18.27}$*	$11.68_{14.96}$*
Ours	**93.78**	**92.71**	**90.59**	**91.84**	**87.23**	59.70	52.62	**74.05**	**60.94**	**79.43**	**80.06**	**49.81**	65.50	**82.48**	**85.93**	**87.13**	$\mathbf{77.11}_{17.33}$	$\mathbf{7.26}_{6.49}$

TABLE III
QUANTITATIVE COMPARISON RESULTS ON THE ULTRASOUND DATASET USING 5-FOLD CROSS-VALIDATION. *: THE RESULT IS SIGNIFICANTLY DIFFERENT FROM OURS WITH $P < 0.05$ VIA PAIRED T-TEST

Methods	Heart	Breast	Thyroid	Avg. DSC	Avg. 95HD
U-Net [14]	$92.68_{0.08}$*	$78.02_{2.52}$*	$79.31_{1.69}$*	$83.34_{8.12}$*	$19.82_{14.78}$*
pCE [35]	$68.71_{0.74}$*	$63.02_{2.94}$*	$64.41_{1.48}$*	$65.38_{2.97}$*	$32.91_{18.19}$*
Gated CRF [15]	$84.43_{0.61}$	$71.46_{2.58}$*	$66.69_{1.08}$	$74.19_{9.18}$	$25.81_{17.12}$*
DMPLS [3]	$74.58_{1.13}$*	$65.25_{3.36}$*	$63.02_{1.62}$*	$67.62_{6.13}$*	$29.19_{16.52}$*
USTM [8]	$71.76_{3.48}$*	$62.13_{3.09}$*	$63.98_{1.74}$	$65.96_{5.11}$*	$33.08_{18.39}$*
ScribbleVC [46]	$84.79_{0.56}$	$70.03_{3.90}$*	$62.84_{1.68}$*	$72.55_{11.19}$*	$25.68_{17.56}$*
ScribFromer [12]	$63.68_{1.77}$*	$65.15_{3.39}$*	$60.03_{1.88}$*	$62.95_{2.64}$*	$31.58_{14.47}$*
DMSPS [7]	$80.77_{1.58}$*	$71.83_{2.16}$*	$63.55_{1.26}$*	$72.05_{8.61}$*	$27.34_{17.80}$*
Bayesian_WSS [47]	$78.76_{0.21}$*	$68.83_{3.09}$*	$63.85_{2.16}$	$70.28_{7.72}$*	$26.91_{15.17}$*
EFFDNet [39]	$79.95_{1.17}$*	$70.79_{2.98}$*	$63.74_{1.53}$*	$71.50_{8.20}$*	$27.06_{16.21}$*
Ours	$\mathbf{85.26}_{0.47}$	$\mathbf{74.27}_{3.18}$	$\mathbf{68.32}_{2.95}$	$\mathbf{75.95}_{8.59}$	$\mathbf{23.74}_{15.58}$

TABLE IV
QUANTITATIVE COMPARISON RESULTS ON THE BTCV, WORD, AND ULTRASOUND DATASETS WITH THE NNU-NET PIPELINE. NOTE: OUR PROPOSED METHOD DOES NOT UTILIZE NNU-NET. *: THE RESULT IS SIGNIFICANTLY DIFFERENT FROM OURS WITH $P < 0.05$ VIA PAIRED T-TEST

Methods	BTCV (3D)		WORD (3D)		Ultrasound (2D)	
	DSC	95HD	DSC	95HD	DSC	95HD
nnU-Net [48]	$85.73_{9.02}$*	$6.50_{15.48}$*	$87.32_{7.95}$*	$6.91_{15.27}$	$85.47_{7.54}$*	$17.96_{15.61}$*
nnU-Net + TV	$58.94_{22.67}$*	$55.89_{33.15}$*	$63.30_{20.08}$*	$64.51_{36.47}$*	$66.16_{2.82}$*	$30.98_{17.48}$*
nnU-Net + pCE	$56.62_{20.12}$*	$61.93_{35.30}$*	$63.33_{18.28}$*	$65.85_{34.82}$*	$66.88_{3.64}$*	$30.02_{17.65}$*
nnU-Net + EM	$69.93_{18.85}$*	$45.51_{42.74}$*	$69.86_{18.13}$*	$18.98_{20.75}$*	$67.51_{5.14}$*	$21.47_{16.58}$*
nnU-Net + Ours	$\mathbf{77.89}_{14.41}$	$\mathbf{12.63}_{20.08}$	$\mathbf{77.34}_{17.21}$	$7.31_{6.58}$	$\mathbf{76.02}_{8.71}$	$\mathbf{23.51}_{15.47}$
Ours	$77.67_{14.37}$	$12.84_{20.16}$	$77.11_{17.33}$	$\mathbf{7.28}_{6.49}$	$75.95_{8.59}$	$23.74_{15.58}$

TABLE V
ABLATION STUDY RESULTS OF KEY COMPONENTS (BASELINE IS CPS, EDHE: EPISTEMIC-DRIVEN HARDNESS ESTIMATION, CP: COPY-PASTE AUGMENTATION, FA: FEATURE ALIGNMENT) ON CT AND ULTRASOUND DATASETS

EDHE	CP	FA	BTCV (DSC)	WORD (DSC)	Ultrasound (DSC)
			$63.38_{22.34}$	$66.27_{21.72}$	$69.83_{8.81}$
✓			$74.93_{14.87}$	$74.16_{17.54}$	$71.98_{7.95}$
✓	✓		$76.62_{15.53}$	$76.33_{16.41}$	$75.06_{8.63}$
✓	✓	✓	$77.67_{14.37}$	$77.11_{17.33}$	$75.95_{8.59}$

our method under scribble supervision, achieving statistically significant improvements ($p<0.05$) in both datasets.

2) 2D Ultrasound Datasets: We compared our method with 8 advanced approaches on the 2D multi-organ (heart, breast, thyroid) ultrasound dataset. These methods include **pCE**, **Gated CRF** [15], **DMPLS**, **USTM**, **ScribbleVC** [46], **ScribFormer** [12], **DMSPS**, and **Bayesian_WSS** [47]. Results are shown in Table III. Our model achieved the highest DSC of 85.26% on the heart, 74.27% on the breast, and 68.32% on the thyroid, along with the lowest average 95HD value of 23.74 pixels. These results represent statistically significant improvements ($p<0.05$) over most competing methods in key categories. Notably, our approach demonstrated strong performance in boundary-aware segmentation, particularly on challenging breast and thyroid cases, where it reduced Hausdorff distance errors compared to methods such as Gated CRF, DMSPS, and Bayesian_WSS. The consistent superiority across diverse organ types confirms the generalizability and robustness of our epistemic-driven hardness-adaptive focusing framework under sparse scribble supervision.

3) Comparison With nnU-Net Pipeline Implementations: Table IV compares our method with nnU-Net combined with various weakly supervised approaches. While fully supervised nnU-Net provides strong upper-bound performance, our method achieves competitive results without relying on this complex pipeline. Notably, it substantially outperforms nnU-Net with conventional scribble-supervised losses (TV, pCE, and EM), which are selected because they can be integrated into nnU-Net without architectural changes. Although nnU-Net + Ours yields additional gains, the improvement over Ours alone is modest, suggesting that our method is already near-optimal. Moreover, nnU-Net requires extensive automated configuration and training, incurring higher computational cost, whereas our epistemic-driven framework effectively leverages sparse scribble annotations without dependence on nnU-Net's adaptive preprocessing and training strategies.

D. Ablation Study

1) Component Analysis: We systematically evaluated our framework's components on CT, and Ultrasound datasets, using CPS as the baseline. The results (Table V) demonstrate progressive performance improvements with each added module. Adding epistemic-driven hardness estimation (EDHE) improved DSC by 11.55% (BTCV), 7.89% (WORD), and 2.15% (Ultrasound), demonstrating its effectiveness in identifying challenging regions across diverse imaging modalities. Incorporating copy-paste augmentation (CP) further boosted DSC by 1.69% (BTCV), 2.17% (WORD), and 3.08% (Ultrasound), validating its efficacy in diversifying hard samples and enhancing generalization, particularly in ultrasound images

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on June 20,2026 at 13:01:41 UTC from IEEE Xplore. Restrictions apply.

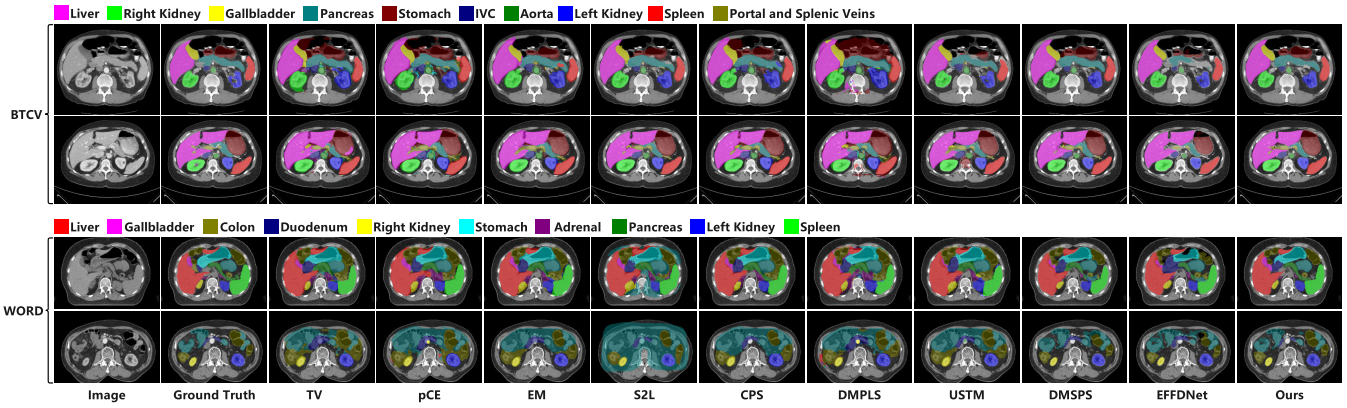

Fig. 4. The visualization of segmentation results by various methods on the CT (BTCV and WORD) datasets.

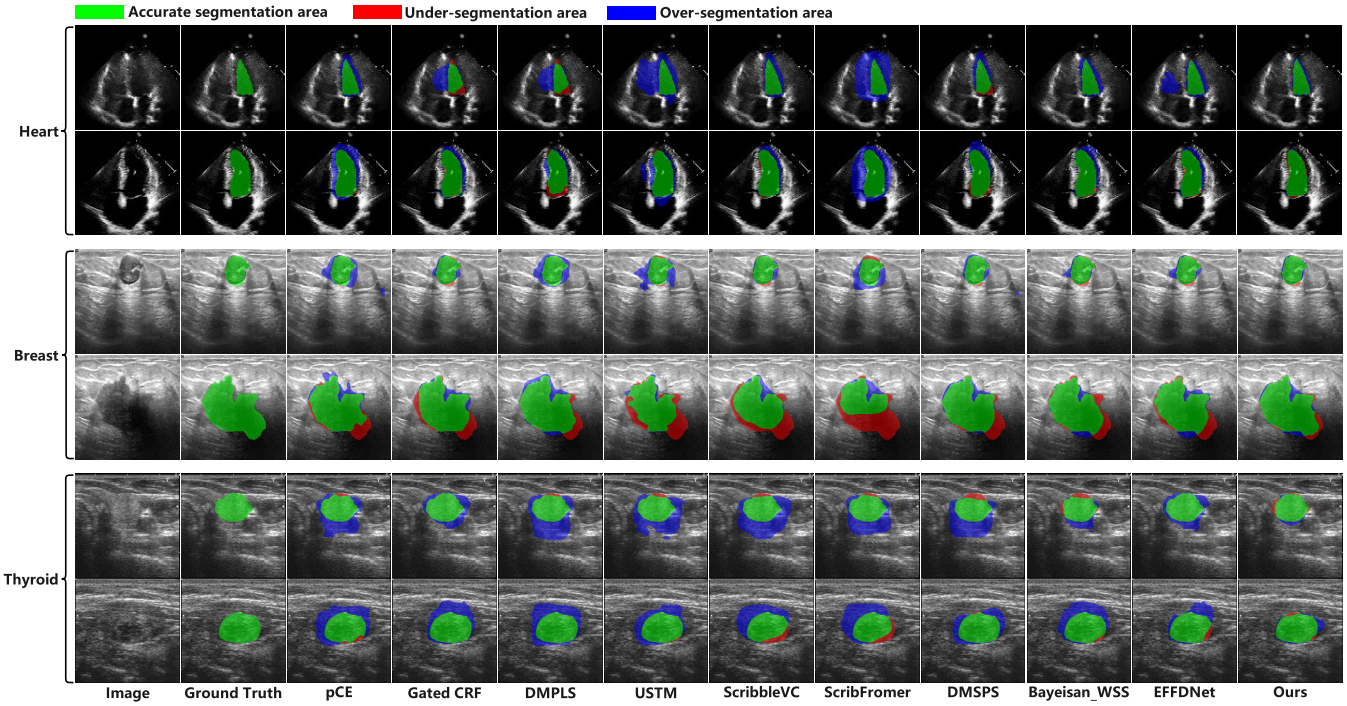

Fig. 5. The visualization of segmentation results by various methods on the ultrasound datasets.

with lower contrast and higher noise. Finally, integrating feature alignment (FA) led to additional gains of 1.05% (BTCV), 0.78% (WORD), and 0.89% (Ultrasound), achieving the optimal performance. The complete framework excels in segmenting small organs and refining object boundaries, with consistent improvements observed across all three datasets, highlighting its robustness and modality-agnostic applicability.

2) Hard Sample Selection Strategy: We compared five selection strategies to validate our epistemic-driven approach (Table VI): (1) Random selection (baseline), (2) Deep Ensembles (5 V-Nets with different initializations), (3) MC Dropout (10 passes), (4) Entropy-based sampling $\left(\mathbf{U}_{ent} = -\sum P(x)\log P(x)\right)$, and (5) our method. Epistemic uncertainty methods (Ensembles, MC Dropout, ours) outperformed aleatoric (Entropy) by 1.2–2.7% DSC, with ours achieving the highest scores (77.67% BTCV, 77.11% WORD) via evidence-aware hardness estimation. While deep ensembles (76.28% DSC) and MC Dropout (76.16% DSC) were competitive, their training costs were 4.3× and 1.8× higher than ours, demonstrating our method's optimal performance-efficiency balance.

TABLE VI
COMPARISON OF HARD SAMPLE SELECTION STRATEGIES. STRATEGY DENOTES THE METHOD FOR HARDNESS (UNCERTAINTY) ESTIMATION. UNC TYPE INDICATES THE TYPE OF UNCERTAINTY, SUCH AS EPISTEMIC UNCERTAINTY OR ALEATORIC UNCERTAINTY

Strategy	Unc Type	BTCV (DSC)	WORD (DSC)	Time (h)
Random	-	$74.15_{15.84}$	$73.55_{18.14}$	12.3
Deep Ensembles	Epistemic	$76.28_{15.03}$	$75.76_{17.24}$	58.7
MC Dropout	Epistemic	$76.16_{14.57}$	$75.59_{17.48}$	24.5
Entropy	Aleatoric	$75.02_{14.94}$	$74.85_{17.62}$	13.1
Ours	Epistemic	$77.67_{14.37}$	$77.11_{17.33}$	13.8

3) Strong Perturbation Strategy Comparison: We visualize two original samples and their EAUP-perturbed counterparts. Since the perturbations may be imperceptible to the human

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on June 20,2026 at 13:01:41 UTC from IEEE Xplore. Restrictions apply.

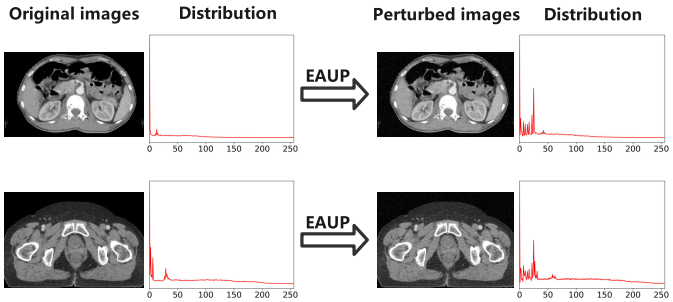

Fig. 6. Visualization of Unadversarial Perturbation Effects. The Distribution plot displays the grayscale histogram of the image, highlighting subtle perturbations that remain imperceptible to the human eye.

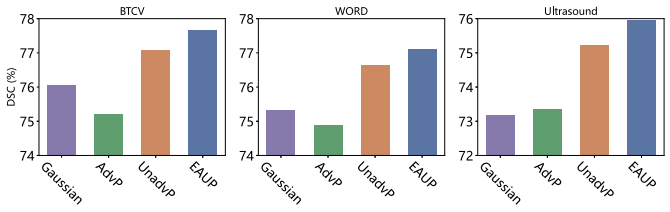

Fig. 7. Impact of different strong perturbation methods on results. Gaussian: random Gaussian noise, AdvP: indicates adversarial perturbation, UnadvP: unadversarial perturbation.

eye, we quantify these changes using grayscale statistics, as shown in Fig. 6. The statistics significantly differ before and after perturbation. Each sample is perturbed to emphasize key features for segmentation. To validate the effectiveness of EAUP, we compared it with random Gaussian noise, adversarial perturbation, and unadversarial perturbation. As shown in Fig. 7, while adversarial perturbation degrades performance by disrupting beneficial features, unadversarial perturbation enhances them. Our method further improves performance by minimizing perturbations to sensitive features.

4) Cube Partitioning Analysis: To evaluate the impact of cube partitioning granularity on segmentation performance, we evaluated three configurations ($N = 2, 3, 4$) on the BTCV and WORD datasets (Table VII). The 3^3 partition (27 cubes of $32 \times 32 \times 32$ voxels) performed best, outperforming both coarser 2^3 and finer 4^3 partitions. This indicates a trade-off: smaller cubes ($24 \times 24 \times 24$ in 4^3) may lose contextual information essential for organ coherence, while larger cubes ($48 \times 48 \times 48$ in 2^3) reduce hard sample localization precision. The 3^3 configuration balances spatial resolution and anatomical context, effectively supporting hardness-guided augmentation while maintaining organ structural integrity. A similar analysis was performed on ultrasound data, where images were divided into patches under three configurations (2^2, 3^2, and 4^2). Consistently, the 3^2 partitioning (yielding 9 patches) achieved the highest performance.

5) Sensitivity Analysis of TopK: To investigate the impact of the number of hard samples (TopK), we present the experimental results in Fig. 8. On the CT dataset, optimal performance is achieved when TopK= 3 (with the volume partitioned into 27 small cubes using a 3^3 grid), whereas on the ultrasound dataset, the best performance is observed at TopK= 2 (with the image divided into 9 small patches using a 3^2 grid). These results indicate that an insufficient number of hard samples may

TABLE VII
COMPARISON OF DIFFERENT CUBE PARTITIONING SIZES (N^3) AND THEIR IMPACT ON SEGMENTATION PERFORMANCE FOR 3D CT (BTCV AND WORD) DATASETS

Partition Size (N^3)	Cube Size	BTCV (DSC)	WORD (DSC)
2^3 (8 cubes)	$48 \times 48 \times 48$	$76.32_{14.80}$	$75.87_{17.62}$
3^3 (27 cubes)	$32 \times 32 \times 32$	$\mathbf{77.67}_{14.37}$	$\mathbf{77.11}_{17.33}$
4^3 (64 cubes)	$24 \times 24 \times 24$	$76.98_{14.59}$	$76.44_{17.35}$

TABLE VIII
ABLATION STUDY OF DIFFERENT LOSS COMPONENTS ON CT (BTCV AND WORD) AND ULTRASOUND DATASETS. THE PERFORMANCE METRIC IS DSC

$\mathcal{L}_{sup}$	$\mathcal{L}_{dis}$	$\mathcal{L}_{align}$	$\mathcal{L}_{ps}$	$\mathcal{L}_{reg}$	$\mathcal{L}_{edge}$	BTCV	WORD	Ultrasound
✓						$50.94_{27.51}$	$60.00_{22.40}$	$65.38_{2.97}$
✓			✓			$73.42_{16.62}$	$70.14_{18.85}$	$73.24_{6.38}$
✓	✓	✓				$67.45_{20.42}$	$66.56_{21.83}$	$70.94_{5.78}$
✓			✓	✓		$74.02_{16.05}$	$73.12_{17.79}$	$74.11_{9.19}$
✓	✓	✓	✓	✓	✓	$77.67_{14.37}$	$77.11_{17.33}$	$75.95_{8.59}$

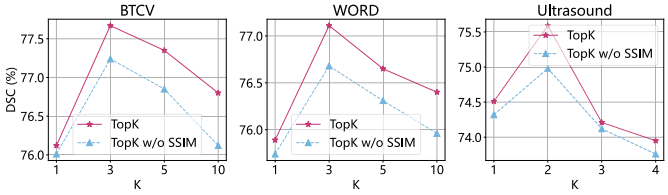

Fig. 8. The experimental results on the impact of the number of selected hard samples (TopK) on performance, analyzing the adjustment of TopK and the reserve or removal of SSIM (distribution discrepancy maximization).

lead to the omission of critical difficult examples, resulting in inadequate learning; conversely, an excessive number of hard samples introduces noise and increases computational overhead, thereby diluting the discriminative power of hard samples. Furthermore, the findings corroborate that maximizing SSIM to enhance the diversity of sample distributions can further improve model performance.

6) Effectiveness of Loss Function: Table VIII shows that each loss component contributes critically to performance. Using only $\mathcal{L}_{sup}$ yields 50.94% DSC on BTCV. Adding $\mathcal{L}_{ps}$ raises it to 73.42%, while further incorporating $\mathcal{L}_{align}$ and $\mathcal{L}_{dis}$ improves results, confirming their role in reducing feature bias from scribbles. The full combination performs best, demonstrating that epistemic-driven hardness estimation ($\mathcal{L}_{dis} + \mathcal{L}_{align}$) and feature alignment ($\mathcal{L}_{edge}$) synergize with pseudo-supervision ($\mathcal{L}_{ps}$) to enhance scribble usage and boundary refinement. The visualization of the experimental results for each dataset is shown in Fig. 9.

7) Hyperparameter Analysis: Our hyperparameter analysis (Fig. 10) revealed optimal performance with $\beta = 1.5$ for controlling uncertainty weighting in pseudo-supervised loss, $\gamma = 3.0$ for adaptive regularization strength, $\varrho = 0.9$ as the threshold for identifying hard boundary voxels, $\eta = 0.05$ for balancing conflict measurement in the discrepancy loss, $\tau = 0.3$ to focus on high-conflict regions, and $\varsigma = 0.1$ to appropriately weight the boundary refinement components. Systematic experimentation across the parameter space demonstrated that these carefully calibrated values effectively balanced model

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on June 20,2026 at 13:01:41 UTC from IEEE Xplore. Restrictions apply.

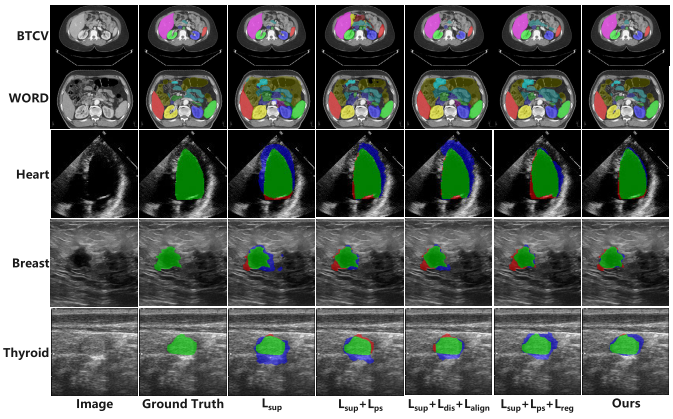

Fig. 9. Qualitative comparison of segmentation results with different loss function components on CT (BTCV and WORD) and ultrasound (Heart, Breast and Thyroid) datasets.

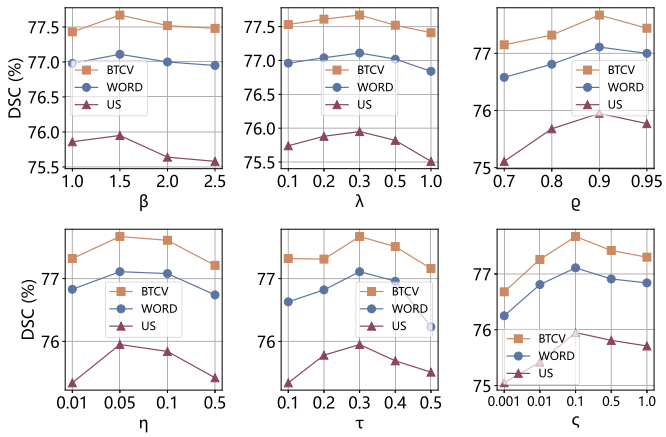

Fig. 10. Hyperparameter analysis showing performance (DSC) across different parameter values for β, γ, ϱ, η, τ, and ς. US denotes ultrasound dataset.

TABLE IX
QUANTITATIVE COMPARISON OF DIFFERENT NETWORK UPDATE STRATEGIES. ALL VARIANTS WERE TRAINED UNDER IDENTICAL CONDITIONS TO ENSURE A FAIR COMPARISON

Methods	BTCV (DSC)	WORD (DSC)	Ultrasound (DSC)
Ours (Independent)	$77.67_{14.37}$	$77.11_{17.33}$	$75.95_{8.59}$
Mean Teacher [49]	$75.84_{15.13}$	$75.22_{17.68}$	$74.09_{8.89}$
EMA Sharing	$76.12_{14.56}$	$75.68_{17.47}$	$74.46_{8.68}$

sensitivity to hard samples while maintaining training stability, with consistent performance across both CT (BTCV and WORD) and ultrasound datasets.

8) Comparison With Mean Teacher Framework: To justify our dual-branch design, we compared it with Mean Teacher [49] and an EMA-sharing variant. Our method uses two independently initialized V-Nets. In **Mean Teacher**, a student network is trained with gradient descent while the teacher network is updated via unidirectional EMA ($\alpha = 0.99$), using the teacher's prediction to supervise the student. In **EMA sharing**, both networks update each other via bidirectional EMA to encourage weight similarity. As shown in Table IX, our independent initialization achieves superior performance. EMA-based alignment reduces representational diversity,

TABLE X
PERFORMANCE COMPARISON OF DIFFERENT TRAINING PHASE TRANSITION STRATEGIES

Training Strategy	BTCV (DSC)	WORD (DSC)	Ultrasound (DSC)
Fixed Early Phase	$75.63_{15.04}$	$75.12_{18.05}$	$74.17_{9.12}$
Fixed Late Phase	$75.21_{15.28}$	$74.75_{18.21}$	$73.86_{9.32}$
Linear Transition	$76.35_{14.89}$	$75.84_{17.65}$	$74.89_{8.87}$
Our Adaptive Transition	$77.67_{14.37}$	$77.11_{17.33}$	$75.95_{8.59}$

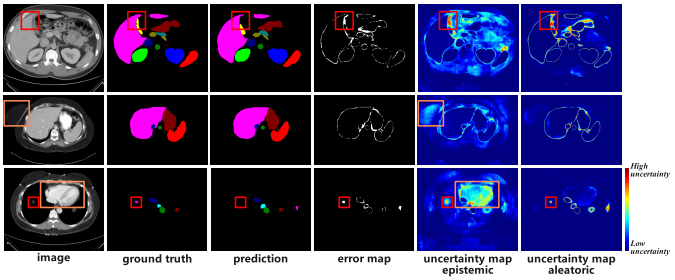

Fig. 11. Visualization of uncertainty map and prediction error map. The epistemic uncertainty map exhibits distinct patterns from the aleatoric uncertainty map. Red boxes highlight regions with prediction errors, while orange boxes indicate uncommon anatomical structures.

limiting the model's ability to capture epistemic uncertainty. Although EMA-sharing outperforms Mean Teacher, it still falls short of our approach. This experiment validates our core argument: preserving parameter independence between the two branches is crucial for generating meaningful disagreement, which serves as a reliable signal for epistemic uncertainty estimation. This inherent diversity prevents confirmation bias and enables more robust pseudo-label generation through mutual calibration, ultimately leading to higher segmentation accuracy under scribble supervision.

9) Impact of the Transition Mechanism on Performance: We analyzed the impact of different training phase transition strategies on segmentation performance. Table X demonstrates that our adaptive transition mechanism significantly outperforms fixed early-phase, fixed late-phase, and linear transition approaches across all evaluated datasets. The adaptive strategy dynamically adjusts the contribution of loss components according to training progress, effectively leveraging both uncertainty-guided learning in early phases and pseudo-label refinement in later phases. This dynamic adaptation prevents premature reliance on potentially noisy pseudo-labels while ensuring the model eventually leverages all available information, creating an optimal learning trajectory that maximizes segmentation accuracy under sparse annotation constraints.

E. Uncertainty and Error Analysis

Epistemic uncertainty reflects model ignorance (e.g., rare anatomical structures), while aleatoric uncertainty captures data noise (e.g., ambiguous boundaries). Fig. 11 displays prediction errors, epistemic uncertainty maps, and aleatoric uncertainty maps for several cases. The patterns of epistemic and aleatoric uncertainty differ, with high aleatoric uncertainty primarily concentrated at the segmentation boundaries. In the first case, prediction errors coincide with high epistemic uncertainty. The second case shows elevated epistemic uncertainty

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on June 20,2026 at 13:01:41 UTC from IEEE Xplore. Restrictions apply.

TABLE XI
COMPARISON OF CALIBRATION QUALITY ACROSS 3 UNCERTAINTY ESTIMATION METHODS ON CT (BTCV AND WORD) AND ULTRASOUND DATASETS. (ECE: EXPECTED CALIBRATION ERROR, BS: BRIER SCORE)

Method	BTCV		WORD		Ultrasound	
	ECE	BS	ECE	BS	ECE	BS
Ours (EDL)	0.048	0.132	0.044	0.128	0.057	0.135
MC-Dropout	0.067	0.152	0.071	0.158	0.089	0.171
Deep Ensembles	0.051	0.139	0.049	0.134	0.067	0.148

in an atypical anatomical location. The third case highlights high epistemic uncertainty due to an unannotated and uncommon heart structure in an abdominal CT dataset.

To further validate the reliability of our epistemic uncertainty estimation, we conducted a comprehensive calibration analysis comparing our EDL-based approach with MC-Dropout and Deep Ensembles across all three datasets (Table XI). Our method consistently achieves the lowest Expected Calibration Error (ECE) and Brier Score (BS), demonstrating superior uncertainty calibration quality, while requiring lower computational cost than MC-Dropout and Deep Ensembles. This precise uncertainty quantification is clinically significant as it ensures that when our model is uncertain about a segmentation (particularly in challenging regions like organ boundaries or rare anatomical structures), this uncertainty is accurately reflected in the prediction confidence. In medical applications where overconfident but incorrect predictions could lead to misdiagnosis, our well-calibrated uncertainty estimates provide an essential safety mechanism.

F. Generalization and Robustness Analysis

1) Generalization Analysis: To evaluate cross-dataset generalization capability, we conduct a separate experiment where the model is trained solely on the BTCV training set and directly evaluated on the WORD test set, considering only the shared organ classes between datasets. This protocol deliberately avoids using any samples from WORD's training or validation sets to assess the model's ability to generalize across different data distributions, scanning protocols, and population characteristics under scribble supervision. All hyperparameters and model selection criteria are determined using only the BTCV training set, ensuring a strict evaluation of out-of-distribution generalization performance. The results in Table XII show that our method achieves a DSC of 70.33% and a 95HD of 22.74, surpassing all baselines ($p < 0.05$). It excelled on challenging organs (stomach: 76.37%, pancreas: 56.68%), outperforming DMSPS by 9.6% and 8.21%. Despite cross-dataset performance drops, our method showed the smallest gap (7.34% DSC decline vs. 8.94% for DMSPS) and 19.3% better boundary precision (95HD). Results validate that our epistemic-driven hardness adaptation and feature alignment strategies effectively learn transferable anatomical representations from scribble labels.

2) Robustness Analysis: As shown in Table XIII, our method demonstrates superior robustness against various noise types (Gaussian, Poisson, impulse, motion blur) compared to other methods, exhibiting the smallest performance drops. Our strategies mitigate sensitivity to noise, particularly preserving boundary precision under strong perturbations (e.g., motion blur). This validates the framework's stability in real-world scenarios with imperfect data.

TABLE XII
CROSS-DATASET GENERALIZATION PERFORMANCE ON WORD TEST SET USING MODELS TRAINED ON BTCV (EVALUATING ONLY SHARED ORGAN CLASSES BETWEEN DATASETS). NOTE: R.KID: RIGHT KIDNEY, L.KID: LEFT KIDNEY, GALL: GALLBLADDER, ESO: ESOPHAGUS, STO: STOMACH, PAN: PANCREAS, AVG: AVERAGE. *: $P < 0.05$ VIA PAIRED T-TEST

Methods	spl	R.kid	L.kid	Gall	Eso	Liver	Sto	Pan	Avg.DSC	Avg.95HD
EM [36]	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 [8]	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 [35]	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 [3]	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 [34]	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 [38]	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 [37]	78.02	**82.99**	68.03	35.04	36.60	90.85	63.69	45.45	$63.15_{41.24}$*	$25.78_{48.47}$*
DMSPS [7]	81.82	79.33	**81.67**	41.89	41.42	84.42	66.77	48.47	$65.72_{24.04}$*	$28.19_{30.37}$*
EFFDNet [39]	82.35	80.12	80.47	43.28	44.63	87.15	69.84	43.56	$66.43_{24.18}$*	$27.36_{29.84}$*
Ours	**83.76**	82.84	79.97	**44.40**	**46.71**	**91.89**	**76.37**	**56.68**	$\mathbf{70.33}_{22.74}$	$\mathbf{26.87}_{36.67}$

TABLE XIII
PERFORMANCE COMPARISON UNDER DIFFERENT NOISE TYPES AND LEVELS, SHOWING THE RELATIVE PERFORMANCE DROP (ΔDSC) OF OURS, DMSPS, AND CPS METHODS ON THE BTCV DATASET

Noise Type	Level	Ours (ΔDSC)	DMSPS (ΔDSC)	CPS (ΔDSC)
Gaussian	σ=0.05	-1.23%	-3.17%	-4.85%
	σ=0.1	-2.87%	-6.92%	-9.34%
Poisson	peak=50	-0.98%	-2.45%	-3.77%
Impulse	10%	-3.12%	-7.83%	-11.02%
Motion Blur	kernel=5	-4.25%	-8.71%	-12.63%

TABLE XIV
SEGMENTATION PERFORMANCE UNDER VARYING SCRIBBLE SPARSITY LEVELS

Scribble Sparsity	BTCV (DSC)	WORD (DSC)	Ultrasound (DSC)
50%	$72.52_{17.43}$	$71.87_{18.65}$	$70.34_{7.68}$
75%	$75.12_{15.97}$	$74.49_{17.99}$	$73.15_{8.16}$
90%	$76.64_{14.98}$	$76.06_{17.58}$	$74.83_{8.41}$
100%	$77.67_{14.37}$	$77.11_{17.33}$	$75.95_{8.59}$

G. Scribble Sparsity Analysis

To evaluate our method's robustness under varying annotation quality reflecting real clinical conditions, we conducted a scribble sparsity analysis. While our synthesized scribbles mimic physician-drawn annotations, clinician-specific variability remains. To simulate the diverse annotation quality from different clinicians, we systematically shortened original scribbles by randomly selecting one endpoint and varying the other, thereby modeling real-world scenarios where time constraints lead to heterogeneous scribble densities across organs and cases. As shown in Table XIV, our framework demonstrates remarkable resilience to increasing scribble sparsity across all three datasets. When annotation coverage decreases to 50% of the original scribbles, our method maintains competitive performance. Notably, the performance drop is non-linear, with the most significant improvement occurring between 50% and

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on June 20,2026 at 13:01:41 UTC from IEEE Xplore. Restrictions apply.

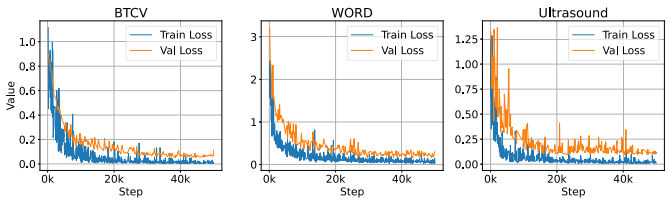

Fig. 12. Training and validation loss curves of our method on CT (BTCV and WORD) and ultrasound datasets.

75% sparsity, while diminishing returns are observed beyond 90% coverage. This suggests our epistemic-driven hardness-adaptive mechanism effectively compensates for extremely sparse annotations by focusing learning on the most challenging regions. The consistent performance trend across different imaging modalities (CT and ultrasound) further validates the generalizability of our approach in practical clinical settings where annotation time and expertise may be limited.

V. Discussion

The key conceptual advance of our framework is a self-improving closed loop in which quantified epistemic uncertainty guides hard sample generation, while hard sample learning and feature alignment jointly reduce uncertainty. Unlike previous scribble-supervised methods that follow a linear pipeline, our approach introduces feedback among components. This synergy enables better uncertainty estimation to produce more informative hard samples, which in turn enhance feature learning and further refine uncertainty. As shown in the ablation results (Table V), the joint effect of these components surpasses the sum of their individual contributions, validating our hypothesis that effective scribble-supervised learning requires a continuously adaptive learning paradigm rather than isolated component improvements.

Training convergence: To better illustrate the optimization behavior of our multi-component loss in Eq. (25), we plot the training and validation losses over epochs on the CT (BTCV and WORD) and ultrasound datasets in Fig. 12. Although the losses exhibit considerable fluctuations in the early phases of training, the overall trends show stable decreases and eventual convergence. Because our loss functions are gradually transitioned across different training phases, no abrupt changes are observed. Moreover, the validation loss closely follows the training loss without divergence, indicating that our training strategy does not suffer from severe overfitting.

Limitations: While our method is computationally more efficient than Bayesian ensembles or MC Dropout, copy-paste augmentation and feature alignment still introduce moderate overhead. Although synthetic scribbles mimic clinical annotation patterns, they cannot fully replicate the nuanced judgment of experienced radiologists; future work will incorporate real radiologist-drawn scribbles via clinical collaboration. Additionally, while validated across multiple organs and modalities, performance on extremely rare anatomical or pathological cases warrants further investigation.

Clinical Applicability: Our method offers substantial practical value for clinical deployment. Compared to time-consuming dense pixel-wise annotation, scribble supervision reduces the annotation burden to sparse curve drawing. Our framework operates with existing backbone networks (V-Net/U-Net) without architectural modifications, facilitating integration into current clinical pipelines. Inference is performed in real time, with an average processing time of 0.3129 ± 0.1248 s per volume. Our open-source implementation and validated performance across both CT and ultrasound modalities further support translational potential, enabling deployment as a fully automated segmentation solution that delivers both AI-generated segmentations and calibrated epistemic uncertainty maps for rapid clinical review, refinement, and risk-aware decision making.

VI. Conclusion

We propose an epistemic-driven hardness-adaptive focusing framework for scribble-supervised 3D multi-organ segmentation. By leveraging Dempster-Shafer theory for uncertainty-aware hardness estimation, our method dynamically identifies challenging regions and enhances learning through targeted copy-paste augmentation and feature distribution alignment. Extensive experiments on multi-organ CT and ultrasound datasets demonstrate superior performance, robustness, and generalizability compared to existing methods. Our framework bridges the gap between weak supervision and precise segmentation, offering a practical solution for clinical applications.

References

[1] X. Luo et al., "WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from CT image," *Med. Image Anal.*, vol. 82, Nov. 2022, Art. no. 102642.

[2] X. Liang, N. Li, Z. Zhang, J. Xiong, S. Zhou, and Y. Xie, "Incorporating the hybrid deformable model for improving the performance of abdominal CT segmentation via multi-scale feature fusion network," *Med. Image Anal.*, vol. 73, Oct. 2021, Art. no. 102156.

[3] X. Luo et al., "Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels supervision," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2022, pp. 528–538, doi: 10.1007/978-3-031-16431-6_50.

[4] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3159–3167.

[5] L. Grady, "Random walks for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1768–1783, Nov. 2006.

[6] V. Vezhnevets and V. Konouchine, "GROWCUT: Interactive multi-label ND image segmentation by cellular automata," in *Proc. Graphicon*, 2005, pp. 150–156.

[7] M. Han et al., "DMSPS: Dynamically mixed soft pseudo-label supervision for scribble-supervised medical image segmentation," *Med. Image Anal.*, vol. 97, Oct. 2024, Art. no. 103274.

[8] X. Liu et al., "Weakly supervised segmentation of COVID19 infection with scribble annotation on CT images," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108341.

[9] W. Li, R. Bian, W. Zhao, W. Xu, and H. Yang, "Diversity matters: Cross-head mutual mean-teaching for semi-supervised medical image segmentation," *Med. Image Anal.*, vol. 97, Oct. 2024, Art. no. 103302.

[10] K. Zhang and X. Zhuang, "CycleMix: A holistic strategy for medical image segmentation from scribble supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11646–11655.

[11] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12674–12684.

[12] Z. Li et al., "ScribFormer: Transformer makes CNN work better for scribble-based medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 43, no. 6, pp. 2254–2265, Jun. 2024.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on June 20,2026 at 13:01:41 UTC from IEEE Xplore. Restrictions apply.

[13] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.
[14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.
[15] A. Obukhov, S. Georgoulis, D. Dai, and L. Van Gool, "Gated CRF loss for weakly supervised semantic image segmentation," 2019, *arXiv:1906.04651*.
[16] B. Kim and J. C. Ye, "Mumford–Shah loss functional for image segmentation with deep learning," *IEEE Trans. Image Process.*, vol. 29, pp. 1856–1866, 2020.
[17] F. Gao et al., "Segmentation only uses sparse annotations: Unified weakly and semi-supervised learning in medical images," *Med. Image Anal.*, vol. 80, Aug. 2022, Art. no. 102515.
[18] Q. Chen and Y. Hong, "Scribble2D5: Weakly-supervised volumetric image segmentation via scribble annotations," in *Proc. 25th Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, Sep. 2022, pp. 234–243.
[19] M. Zhou, Z. Xu, K. Zhou, and R. K.-Y. Tong, "Weakly supervised medical image segmentation via superpixel-guided scribble walking and class-wise contrastive regularization," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2023, pp. 137–147.
[20] M. Han, X. Luo, W. Liao, S. Zhang, S. Zhang, and G. Wang, "Scribble-based 3D multiple abdominal organ segmentation via triple-branch multi-dilated network with pixel-and class-wise consistency," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2023, pp. 33–42.
[21] H. E. Wong, M. Rakic, J. V. Guttag, and A. V. Dalca, "ScribblePrompt: Fast and flexible interactive segmentation for any biomedical image," in *Proc. Eur. Conf. Comput. Vis.*, 2023, pp. 207–229.
[22] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.
[23] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6023–6032.
[24] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "ClassMix: Segmentation-based data augmentation for semi-supervised learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2021, pp. 1369–1378.
[25] J.-H. Kim, W. Choo, and H. O. Song, "Puzzle mix: Exploiting saliency and local statistics for optimal mixup," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5275–5285.
[26] D. Chen, Y. Bai, W. Shen, Q. Li, L. Yu, and Y. Wang, "MagicNet: Semi-supervised multi-organ segmentation via magic-cube partition and recovery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23869–23878.
[27] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.
[28] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," *Med. Image Anal.*, vol. 59, Jan. 2020, Art. no. 101557.
[29] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 6405–6416.
[30] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," *Neurocomputing*, vol. 338, pp. 34–45, Apr. 2019.
[31] M. Şensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 3183–3193.
[32] Y. Chen et al., "Evidence-based uncertainty-aware semi-supervised medical image segmentation," *Comput. Biol. Med.*, vol. 170, Mar. 2024, Art. no. 108004.
[33] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[34] M. Javanmardi, M. Sajjadi, T. Liu, and T. Tasdizen, "Unsupervised total variation loss for semi-supervised deep learning of semantic segmentation," 2016, *arXiv:1605.01368*.
[35] M. Tang, A. Djelouah, F. Perazzi, Y. Boykov, and C. Schroers, "Normalized cut loss for weakly-supervised CNN segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1818–1827.
[36] S. Yu, B. Zhang, J. Xiao, and E. G. Lim, "Structure-consistent weakly supervised salient object detection with local saliency coherence," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 4, pp. 3234–3242.
[37] H. Lee and W.-K. Jeong, "Scribble2Label: Scribble-supervised cell segmentation via self-generating pseudo-labels with consistency," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2020, pp. 14–23.
[38] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2613–2622.
[39] J. Liu, S. Y. Tan, X. Yang, Y. Xu, and S. Y. Yeo, "Effdnet: A scribble-supervised medical image segmentation method with enhanced foreground feature discrimination," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2025, pp. 194–204.
[40] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein, "MICCAI multi-atlas labeling beyond the cranial vault–workshop and challenge," in *Proc. Multi-Atlas Labeling Beyond Cranial Vault Workshop Challenge (MICCAI)*, vol. 5, 2015, p. 12.
[41] A. Hatamizadeh et al., "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 574–584.
[42] D. Ouyang et al., "Video-based AI for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, no. 7802, pp. 252–256, Apr. 2020.
[43] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data brief*, vol. 28, Feb. 2020, Art. no. 104863.
[44] L. Pedraza, C. Vargas, F. Narváez, O. Durán, E. Muñoz, and E. Romero, "An open access thyroid ultrasound image database," *Proc. SPIE*, vol. 9287, pp. 188–193, Jan. 2015.
[45] K. Gotkowski, C. Lüth, and P. F. Jäger, "Revisiting 3D medical scribble supervision: Benchmarking beyond cardiac segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2024, pp. 436–446.
[46] Z. Li, Y. Zheng, X. Luo, D. Shan, and Q. Hong, "ScribbleVC: Scribble-supervised medical image segmentation with vision-class embedding," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 3384–3393.
[47] Z. Zhou, Y. Hayashi, M. Oda, T. Kitasaka, and K. Mori, "A Bayesian approach to weakly-supervised laparoscopic image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2024, pp. 14–24.
[48] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "NnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021.
[49] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1195–1204.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on June 20,2026 at 13:01:41 UTC from IEEE Xplore. Restrictions apply.