Asymmetric Dual-Path Structural Consistency for Semi-Supervised Medical Image Segmentation

Anonymous submission

Abstract

Deep neural networks have achieved remarkable progress in medical image segmentation, but their performance still depends heavily on large-scale pixel-level annotations. Semi-supervised learning (SSL) alleviates this burden by leveraging unlabeled data, yet conventional teacher–student frameworks often suffer from error accumulation due to unidirectional supervision and noisy pseudo-labels. To overcome these limitations, we propose an Asymmetric Dual-Path Mutual Supervision (BiAsy-MS) framework that enables two models to collaborate through structurally and contextually diverse views. We introduce an irregular region-mixing augmentation (BezierMix) that generates anatomically aligned masks and asymmetric semantic contexts, promoting the exchange of complementary structural priors across labeled and unlabeled domains. In addition, an Information-driven Pseudo-Label Weighted mechanism adaptively emphasizes reliable pseudo-labels by accounting for information value, prediction stability, and confidence distribution, thereby suppressing noise and improving boundary recognition. Extensive experiments on three benchmark datasets (LA, Pancreas-NIH, and ACDC) show that our method consistently surpasses state-of-the-art SSL approaches under limited supervision (5% and 10% labeled data), achieving superior robustness, generalization, and data efficiency.

Keywords: Semi-Supervised, Medical Image Segmentation, Consistency Regularization, Data Augmentation

1 Introduction

Although deep neural networks have achieved remarkable progress in visual understanding tasks, their performance remains strongly dependent on large amounts of high-quality annotated data [1]. This requirement is especially critical for semantic segmentation, which demands fine-grained pixel-level supervision. However, in medical imaging, obtaining such annotations is particularly challenging: data acquisition is constrained by patient privacy and institutional data silos, and annotation requires domain expertise, structural interpretation, and extensive cross-validation

[2]. Moreover, unlike natural images with clearly defined object boundaries, medical images often contain fuzzy edges, high inter-class similarity, and substantial anatomical variability, making reliable annotation both labor-intensive and error-prone.

To mitigate the reliance on large-scale labeled data, several alternative learning paradigms have been explored. Meta-learning aims to enhance generalization under limited supervision by enabling rapid task adaptation, and has shown promising applications in medical image analysis [3]. Reinforcement learning has also been introduced to optimize segmentation and registration pipelines through sequential decision-making processes [4]. Additionally, metaheuristic and intelligent optimization algorithms offer flexible global search strategies for improving robustness and parameter efficiency in image processing and computer vision, as summarized in the comprehensive survey in [5].

Despite their advantages, these alternative learning paradigms often require carefully designed task distributions, handcrafted reward functions, or substantial computational overhead, and their effectiveness heavily depends on domain-specific priors. Therefore, in practical medical imaging scenarios with limited annotated samples, semi-supervised learning (SSL) remains a more direct and scalable solution. Leveraging a small amount of labeled data together with abundant unlabeled data, SSL has been widely adopted to improve model generalization and data utilization efficiency [6]. Recent strategies such as pseudo-label generation [7], consistency regularization [8–19], and contrastive learning [20–22] have demonstrated notable progress in the medical imaging domain. However, mainstream SSL frameworks still commonly rely on a single teacher–student paradigm, guiding the student through consistency constraints or pseudo-labels. While these approaches alleviate the scarcity of labeled data to some extent, they suffer from two fundamental limitations.

First, the supervision structure is essentially unidirectional, restricting model collaboration. In typical teacher–student methods, the teacher is updated via exponential moving averages of the student, resulting in delayed feedback and preventing it from capturing the student's most recent structural knowledge. When pseudo-labels contain noise—especially around boundaries or anatomically ambiguous regions—the one-way mechanism tends to magnify such errors and may lead to suboptimal convergence.

Second, the stability of pseudo-labels is inherently fragile. Most methods depend on maximum-probability (argmax) predictions, which are especially unreliable during early training. Even with confidence thresholding, noisy pseudo-labels can still be reinforced through iterative supervision, forming a "self-reinforcing pseudo-label noise loop" that compromises stability and generalization [15, 19].

Several recent works attempt to mitigate these issues through multi-path or dual-task designs—for example, dual-task consistency (DTC) [13], bidirectional copy–paste (BCP) [17], or multi-teacher switching strategies (CS-MT) [19]. However, these approaches typically operate within a single semantic space (dual-task), disrupt anatomical topology with rectangular region mixing, or rely on homogeneous model architectures that do not introduce genuinely diverse contextual views. Consequently, they do not explicitly construct asymmetric semantic pathways or enable collaboration based on structurally complementary information.

To overcome these limitations, we propose a novel Asymmetric Dual-Path Mutual supervision (BiAsy-MS) framework, where two models are trained using structurally and signal-intensity asymmetric views. Central to this framework is a Bezier-based irregular region-mixing augmentation (BezierMix). Unlike patch-based (CutMix [23]) or class-based (ClassMix [24]) approaches, BezierMix generates smooth, anatomically aligned free-form masks. Although labeled and unlabeled images are drawn from the same distribution, we deliberately construct information-asymmetric augmentation, exposing the two models to distinct contextual spaces. One model inserts unlabeled foregrounds into labeled contexts, while the other inserts labeled structures into unlabeled backgrounds. This directional asymmetry yields divergent spatial priors, enhances model complementarity, and explicitly embeds semantic contextual diversity into training—forming a collaboration mechanism absent in previous SSL frameworks.

To further ensure reliable mutual supervision, we introduce an information-driven pseudo-label weighted mechanism. This mechanism evaluates the reliability of pseudo-labels from three complementary perspectives: information value, prediction stability, and confidence distribution. Prediction divergence highlights structurally informative regions (e.g., boundaries), entropy-based suppression filters unstable areas, and confidence calibration mitigates misleading overconfidence. By integrating these components into a unified per-pixel weight, the model is guided to focus on informative and trustworthy signals, thereby improving robustness and boundary perception in both 2D and 3D settings.

We systematically evaluate the proposed framework on the LA, Pancreas-NIH, and ACDC datasets under 5% and 10% labeled settings. The results demonstrate consistent improvements over state-of-the-art SSL approaches, confirming the effectiveness of BezierMix augmentation, asymmetric dual-path supervision, and information-driven weighting in enhancing data efficiency and generalization. To the best of our knowledge, this is the first SSL framework that jointly integrates anatomically aligned region-level mixing, asymmetric dual-path semantic supervision, and information-driven consistency weighting, providing new insights into collaborative learning under limited annotations.

The remainder of this paper is organized as follows. Section 2 reviews the related work on semi-supervised medical image segmentation and region-level augmentation. Section 3 presents the proposed methodology, including the BezierMix data augmentation module, the asymmetric dual-path mutual supervision framework, and the information-driven pseudo-label weighting strategy. Section 4 describes the experimental setup, covering datasets, implementation details, evaluation metrics, and comparisons with state-of-the-art methods on the LA, Pancreas-NIH, and ACDC datasets, followed by ablation studies evaluating each component of our framework. Finally, Section 5 concludes the paper.

2 Related Work

Semi-supervised learning (SSL) has been widely explored in medical image segmentation to reduce dependence on large-scale annotated datasets. Existing SSL methods can be broadly categorized into three groups: contrastive-learning-based approaches

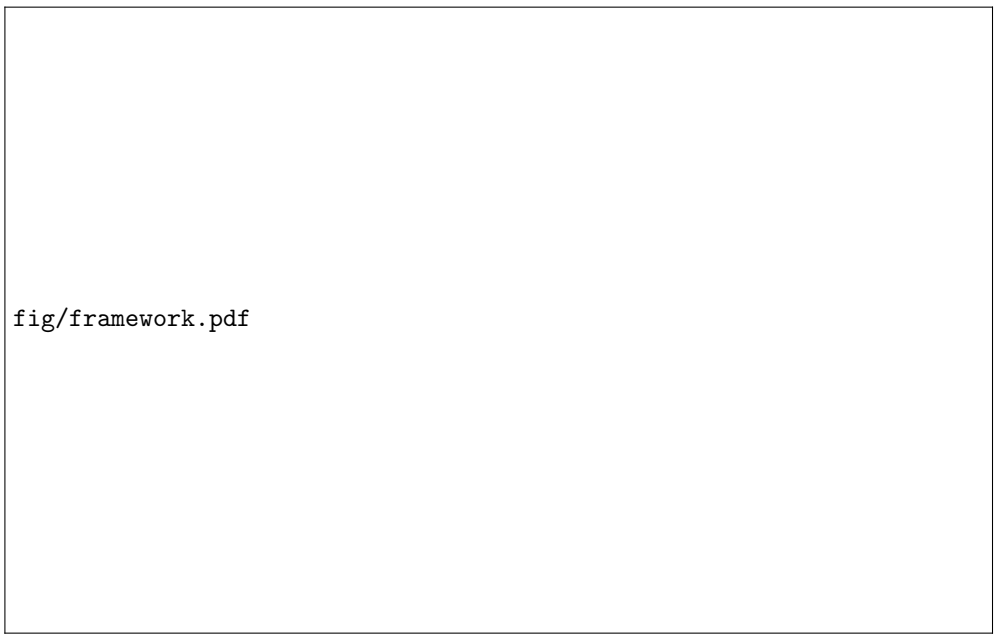


Fig. 1 A diagram of our BiAsy-MS framework, which employs an asymmetric dual-path student-teacher training paradigm. The teacher model and student model are trained with distinct data flows: A-flow and B-flow. The framework uses a BezierMask to create mixed data by combining labeled and unlabeled data. The two models generate pseudo-labels for the other's training, and a consistency loss is used for supervision. The model incorporates a new complementarity module to evaluate the information gain between predictions. Additionally, it includes a stability module to assess the uncertainty of the pseudo-labels and a calibration module to adjust confidence. This design enhances structural recognition and avoids training bias. The overall framework aims to improve segmentation quality and learning stability.

[20–22], pseudo-labeling strategies, and consistency-regularization frameworks. Among these, consistency regularization has become particularly influential due to its simplicity and empirical effectiveness. The central idea is to encourage a model to produce consistent predictions under perturbations of the same unlabeled sample. Early works generally adopted single-network formulations [9–11, 13], enforcing prediction consistency across augmented views. However, because the model learns only from its own outputs, such approaches are prone to confirmation bias and rapid error accumulation.

To alleviate these issues, more advanced frameworks introduced the teacher–student paradigm, with the Mean Teacher (MT) model [8] becoming a foundational solution. Numerous extensions have since been proposed. French et al. incorporated CutOut and CutMix augmentations to encourage the student to learn beyond the teacher's limited spatial context, while Olsson et al. developed ClassMix to replace rectangular masks with class-aware region mixing, enabling more effective exploitation of pseudo-labels. Additional regularization constraints have also been explored: SASSNet [12] integrated geometric shape priors, DTC [13] enforced dual-task consistency through auxiliary boundary prediction, and SimCVD [20] explicitly

modeled geometric and semantic structure relationships. UA-MT [14] leveraged uncertainty estimates to guide reliable pseudo-label usage, and cross-level consistency [25] refined feature similarity modeling across multi-scale representations. Other frameworks such as CoraNet [18] and UMCT [26] addressed regional uncertainty by distinguishing between reliable and unreliable regions or leveraging multi-view predictions.

Beyond consistency alone, recent works combine uncertainty modeling or representation learning to further improve robustness. Luo et al. [15] separated unlabeled pixels into certain and uncertain subsets, applying different supervisory signals to each. Xiang et al. [16] incorporated epistemic and aleatoric uncertainty into the consistency pipeline. Basak et al. [27] integrated contrastive learning to enhance feature discriminability, while Bai et al. [17] proposed bidirectional copy–paste augmentation to reduce distribution mismatch. Deng et al. [19] introduced a correlation-guided switching mean teacher (CS-MT) framework to mitigate error propagation by dynamically selecting the most reliable teacher.

These advances significantly broaden the utility of consistency regularization in medical image segmentation. However, most methods still rely on either homogeneous teacher–student pairs or symmetric dual-task designs, which limits their ability to leverage structurally complementary cues or explicitly model asymmetric semantic pathways between labeled and unlabeled domains. Enabling more effective and structurally informed knowledge transfer thus remains an open challenge.

3 Method

3.1 BezierMix Data Augmentation Module

To generate smoother and more anatomically plausible free-form shapes, we leverage the properties of Bezier curves to design an irregular mask generation strategy. In brief, we define a set of control points on a blank mask, where both the number and spatial distribution of points determine the shape of the Bezier curve. The enclosed area formed by the curve is then filled to define the cropping region, enabling the construction of augmented samples with greater smoothness, stability, and diversity.

To avoid cases where too few or overly clustered control points in the set P result in excessively small cropped regions, we first divide the entire slice of size $H \times W$ into grids and ensure that the control points are sampled across different grids. This design yields relatively stable cropping shapes. We observe that the complexity of the generated shapes is positively correlated with the number of control points—more points lead to more complex structures. The essence of the Bezier curve lies in using a parameter $t \in [0, 1]$ to perform linear interpolation along the line segments defined by the control point set $P = \{P_i\}_{i=1}^{n}$, thereby producing new vertices. This recursive process continues on the newly formed line segments until a final vertex is obtained, which corresponds to a point on the Bezier curve for a given t. By iterating through all possible values of t, the complete Bezier curve can be constructed. Figure 2 illustrates the drawing process of a quadratic Bezier curve. This process can be expressed using



Fig. 2 Visualization example of a quadratic Bezier curve. The control points P_0, P_1, and P_2 jointly determine the curve trajectory through linear interpolations (P'_{01}, P'_{12}) and the final Bezier point P''_t. The shaded region illustrates the smooth and anatomically aligned free-form shape that BezierMix leverages for structured region-level augmentation.

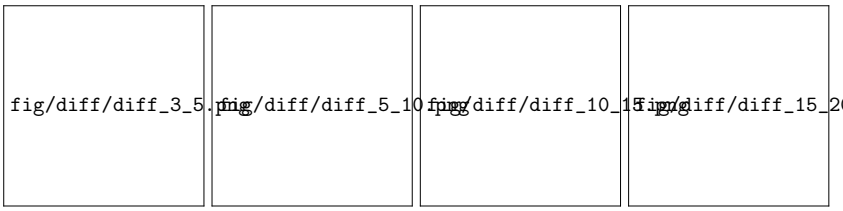


Fig. 3 Examples of BezierMix masks generated using different control-point intervals. As the number of control points increases (3-5, 5-10, 10-15, 15-20), the resulting shapes exhibit progressively higher geometric complexity and boundary variability, enabling structurally diverse region-level augmentation.

the Bernstein polynomial form:

$$B(t) = \sum_{i=0}^{n} \frac{n!}{i!(n-i)!}(1-t)^{n-i}t^i P_i, \tag{1}$$

where $\frac{n!}{i!(n-i)!}$ is the binomial coefficient, P_i is the i-th control point, and n is the order of the curve. Since it is infeasible to traverse all $t \in [0, 1]$, we sample a finite set of t values to render the curve. To alleviate aliasing on irregular boundaries, we apply Gaussian filtering after rendering and filling the curve region. The overall BezierMask generation process is summarized in Algorithm 1.The visualization of BezierMasks with different numbers of control points is shown in Figure 3; generally, more control points lead to more complex shapes.

Algorithm 1 BezierMix Algorithm

Input: number of endpoints n, number of control points m, number of grids g, kernel size K, threshold τ
Output: final mask M
Initialize empty control point set P and Bezier point set B
Compute grid interval $s = \text{image_size}//g$
for $i = 0$ to image_size step s **do**
    for $j = 0$ to image_size step s **do**
        Add grid point (i, j) to candidate grid set Grid_P^g
    end for
end for
$\text{Grid}_P^m = \text{RandomSampling}(\text{Grid}_P^g, m)$
for $(i, j) \in \text{Grid}_P^m$ **do**
    Generate control point $p = (\text{Randint}(j, j + s), \text{Randint}(i, i + s))$
    Add p to P
end for
for $t = 0$ to 1 step $1/n$ **do**
    Initialize $B_t = 0$
    for $i = 0$ to m **do**
        $B_t = B_t + \frac{m!}{i!(m-i)!}(1-t)^{m-i}t^i P_i$ ▷ binomial expansion
    end for
    Add B_t to B
end for
Connect and fill Bezier points in B to form raw mask M_b
Apply Gaussian smoothing to M_b with kernel size K to get M_s
Binarize M_s using threshold τ to obtain final mask M
return M

To construct cropping regions that preserve structural continuity and spatial consistency in volumetric data (e.g., 3D medical images), we extend the 2D BezierMix into a 3D continuous cropping strategy. Specifically, along the depth axis (D-axis), we introduce a cross-slice mask interpolation mechanism, ensuring that the generated cropping regions maintain smooth and coherent structures in three dimensions, better matching the anatomical morphology of real organs.

Concretely, we first randomly select several key slices in the depth dimension and apply the 2D Bezier-based cropping strategy to obtain corresponding 2D mask regions. For the non-key slices, masks are generated via linear interpolation between neighboring key slices along the depth axis, such that the structural contours undergo a spatially coherent deformation process across slices. This procedure can be formulated as:

$$\mathcal{M}_d = \frac{d_{\text{high}} - d}{d_{\text{high}} - d_{\text{low}}}\mathcal{M}_{\text{low}} + \frac{d - d_{\text{low}}}{d_{\text{high}} - d_{\text{low}}}\mathcal{M}_{\text{high}}, \tag{2}$$

where $\mathcal{M}_d$ denotes the interpolated mask at slice d, d_{low} and d_{high} are the nearest key slices before and after d, and $\mathcal{M}_{\text{low}}$, $\mathcal{M}_{\text{high}}$ are their corresponding 2D masks.

To further enhance the smoothness of cropping regions in spatial structure, we apply Gaussian filtering to the generated 3D masks across the (D, H, W) dimensions, ensuring voxel-wise consistency. Finally, we perform binarization to obtain irregular 3D cropping regions with structural continuity.

The augmented samples and their corresponding pseudo-labels are then formulated as:

$$\tilde{x} = m \odot x_s + (1 - m) \odot x_t, \quad \tilde{y} = m \odot y_s + (1 - m) \odot \hat{y}_t, \tag{3}$$

where $\odot$ denotes element-wise multiplication, x_s and x_t represent source and target samples, y_s is the ground truth label from the source domain, and $\hat{y}_t$ is the pseudo-label generated by the teacher model for the target domain.

3.2 Asymmetric Dual-Path Mutual Supervision

To more effectively capture complementary structural information and enhance cross-model consistency, we propose an Asymmetric Dual-Path Mutual Supervision (BiAsy-MS) framework comprising two parallel networks, $\mathcal{M}_A$ and $\mathcal{M}_B$. Each model is trained using asymmetric semantic-context views generated via BezierMix and learns through reciprocal cross-supervision.

Path-Level Structural Information Asymmetry. Leveraging the proposed BezierMix augmentation strategy, we construct two asymmetric data views with opposite structural migration directions for the two supervision paths.

For **Model A**, the structural region x_f^u cropped from an unlabeled image x^u is pasted onto a labeled image x^l, forming:

$$\tilde{x}_A = \text{BezierMix}(x^l, x_f^u). \tag{4}$$

Conversely, for **Model B**, the structural region x_f^l cropped from a labeled image x^l is pasted onto an unlabeled image x^u, resulting in:

$$\tilde{x}_B = \text{BezierMix}(x^u, x_f^l). \tag{5}$$

Through this directional mixing, the two models learn from complementary semantic contexts: unlabeled structures within labeled contexts for Model A, and labeled structures within unlabeled contexts for Model B. This design introduces a deliberate structural-information asymmetry, enabling the models to acquire diverse yet complementary representations and enforcing stronger cross-model consistency.

Structure Alignment via Foreground Centroid Migration. To ensure that the pasted structural regions remain spatially reasonable within the mixed images, we incorporate a foreground centroid alignment mechanism into the BezierMix process, providing structure-aware spatial guidance.

Given a source image x_s (from which the structure is extracted) and a target image x_t (providing the background), their structure masks are denoted by m_s and m_t, respectively. For labeled images, m_s is obtained from ground-truth annotations; for unlabeled images, it is derived from pseudo-labels predicted by the counterpart model.

We represent the geometric center of each structural region n using its centroid $c_n \in \mathbb{R}^2$ (or $\mathbb{R}^3$), computed as the mean of pixel coordinates within that region:

$$c_n = \frac{1}{|P_n|} \sum_{p \in P_n} p, \tag{6}$$

where $P_n = p \mid m_n(p) = 1$ denotes the set of pixels belonging to class n under mask m_n.

To align structural placement between the source and target images, an affine transformation $\mathcal{T}$ is applied to the generated Bezier mask M, shifting its centroid c_M toward the target centroid c_t:

$$M' = \mathcal{T}(M), \qquad \mathcal{T} = \text{Affine}(c_M \to c_t). \tag{7}$$

This centroid-guided alignment ensures that pasted structures appear in anatomically plausible locations, preventing spatial drift or unrealistic placement. As a result, it enhances structural consistency between the two supervision paths and improves the stability of the resulting pseudo-labels.

Pseudo-Label Based Mutual Supervision Mechanism. To enable structural compensation and collaborative learning between models, we further adopt a cross pseudo-label supervision strategy in the dual-path setting. Specifically, predictions from one model are used as supervisory signals for the other:

$$\mathcal{L}^A_{\text{cons}} = \mathcal{D}(\mathcal{M}_B(\tilde{x}_B), \hat{y}_A), \quad \mathcal{L}^B_{\text{cons}} = \mathcal{D}(\mathcal{M}_A(\tilde{x}_A), \hat{y}_B) \tag{8}$$

where $\hat{y}_A$ and $\hat{y}_B$ denote pseudo-labels obtained (sampled or softened) from $\mathcal{P}_A = \mathcal{M}_A(\tilde{x}_A)$ and $\mathcal{P}_B = \mathcal{M}_B(\tilde{x}_B)$, respectively. $\mathcal{D}$ represents the consistency loss function, such as Dice loss or KL divergence.

This mutual supervision mechanism facilitates information exchange across structural boundary regions while avoiding error accumulation that often arises from relying on a single teacher model with delayed updates.

3.3 Information-Driven Pseudo-Label Weighted Mutual Supervision

Although the mutual supervision framework enables complementary learning between two models, its effectiveness fundamentally depends on the reliability of pseudo-labels, especially in structurally ambiguous or boundary regions. During the early training stages, blindly enforcing supervision using noisy or unstable pseudo-labels may hinder convergence and degrade segmentation quality.

To address this issue, we propose an information-driven pseudo-label weighting mechanism that adaptively adjusts the supervisory strength at the pixel level. Instead of treating all pseudo-labels equally, our method integrates three complementary criteria—discrepancy guidance, uncertainty suppression, and confidence stability calibration—to estimate the supervisory potential of each pseudo-label and assign

pixel-wise weights accordingly. This unified strategy suppresses unreliable signals while emphasizing information-rich and structure-aware regions.

Discrepancy Guidance. During dual-path learning, the two models typically agree in confident regions but may diverge near structural boundaries or anatomically complex areas. Such moderate discrepancies often carry useful structural cues. To capture these informative differences, we introduce a divergence metric that measures the deviation between the prediction distributions p_A and p_B:

$$\text{Diver} = \mathcal{H}(p_A) + \mathcal{H}(p_B) - \mathcal{H}\Big((p_A + p_B)/2\Big),$$

where $\mathcal{H}(p) = -\sum_{c=1}^{C} p_c \log p_c$ denotes the pixel-wise entropy.

This formulation quantifies the information loss incurred when averaging the two distributions. Its behavior naturally highlights structurally informative regions. When the two models confidently agree (e.g., inside homogeneous organs), we have $p_A \approx p_B$ and thus $\mathcal{H}((p_A+p_B)/2) \approx \mathcal{H}(p_A)$, leading to Diver ≈ 0; these pixels offer limited value for complementary learning. In contrast, near anatomical boundaries or ambiguous regions, $p_A \neq p_B$. Due to the strict concavity of entropy,

$$\mathcal{H}\Big(\frac{p_A + p_B}{2}\Big) > \tfrac{1}{2}\mathcal{H}(p_A) + \tfrac{1}{2}\mathcal{H}(p_B),$$

which results in Diver > 0. Thus, the divergence metric serves as a principled indicator of model disagreement, revealing regions that may contain meaningful structural information.

Uncertainty Suppression. Although discrepancy highlights potentially informative regions, not all disagreements are beneficial. In early training or in low-quality image regions, both models may exhibit high entropy, producing "spurious discrepancies" that stem from prediction instability rather than complementary structural cues. To prevent leveraging such noise, we introduce an uncertainty suppression term based on the maximum entropy:

$$\text{Uncert} = \max(\mathcal{H}(p_A),\, \mathcal{H}(p_B)).$$

This design follows a conservative reliability assumption: the pseudo-label reliability is constrained by the weaker model. Even when one model appears confident, supervision should be attenuated if the other exhibits high uncertainty. Maximum entropy thus acts as a safety-aware measure that prevents unreliable supervision.

To combine discrepancy information with reliability, we adopt an exponential decay function as a soft gating mechanism. This formulation prevents high-entropy pixels from being assigned large weights while allowing confident discrepancies to be preserved. Additionally, to normalize the discrepancy response, we apply a sigmoid function to Diver. The resulting discrepancy-aware reliability weight is defined as:

$$w_{\text{diff}}(p_A, p_B) = \underbrace{\sigma(\lambda_1 \cdot \text{Diver})}_{\text{Discrepancy Guidance}} \cdot \underbrace{\exp(-\lambda_2 \cdot \text{Uncert})}_{\text{Uncertainty Suppression}},$$

where λ_1 and λ_2 control the trade-off between discrepancy exploitation and uncertainty suppression.

This formulation ensures that model disagreement contributes to supervision only under controlled and trustworthy conditions. In effect, it filters out uninformed uncertainty while emphasizing structurally meaningful discrepancies, enabling the framework to adaptively prioritize pixels that are both informative and reliable. This weighting strategy therefore improves pseudo-label quality and enhances the stability of mutual supervision throughout training.

Algorithm 2 Information-Driven Pseudo-Label Weighted Mutual Supervision

Input: model predictions p_A, p_B; BezierMix mask m; pseudo-labels $y_{\text{main}}, y_{\text{mix}}$; EMA statistics (μ, σ); hyperparameters $\lambda_1, \lambda_2, \gamma, \alpha, \beta$
Output: segmentation loss $\mathcal{L}_{\text{seg}}$ and updated (μ, σ)
Compute entropy maps $\mathcal{H}(p_A)$ and $\mathcal{H}(p_B)$
Compute averaged distribution $p_{\text{avg}} = \frac{1}{2}(p_A + p_B)$
Compute divergence:
Diver $= \mathcal{H}(p_A) + \mathcal{H}(p_B) - \mathcal{H}(p_{\text{avg}})$
Compute uncertainty:
Uncert $= \max\big(\mathcal{H}(p_A), \mathcal{H}(p_B)\big)$
Compute discrepancy-aware weight:
$w_{\text{diff}} = \sigma(\lambda_1 \cdot \text{Diver}) \cdot \exp(-\lambda_2 \cdot \text{Uncert})$
Compute pixel confidence:
$\text{conf}(x) = \max_c p_{\text{avg},c}(x)$
Compute batch mean μ_{batch} and std σ_{batch}
Update EMA confidence statistics:
$\mu \leftarrow \gamma\mu + (1-\gamma)\mu_{\text{batch}}, \quad \sigma \leftarrow \gamma\sigma + (1-\gamma)\sigma_{\text{batch}}$
Compute confidence stability weight:
$$w_{\text{conf}}(x) = \begin{cases} \exp\left(-\dfrac{(\text{conf}(x) - \mu)^2}{2\sigma^2}\right), & \text{conf}(x) < \mu \\ 1, & \text{otherwise} \end{cases}$$
Compute unified pixel-wise weight:
$w(x) = w_{\text{diff}}(x) \cdot w_{\text{conf}}(x)$
Compute weighted segmentation loss:
$\mathcal{L}_{\text{seg}} = \alpha \cdot \mathcal{L}\big(o, y_{\text{main}}, w \cdot m\big) + \beta \cdot \mathcal{L}\big(o, y_{\text{mix}}, w \cdot (1-m)\big)$
return $\mathcal{L}_{\text{seg}}$, updated (μ, σ)

Confidence Stability Calibration. Even after suppressing high-uncertainty regions, an additional challenge remains: highly confident predictions are not necessarily reliable. In structurally ambiguous areas or near anatomical boundaries, models may produce overconfident yet incorrect predictions due to bias, shortcut learning, or local appearance artifacts. Such "pathological confidence" can still misguide mutual supervision if left unregulated.

To mitigate this issue, we introduce a *distribution-aware confidence calibration* mechanism that evaluates confidence not solely by its magnitude but by its statistical

consistency with the model's global prediction behavior. This reflects the intuition that meaningful confidence should be both high and stable relative to the evolving confidence distribution in the mini-batch.

For each pixel, confidence is computed as:

$$\mathrm{conf}(x) = \max_c p_c(x).$$

We track the batch-level mean μ and standard deviation σ of confidence values through exponential moving averages:

$$\mu \leftarrow \gamma\mu + (1-\gamma)\mu_{\text{batch}}, \qquad \sigma \leftarrow \gamma\sigma + (1-\gamma)\sigma_{\text{batch}}.$$

These quantities provide a smooth and adaptive estimate of the model's confidence distribution across training.

To penalize abnormally low or unstable confidence values—even when entropy is low—we adopt a Gaussian-shaped weighting function:

$$w_{\text{conf}}(x) = \begin{cases} \exp\left(-\frac{(\mathrm{conf}(x)-\mu)^2}{2\sigma^2}\right), & \text{if } \mathrm{conf}(x) < \mu, \\ 1, & \text{otherwise.} \end{cases}$$

This formulation identifies pseudo-labels exhibiting "false confidence" and attenuates their influence. The weighting is intentionally asymmetric: only below-average or unstable confidence values are suppressed, while stronger-than-average confidence is preserved to avoid weakening genuinely reliable predictions.

Unified Weighting. By integrating the three complementary criteria—informative discrepancy, uncertainty-aware reliability, and confidence stability—the final pixel-wise weight is defined as:

$$w(p) = w_{\text{diff}}(p_A, p_B) \cdot \underbrace{w_{\text{conf}}(p)}_{\text{Confidence Calibration}}.$$

This unified formulation acts as a *multi-factor information gate*, selectively amplifying learning signals from pixels that are: (1) structurally informative (high discrepancy), (2) prediction-stable and reliable (low entropy), and (3) statistically meaningful (confidence consistent with the batch-level distribution). Together, these components follow a coherent information-theoretic principle: pseudo-labels should contribute to supervision only when they satisfy informativeness, stability, and statistical reliability. Accordingly, $w(p)$ provides a principled estimator of pixel-wise pseudo-label trustworthiness.

Application to the Supervision Mechanism. The resulting pixel-wise weighting map $w(x)$ is integrated into the loss computation to emphasize informative supervisory signals. For each model path, the segmentation loss is defined as:

$$\mathcal{L}_{\text{seg}} = \alpha \cdot \mathcal{L}(o, y_{\text{main}}, w \cdot m) + \beta \cdot \mathcal{L}(o, y_{\text{mix}}, w \cdot (1-m)),$$



Fig. 4 Visualization results on the LA dataset using 10% labeled data. The first row shows the segmentation outputs of the BCP method, which often exhibit boundary distortions and incomplete structural reconstruction. The second row displays the results produced by the proposed BiAsy-MS framework, demonstrating improved structural integrity and smoother anatomical surfaces. The third row presents the ground truth annotations. These comparisons highlight the advantage of our method in capturing fine-grained anatomical details under limited supervision.

where m denotes the Bezier-based mixing mask, and y_{main} and y_{mix} correspond to pseudo-labels from the original and mixed regions, respectively. The loss $\mathcal{L}$ represents Cross-Entropy or Dice loss.

Because $w(x)$ suppresses unreliable pseudo-label signals while highlighting structurally meaningful ones, the model benefits in several ways:

1. It effectively leverages complementary cues between the two models in regions where discrepancies are informative and stable.
2. It avoids being misled by boundary noise, ambiguous structures, or early-stage prediction instability.
3. Its attention gradually shifts toward anatomically consistent regions as training progresses, strengthening structural awareness.

These advantages collectively enhance the robustness, stability, and segmentation quality of the mutual supervision framework, enabling the system to selectively trust pseudo-labels that are both informative and reliable throughout training.

4 Experiments

In this section, we detail the experimental setup, implementation specifics, evaluation metrics, and comparative results of our proposed method. Our aim is to demonstrate the effectiveness of our approach in semi-supervised medical image segmentation and to benchmark its performance against state-of-the-art methods.

4.1 Datasets

We conducted extensive experiments on several publicly available medical image segmentation datasets to comprehensively evaluate the generalizability and robustness

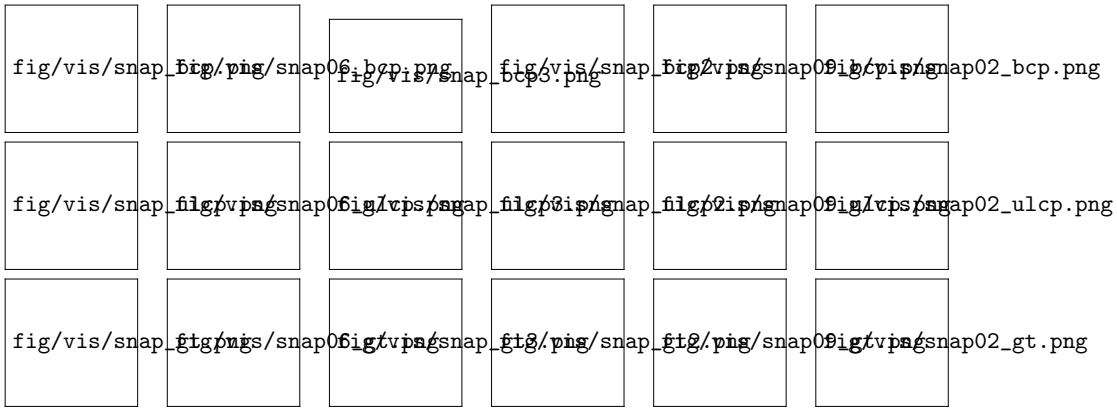


Fig. 5 Visualization results on the Pancreas-NIH dataset under 20% labeled data. The first row shows the segmentation outputs produced by the BCP method, which often exhibit fragmented shapes and inconsistent boundary reconstruction due to the irregular morphology of the pancreas. The second row displays the results generated by the proposed BiAsy-MS framework, demonstrating improved continuity, smoother surfaces, and better preservation of elongated anatomical structures. The third row provides the ground truth annotations. These comparisons highlight the robustness of our method when handling highly variable organ geometries under limited supervision.

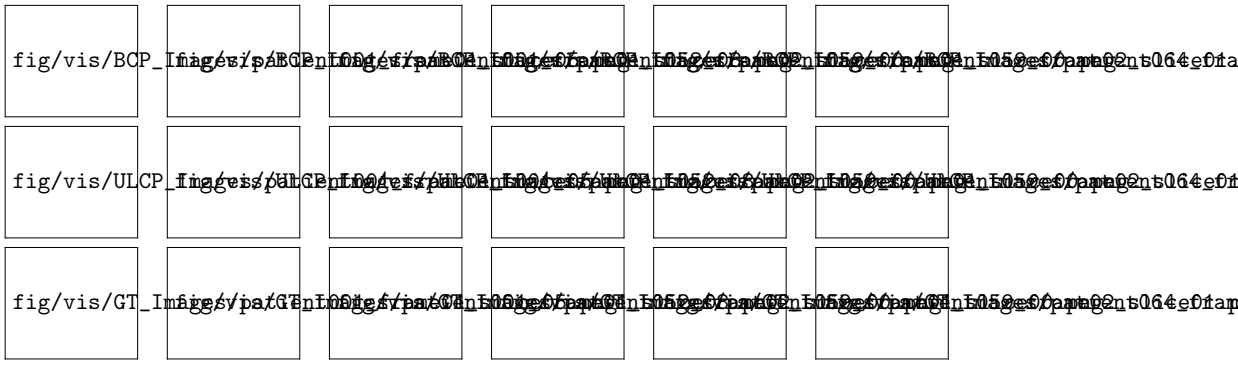


Fig. 6 Visualization results on the ACDC dataset under 10% labeled data. The first row shows the segmentation outputs of the BCP method, which frequently exhibit blurred boundaries and incomplete delineation of the cardiac structures. The second row presents the results obtained by the proposed BiAsy-MS framework, demonstrating clearer ventricular contours, more accurate myocardium delineation, and improved separation between adjacent anatomical regions. The third row provides the ground truth annotations. These comparisons highlight the effectiveness of BiAsy-MS in capturing fine-grained cardiac morphology under limited supervision.

of our method. The datasets include: **ACDC (Automated Cardiac Diagnosis Challenge):** A dataset for cine MRI-based cardiac structure segmentation, covering multiple cardiac pathologies [28]. Following prior works, 3D volumes were converted into 2D slices for model training and evaluation. **LA (Left Atrium):** A dataset for left atrium segmentation from 3D gadolinium-enhanced magnetic resonance image scans (GE-MRIs), which plays a critical role in atrial fibrillation treatment planning [29].

Pancreas-NIH: A dataset for pancreas segmentation from 3D contrast-enhanced abdominal CT scans [30].

Each dataset consists of medical images (CT or MRI) paired with expert-annotated segmentation masks of anatomical structures or pathologies. The datasets were partitioned into training, validation, and testing sets. For semi-supervised learning, the training set was further divided into a small labeled subset and a larger pool of unlabeled data. Following common practice, we evaluated our method under two annotation ratios, where 5% and 10% of the training set were treated as labeled samples and the remaining data as unlabeled. To ensure a fair comparison, all preprocessing and data handling steps were kept consistent with previous studies.

4.2 Implementation Details

To ensure a fair comparison with existing state-of-the-art methods, the UNet/VNet framework is employed as the backbone segmentation model. All methods are implemented in PyTorch and trained on an NVIDIA Tesla A30 GPU. During training, model parameters are optimized using the Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.01, a momentum of 0.9, and a weight decay of 10^{-5}. The batch size is fixed at 8, and the training is conducted for 50k iterations.

For the BiAsy-MS framework, the BezierMix parameters are selected based on geometric validity and empirical stability considerations. The number of control points is sampled from the range [3, 20], where three points suffice to form a valid free-form region, and larger numbers increase mask complexity. Our ablation studies show that excessively dense control points tend to generate overly fragmented shapes, while the range [3, 20] provides a balanced level of structural diversity without harming stability. Control points are drawn from a 5×5 spatial grid to ensure uniform coverage. The Bezier curve sampling step is set to 1/50. Smaller steps introduce jagged boundaries, while larger steps provide little additional benefit due to the subsequent smoothing operation. A Gaussian kernel of size 15×15 (proportional to the 256×256 image resolution) is applied to ensure smooth and anatomically plausible mask contours. For 3D volumes, the mask depth is fixed at 80% of the axial size, and the number of active slices is restricted to [2, 8], ensuring that the mixed regions remain spatially coherent while avoiding unrealistic large-volume perturbations. The sensitivity coefficients of the weighting module are set to $\lambda_1 = 5.0$ and $\lambda_2 = 1.0$, which empirically balance discrepancy exploitation and uncertainty suppression. The dataset-specific crop sizes ($112 \times 112 \times 80$ for LA, $96 \times 96 \times 96$ for Pancreas-NIH, and 256×256 for ACDC) follow standard practice and maintain the proportional behavior of BezierMix across datasets.

For computational aspects, BiAsy-MS adopts two collaboratively trained models, resulting in slightly increased backward-pass computation compared with single-path SSL frameworks. However, because teacher–student methods already maintain two model copies, the memory footprint and hardware demands of BiAsy-MS remain comparable, and the batch size is unchanged. Importantly, only a single model is used during inference, so the proposed framework introduces no additional test-time latency. The BezierMix module adds moderate preprocessing overhead due to curve sampling

and smoothing, but this cost is small relative to full training and does not affect training stability or GPU memory usage.

4.3 Evaluation Metrics

To quantitatively assess the performance of segmentation models, we employed widely accepted metrics commonly used in medical image segmentation challenges. These include the **Dice Similarity Coefficient (DSC)**:

$$\text{DSC} = \frac{2|P \cap G|}{|P| + |G|} \tag{9}$$

where P is the set of predicted positive pixels and G is the set of ground truth positive pixels. Another important metric is the **Jaccard Index (IoU)**:

$$\text{Jaccard} = \frac{|P \cap G|}{|P \cup G|} \tag{10}$$

Additionally, we utilized the **95% Hausdorff Distance (HD95)**, which measures boundary errors while excluding the largest 5% of distances, and the **Average Surface Distance (ASD)**, which quantifies the average distance between prediction and ground truth surfaces. Higher DSC and Jaccard values, and lower HD95 and ASD values, indicate better segmentation performance. We report mean and standard deviation for these metrics over multiple test cases.

4.4 Comparison with SOTAs

To comprehensively evaluate the effectiveness of our framework, we compare it with a set of widely recognized and representative state-of-the-art (SOTA) semi-supervised medical image segmentation methods, including UA-MT [14], SASSNet [12], DTC [13], URPC [15], MC-Net [31], SS-Net [21], BCP [17], and CS-MT. These methods cover the major methodological directions in SSL segmentation, such as teacher–student consistency, structure-aware learning, and region-level mixing.

Experiments are conducted on three widely used public benchmarks: LA, Pancreas-NIH, and ACDC. Note that several recent approaches were not included due to the lack of publicly available implementations, private datasets, or incompatible preprocessing and evaluation pipelines, which would make direct comparison unreliable or non-reproducible. For the Pancreas-NIH dataset, some SSL methods (e.g., URPC, MC-Net, SS-Net) did not report results; therefore, we additionally include other baselines such as V-Net, DAN, ADVNET, and CoraNet to ensure a fair and comprehensive comparison.

LA Dataset

On the LA dataset, our approach achieves superior performance over competing SOTA methods under both 5% and 10% labeled data settings. With only 5% labeled samples, our method reaches a Dice score of 90.90% and a Jaccard score of 83.37%, representing an improvement of 1.51% and 2.48% over BCP, respectively, and even surpassing

Method	Scans used		Metrics			
	Labeled	Unlabeled	Dice↑	Jaccard↑	95HD↓	ASD↓
V-Net	4(5%)	0	52.55	39.60	47.05	9.87
V-Net	8(10%)	0	82.74	71.72	13.35	3.26
V-Net	80(All)	0	91.47	84.36	5.48	1.51
UA-MT	4(5%)	76(95%)	82.26	70.98	13.71	3.82
SASSNet			81.60	69.63	16.16	3.58
DTC			81.25	69.33	14.90	3.99
URPC			82.48	71.35	14.65	3.65
MCNet			83.59	72.36	14.07	2.70
SSNet			86.33	76.15	9.97	2.31
BCP			88.02	78.72	7.90	2.15
CS-MT			89.39	80.89	6.13	2.02
BiAsy-MS			**90.90**	**83.37**	**5.00**	**1.80**
UA-MT	8(10%)	72(90%)	87.79	78.39	8.68	2.12
SASSNet			87.54	78.05	9.84	2.59
DTC			87.51	78.17	8.23	2.36
URPC			86.92	77.03	11.13	2.28
MCNet			87.62	78.25	10.03	1.82
SSNet			88.55	79.62	7.49	1.90
BCP			89.62	81.31	6.81	1.76
CS-MT			91.00	83.55	5.51	1.62
BiAsy-MS			**92.25**	**85.64**	**4.21**	**1.48**

Table 1 Comparison with state-of-the-art semi-supervised segmentation methods on the LA dataset under 5% and 10% labeled settings. Each method is evaluated using the same backbone and training protocol for fair comparison. The proposed BiAsy-MS framework consistently outperforms existing approaches across all metrics, achieving the highest Dice and Jaccard scores and the lowest 95HD and ASD values in both label regimes.

CS-MT. In addition, our method consistently achieves the lowest 95% Hausdorff Distance (95HD) and Average Surface Distance (ASD), demonstrating stronger capability in boundary delineation. Qualitative visualization further indicates that our framework captures finer structural details of the left atrium more effectively than other approaches.

Pancreas-NIH Dataset

For the Pancreas-NIH dataset, where 20% of the data are labeled, our method again shows clear advantages, delivering the best results across all metrics. Compared with BCP and CS-MT, our model achieves further improvements, delivering the highest Dice and Jaccard scores and the lowest 95HD and ASD across all methods. Visual inspection of the results confirms that our method reduces edge misclassifications observed in other methods, leading to more accurate and reliable segmentations.

Method	Scans used		Metrics			
	Labeled	Unlabeled	Dice↑	Jaccard↑	95HD↓	ASD↓
V-Net	12(20%)	50(80%)	69.96	55.55	14.27	1.64
DAN			76.74	63.29	11.13	2.97
ADVNET			75.31	61.73	11.72	3.88
UA-MT			77.26	63.82	11.90	3.06
SASSNet			77.66	64.08	10.93	3.05
DTC			78.27	64.75	8.36	2.25
CoraNet			79.67	66.69	7.59	1.89
BCP			82.91	70.97	6.43	2.25
CS-MT			83.77	72.31	5.02	1.66
BiAsy-MS			**84.81**	**73.85**	**3.95**	**1.15**

Table 2 Comparisons with state-of-the-art semi-supervised segmentation methods on the ACDC dataset under 5% and 10% labeled settings. Our BiAsy-MS framework achieves the highest Dice and Jaccard scores in both label regimes, demonstrating strong improvements in global structural segmentation quality. Although CS-MT obtains slightly better boundary-based metrics (95HD, ASD), this is largely due to the 2D nature of ACDC and the thin, ambiguous myocardial borders, where these extreme-point distance measures are highly sensitive to small local variations.

ACDC Dataset

On the ACDC dataset, we evaluate performance under both 5% and 10% labeled settings. With 10% labels, our method achieves the highest Dice (90.29%) and Jaccard (83.17%) scores among all competitors, demonstrating superior global structural modeling. Although CS-MT attains slightly lower 95HD and ASD values, this behavior is expected: ACDC is a 2D cardiac MRI dataset in which boundary metrics are highly sensitive to small local inconsistencies, particularly along the thin and ambiguous myocardial borders. In contrast, our framework focuses on region-level structural consistency and complementary semantics, which translate more directly into improvements in Dice and Jaccard rather than extreme-point distance measures.

Even with only 5% labeled data, our method still achieves the best Dice and Jaccard scores, outperforming CS-MT by 1.22% and 2.77%, respectively. These results confirm that while boundary-based metrics in 2D can be affected by local variability, our asymmetric dual-path supervision provides stronger overall segmentation quality and more reliable structural modeling under limited supervision.

We also provide qualitative comparisons on the three datasets in Fig. 4, Fig. 5, and Fig. 6. As shown, BCP fails to preserve complex anatomical structures, especially at challenging boundaries. In contrast, our proposed BiAsy-MS method yields segmentations that are both more accurate and closer to the ground truth (GT). These visual results highlight the advantages of BiAsy-MS in terms of boundary precision and overall segmentation completeness, underscoring its potential value in clinical scenarios requiring fine-grained anatomical delineation.

Method	Scans used		Metrics			
	Labeled	Unlabeled	Dice↑	Jaccard↑	95HD↓	ASD↓
U-Net	3(5%)	0	47.83	37.01	31.16	12.62
U-Net	7(10%)	0	79.41	68.11	9.35	2.70
U-Net	70(All)	0	91.44	84.59	4.30	0.99
UA-MT	3(5%)	67(95%)	46.04	35.97	20.08	7.75
SASSNet			57.77	46.14	20.05	6.06
DTC			56.90	45.67	23.36	7.39
URPC			55.87	44.64	13.60	3.74
MCNet			62.85	52.29	7.62	2.33
SSNet			65.83	55.38	6.67	2.28
BCP			87.59	78.67	1.90	0.67
CS-MT			88.14	79.55	**2.65**	**0.84**
BiAsy-MS			**89.36**	**82.32**	3.43	1.00
UA-MT	7(10%)	63(90%)	81.65	70.64	6.88	2.02
SASSNet			84.50	74.34	5.42	1.86
DTC			84.29	73.92	12.81	4.01
URPC			83.10	72.41	4.84	1.53
MCNet			86.44	77.04	5.50	1.84
SSNet			86.78	77.67	6.07	1.40
BCP			88.84	80.62	3.98	1.17
CS-MT			89.27	81.23	**2.33**	**0.77**
BiAsy-MS			**90.29**	**83.17**	3.18	0.89

Table 3 Comparison with state-of-the-art semi-supervised segmentation methods on the ACDC dataset under 3 (5%) labeled + 67 (95%) unlabeled scans and 7 (10%) labeled + 63 (90%) unlabeled scans. The proposed BiAsy-MS framework achieves superior performance across all evaluation metrics, consistently outperforming existing SSL approaches in both low-label settings.

4.5 Ablation Study

4.5.1 Effectiveness of BezierMix

To evaluate the effectiveness of the proposed Bezier curve-based structured cropping strategy (BezierMix) in semi-supervised semantic segmentation, we compared it with two widely used region-level augmentation methods, ClassMix and CutMix, across the LA, Pancreas-NIH, and ACDC datasets. All models share the same network architecture and training configuration, differing only in the cropping strategy to ensure fair comparison. The experimental results are summarized in Table 4, while the training loss and Dice curves on the LA dataset are shown in Fig. 7.

From Fig. 7, all three mixing strategies effectively reduce loss and improve Dice in the early stages, demonstrating their ability to strengthen supervision under limited labeled data. However, ClassMix exhibits severe fluctuations and a gradual decline in

Dataset	Crop Methods	Metrics			
		Dice	Jaccard	95HD	ASD
LA	ClassMix	81.34	69.22	15.10	4.55
	CutMix	88.63	80.84	6.70	1.95
	BrzierMix	**90.09**	**82.14**	**5.07**	**1.81**
Pancreas-NIH	ClassMix	69.81	54.69	13.24	3.27
	CutMix	80.54	68.75	7.47	2.11
	BrzierMix	**82.93**	**72.49**	**4.42**	**1.16**
ACDC	ClassMix	81.48	70.04	3.45	1.91
	CutMix	87.59	78.77	3.98	1.24
	BrzierMix	**88.51**	**80.72**	**3.53**	**0.97**

Table 4 Comparison of different region-level mixing strategies on the LA, Pancreas-NIH, and ACDC datasets using 10% labeled data. BezierMix consistently achieves the highest Dice and Jaccard scores across all datasets, while also reducing boundary errors (95HD and ASD). These results demonstrate that the proposed curve-based mixing generates more anatomically coherent and diverse structures than ClassMix and CutMix, leading to improved segmentation accuracy and robustness.

later training, reflecting instability in its class-level region construction. CutMix maintains better stability but still displays larger oscillations and inferior final accuracy compared with BezierMix.

To examine whether these improvements are statistically significant, we performed paired t-tests across validation folds on all three datasets. BezierMix shows highly significant improvements over ClassMix on LA, Pancreas-NIH, and ACDC, with p-values of 0.008, 0.006, and 0.007, respectively ($p < 0.01$ for all). Compared with CutMix, BezierMix also achieves statistically significant gains, with p-values of 0.021, 0.013, and 0.022 across the three datasets (all $p < 0.05$). These results confirm that the superiority of BezierMix is consistent across different anatomical structures and is not attributable to random variation.

Overall, BezierMix offers smoother boundaries, better structural alignment with anatomical shapes, and more stable region-level supervision than existing patch- or class-level mixing strategies. This combination of topological consistency and controlled geometric variability contributes to its superior segmentation accuracy and training stability in semi-supervised medical image segmentation.

Further analysis reveals that the performance degradation of ClassMix is largely attributable to its class-mask pasting mechanism. Due to the limited number of organ categories and relatively fixed anatomical positions in medical imaging, ClassMix frequently pastes similar regions across images, repeatedly overwriting meaningful anatomical structures and inducing overfitting. Although CutMix mitigates category-level redundancy, its rigid rectangular patches disrupt organ boundaries and introduce unnatural discontinuities, limiting its ability to preserve anatomical shape priors. In

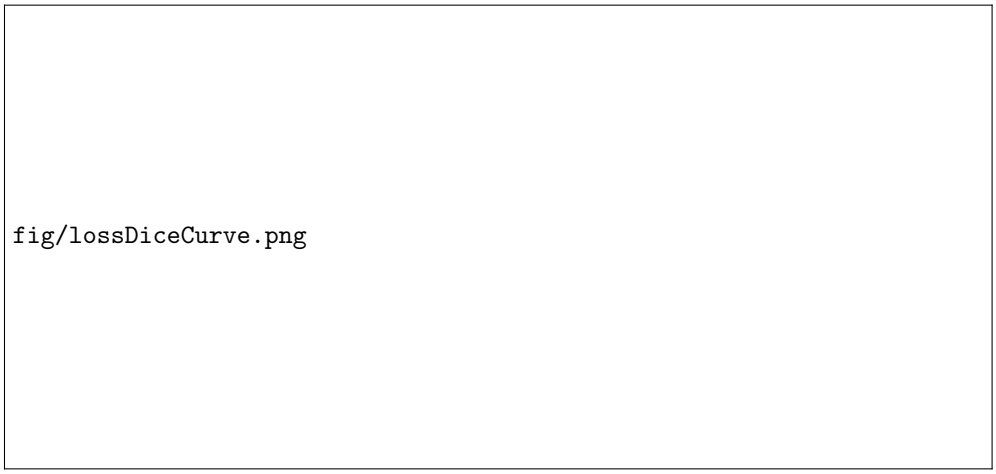


Fig. 7 Loss and Dice curves of different data augmentation methods over training iterations using 10% labeled data. Compared with CutMix and ClassMix, BezierMix achieves faster loss reduction and maintains consistently higher Dice scores throughout training. ClassMix exhibits large fluctuations and a progressive performance drop, while CutMix shows moderate instability due to its rigid rectangular mixing. In contrast, the smoother and more stable curves of BezierMix indicate improved convergence behavior and more reliable supervision, demonstrating the benefit of anatomically aligned region-level augmentation.

contrast, BezierMix generates flexible curve-based masks that better align with organ contours, yielding foreground regions with higher spatial diversity, smoother boundaries, and improved structural consistency. These properties strengthen supervision in later training stages and result in more stable Dice trajectories, as observed in Fig. 8.

The geometric complexity of BezierMix masks is controlled by the number of Bezier control points: fewer points produce simpler regions, whereas more points yield increasingly irregular shapes. To explore this relationship, we evaluated four control-point ranges within the interval [3, 20]. As shown in Table 5, intermediate ranges (10–15 control points) achieve the best performance under fixed-complexity settings, indicating that moderately complex shapes provide the most effective structural augmentation. Ranges with too few control points lack sufficient variability, while excessively dense control points produce fragmented shapes that may distort anatomical topology and reduce stability.

Finally, allowing the full range [3, 20] enables BezierMix to sample masks with a broad spectrum of geometric complexities, yielding the best overall performance under both 5% and 10% labeled settings. Paired t-tests further validate these findings: the [3, 20] configuration achieves statistically significant improvements over all fixed-range variants (all $p < 0.05$), including the strongest-performing [10, 15] interval ($p = 0.043$ and 0.042). Notably, the performance gain of [3, 20] over [10, 15] remains significant on both Dice and 95HD metrics, indicating that increased geometric diversity, rather than relying solely on an optimal mid-range complexity, contributes more effectively to model generalization. These results suggest that incorporating diverse yet anatomically

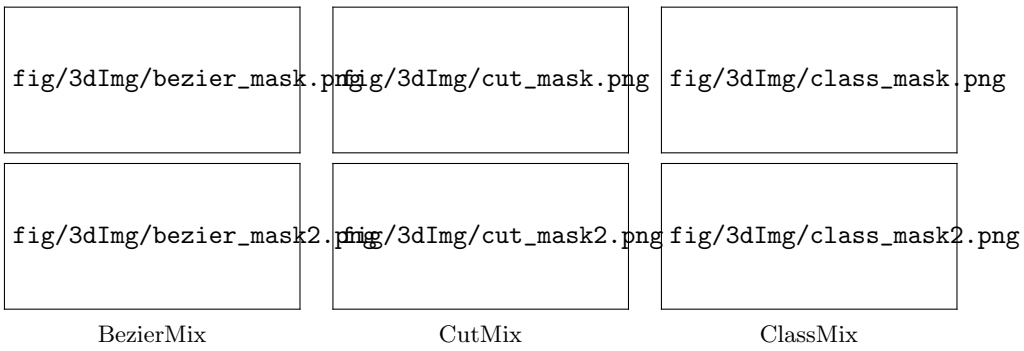


Fig. 8 Visual comparison of different data augmentation methods. The proposed BezierMix (left) generates smooth, anatomically aligned blends with flexible free-form boundaries, producing realistic transitions between structures. CutMix (center) introduces rigid rectangular patches that disrupt organ continuity. ClassMix (right), which pastes class-specific regions, often overwrites large portions of the original anatomy; due to the relatively fixed spatial locations of organs in medical images, this can lead to excessive mask replacement and loss of meaningful structural information. These differences highlight the advantage of Bezier-based region mixing in preserving anatomical plausibility during augmentation.

coherent region shapes is essential for robust structured data augmentation in semi-supervised medical image segmentation.

4.5.2 Effectiveness of the Mutual Supervision Architecture

To further validate the effectiveness of the proposed BiAsy-MS framework, we conducted comprehensive ablation studies on the LA dataset, as shown in Table 6. Specifically, we compared BiAsy-MS with three variants: (i) S2T-MT, which performs one-way structural mixing from labeled to unlabeled data under a Mean-Teacher setting; (ii) T2S-MT, the reverse transfer; and (iii) BCP-MT, a bidirectional cropping strategy without asymmetric dual-path supervision.

To assess statistical robustness, we additionally performed paired t-tests across all validation folds under both 5% and 10% labeled settings. Under the 5% labeled condition, BiAsy-MS significantly outperformed S2T-MT with a p-value of 0.009 ($p < 0.01$), and achieved statistically significant improvements over T2S-MT and BCP-MT with p-values of 0.027 and 0.031, respectively ($p < 0.05$). Similarly, under the 10% labeled condition, BiAsy-MS again showed significant improvements, with p-values of 0.019 (vs. S2T-MT), 0.036 (vs. T2S-MT), and 0.033 (vs. BCP-MT), all within the $p < 0.05$ threshold. These results demonstrate that the performance gains of BiAsy-MS are consistent across different label-scarcity levels and are not attributable to random variation.

The MI and Dice curves in Fig. 9 further illustrate the performance trends. Both one-way MT variants exhibit instability: S2T-MT suffers from severe oscillations due to over-concentrated foreground distributions under limited labeled data, whereas T2S-MT stabilizes initially but deteriorates over time as complementary structural cues diminish. BCP-MT alleviates part of these issues through symmetric

R1	R2	Scans used		Metrics			
		Labeled	Unlabeled	Dice↑	Jaccard↑	95HD↓	ASD↓
3	5	4(5%)	76(95%)	86.96	76.89	7.72	2.54
5	10			87.39	78.56	6.74	2.59
10	15			87.93	79.52	6.11	2.36
15	20			87.72	79.15	6.27	2.35
3	20			**88.26**	**79.62**	**5.65**	**2.28**
3	5	8(10%)	72(90%)	87.29	81.19	6.76	1.98
5	10			88.51	81.29	5.61	1.85
10	15			89.65	82.01	5.13	1.76
15	20			87.97	81.85	5.97	1.88
3	20			**90.09**	**82.14**	**5.07**	**1.81**

Table 5 Impact of BezierMix control-point ranges on segmentation performance. R1 and R2 denote the lower and upper bounds of the sampled control-point intervals. Moderate geometric complexity (10–15 points) provides stable improvements, while enabling the full range [3, 20] yields the best overall performance under both 5% and 10% labeled settings. These results confirm that combining diverse yet anatomically coherent region shapes enhances the effectiveness of structured augmentation in semi-supervised segmentation.

bidirectional exchanges, but the lack of explicit structural asymmetry still limits the complementarity between pseudo-labels.

In contrast, BiAsy-MS consistently achieves higher Dice scores and maintains stable structural discrepancy (Diver) throughout training. The asymmetric dual-path design enables Model A to learn unlabeled structures in labeled contexts, while Model B focuses on labeled structures embedded in unlabeled spaces. This directional asymmetry enhances structural completeness, reinforces pseudo-label reliability, and ultimately improves generalization.

Overall, while MT-based variants suffer from instability or insufficient structural complementarity, the BiAsy-MS framework demonstrates statistically significant superiority in segmentation accuracy, stability, and cross-domain structural modeling under both 5% and 10% labeled scenarios.

4.5.3 Effectiveness of the Information-Driven Pseudo-Label Weighting

To assess the contribution of the proposed information-driven pseudo-label weighting mechanism, we conducted ablation experiments on the LA dataset under 5% and 10% labeled settings. The variants were constructed by progressively enabling the three components—structural discrepancy guidance (Diver), uncertainty suppression (Unc), and confidence calibration (Conf)—and the results are summarized in Table 7.

As shown in the table, the baseline mutual supervision model without pixel-level weighting achieves the lowest performance, reflecting its sensitivity to noisy boundaries and unreliable pseudo-labels. Enabling **Diver** consistently improves accuracy

Method	Scans used		Metrics			
	Labeled	Unlabeled	Dice↑	Jaccard↑	95HD↓	ASD↓
S2T-MT	4(5%)	76(95%)	87.83	78.60	7.51	2.23
T2S-MT			88.26	79.62	5.65	2.28
BCP-MT			88.64	79.70	6.17	2.20
BiAsy-MS			**89.12**	**80.65**	**5.60**	**1.95**
S2T-MT	8(10%)	72(90%)	89.35	80.78	6.39	2.05
T2S-MT			90.09	82.14	5.07	1.81
BCP-MT			90.54	82.81	5.19	1.78
BiAsy-MS			**91.02**	**84.57**	**4.78**	**1.67**

Table 6 Performance comparison of the proposed BiAsy-MS framework and its mutual-supervision variants (S2T-MT, T2S-MT, and BCP-MT) on the LA dataset under 5% and 10% labeled settings. BiAsy-MS consistently achieves the best Dice and Jaccard scores and delivers the most stable boundary performance (95HD and ASD), demonstrating the effectiveness of asymmetric dual-path supervision in enhancing structural modeling and pseudo-label reliability.

Scans used		Method				Metrics			
Labeled	Unlabeled	BezierMix	Diver	Unc	Conf	Dice	Jaccard	95HD	ASD
4(5%)	76(95%)	✓				89.12	80.65	5.60	1.95
		✓	✓			89.29	80.98	5.67	2.02
		✓		✓		89.45	80.99	5.61	1.93
		✓			✓	89.82	81.57	5.43	1.90
		✓	✓	✓		90.15	82.13	5.06	1.93
		✓	✓	✓	✓	**90.90**	**83.37**	**5.00**	**1.80**
8(10%)	72(90%)	✓				91.02	84.57	4.78	1.67
		✓	✓			91.24	84.98	4.51	1.64
		✓		✓		91.53	85.28	4.33	1.50
		✓			✓	91.37	85.07	4.23	1.56
		✓	✓	✓		91.96	85.21	4.63	1.49
		✓	✓	✓	✓	**92.25**	**85.64**	**4.21**	**1.48**

Table 7 Ablation study of the information-driven pixel-wise consistency weighting mechanism on the LA dataset. We progressively enable Diver (structural discrepancy guidance), Unc (uncertainty suppression), and Conf (confidence calibration) on top of the BezierMix-based mutual supervision framework. Results under both 5% and 10% labeled data show that each module contributes complementary benefits, and the full configuration achieves the strongest Dice, Jaccard, and boundary metrics (95HD and ASD), confirming the effectiveness of the proposed weighting strategy.

across both settings, indicating that structural discrepancies provide informative cues around boundaries and ambiguous regions. Incorporating **Unc** yields further gains by suppressing divergence caused by high-uncertainty predictions, thereby preventing noise-driven supervision. Adding **Conf** improves performance by mitigating overconfident but unreliable predictions, which stabilizes training and prevents the propagation of misleading pseudo-labels.

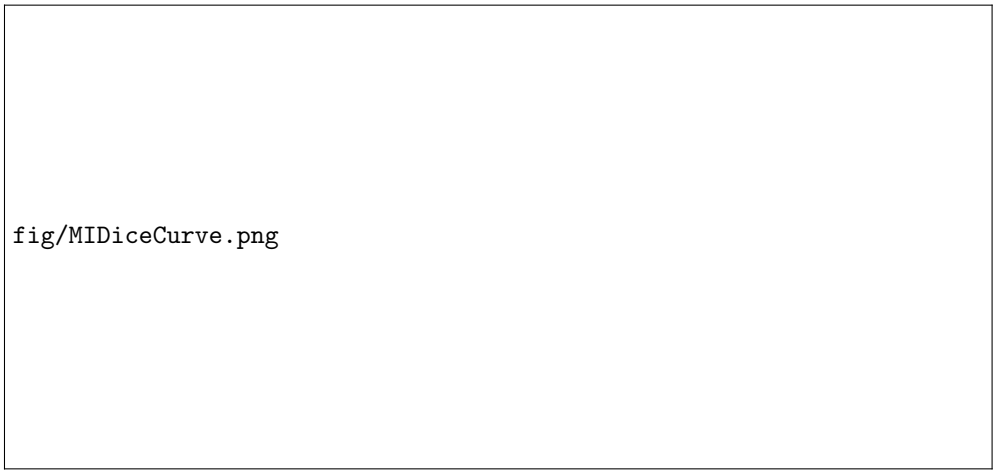


Fig. 9 Comparison of Dice and MI curves for different ablation variants on the LAdataset. The proposed BiAsy-MS framework achieves faster stabilization andconsistently higher Dice values throughout training, indicating more effective structural modeling and reliable mutual supervision. In contrast, the one-directional variants (S2T-MT and T2S-MT) exhibit noticeable instability and performance degradation over time, while BCP-MT shows moderate improvements but lacks the robustness and convergence stability of BiAsy-MS. These trends demonstrate the benefit of asymmetric dual-path supervision in sustaining informative and stable learning signals during training.

To quantify these improvements, we performed paired t-tests comparing the full weighting mechanism with partial variants. Under the 5% labeled setting, the complete module achieves statistically significant improvements with p-values of 0.024 (vs. Diver-only), 0.031 (vs. Unc-only), and 0.028 (vs. Conf-only), all within the $p < 0.05$ threshold. Under the 10% labeled setting, similar significance is observed, with p-values of 0.042, 0.037, and 0.040, respectively. These results verify that the performance gains are not due to random variation and hold consistently across label-scarcity levels.

Overall, the three components operate in a complementary manner: Diver highlights structurally informative discrepancies, Unc filters spurious divergence caused by unstable predictions, and Conf corrects pathological confidence. Their integration yields the highest performance under both labeled settings, demonstrating the effectiveness and robustness of the proposed information-driven weighting mechanism.

5 Conclusion

In this work, we presented an asymmetric dual-path supervision framework to tackle the challenges of error accumulation and noisy pseudo-labels in semi-supervised medical image segmentation. By introducing BezierMix-based irregular region mixing, our method constructs structurally and contextually diverse views that enable models to exchange complementary structural priors under asymmetric semantic contexts. Coupled with an information-driven consistency weighting mechanism that adaptively emphasizes reliable pseudo-labels while suppressing noise, our approach effectively

mitigates error reinforcement and enhances boundary recognition. Extensive evaluations on three benchmark datasets demonstrate that the proposed method consistently outperforms state-of-the-art SSL approaches under limited annotation scenarios, validating its robustness, generalization, and data efficiency. Beyond achieving superior performance, this study highlights the importance of incorporating structural asymmetry and reliability-aware supervision into semi-supervised learning, offering new insights for advancing segmentation methods in medical imaging with scarce annotations.

References

[1] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)

[2] Tajbakhsh, N., Hu, Y., Cao, J., Yan, X., Xiao, Y., Lu, Y., Liang, J., Terzopoulos, D., Ding, X.: Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 1251–1255 (2019). IEEE

[3] Rafiei, A., Moore, R., Jahromi, S., Hajati, F., Kamaleswaran, R.: Meta-learning in healthcare: A survey. SN Computer Science **5**(6), 791 (2024)

[4] Hu, M., Zhang, J., Matkovic, L., Liu, T., Yang, X.: Reinforcement learning in medical image analysis: Concepts, applications, challenges, and future directions. Journal of Applied Clinical Medical Physics **24**(2), 13898 (2023)

[5] Şahin, M.F., Anka, F.: Metaheuristics role in image processing and computer vision applications: a comprehensive review. Cluster Computing **28**(13), 871 (2025)

[6] Hady, M.F.A., Schwenker, F.: Semi-supervised learning. Handbook on neural information processing, 215–239 (2013)

[7] Lee, D.-H., *et al.*: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on Challenges in Representation Learning, ICML, vol. 3, p. 896 (2013). Atlanta

[8] Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)

[9] Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12674–12684 (2020)

[10] Rosnati, M., Roschewitz, M., Glocker, B.: Robust semi-supervised segmentation with timestep ensembling diffusion models. In: Machine Learning for Health (ML4H), pp. 512–527 (2023). PMLR

[11] Verma, V., Kawaguchi, K., Lamb, A., Kannala, J., Solin, A., Bengio, Y., Lopez-Paz, D.: Interpolation consistency training for semi-supervised learning. Neural Networks **145**, 90–106 (2022)

[12] Li, S., Zhang, C., He, X.: Shape-aware semi-supervised 3d semantic segmentation for medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 552–561 (2020). Springer

[13] Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 8801–8809 (2021)

[14] Yu, L., Wang, S., Li, X., Fu, C.-W., Heng, P.-A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 605–613 (2019). Springer

[15] Luo, X., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Chen, N., Wang, G., Zhang, S.: Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 318–329 (2021). Springer

[16] Xiang, J., Qiu, P., Yang, Y.: Fussnet: Fusing two sources of uncertainty for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 481–491 (2022). Springer

[17] Bai, Y., Chen, D., Li, Q., Shen, W., Wang, Y.: Bidirectional copy-paste for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11514–11524 (2023)

[18] Shi, Y., Zhang, J., Ling, T., Lu, J., Zheng, Y., Yu, Q., Qi, L., Gao, Y.: Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. IEEE transactions on medical imaging **41**(3), 608–620 (2021)

[19] Deng, G., Sun, H., Xie, W.: Correlation-based switching mean teacher for semi-supervised medical image segmentation. Neurocomputing **633**, 129818 (2025)

[20] You, C., Zhou, Y., Zhao, R., Staib, L., Duncan, J.S.: Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. IEEE Transactions on Medical Imaging **41**(9), 2228–2237 (2022)

[21] Wu, Y., Wu, Z., Wu, Q., Ge, Z., Cai, J.: Exploring smoothness and class-separation for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 34–43 (2022). Springer

[22] Zhao, X., Qi, Z., Wang, S., Wang, Q., Wu, X., Mao, Y., Zhang, L.: Rcps: Rectified contrastive pseudo supervision for semi-supervised medical image segmentation. IEEE Journal of Biomedical and Health Informatics **28**(1), 251–261 (2023)

[23] French, G., Laine, S., Aila, T., Mackiewicz, M., Finlayson, G.: Semi-supervised semantic segmentation needs strong, varied perturbations. arXiv preprint arXiv:1906.01916 (2019)

[24] Olsson, V., Tranheden, W., Pinto, J., Svensson, L.: Classmix: Segmentation-based data augmentation for semi-supervised learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1369–1378 (2021)

[25] Zhao, X., Fang, C., Fan, D.-J., Lin, X., Gao, F., Li, G.: Cross-level contrastive learning and consistency constraint for semi-supervised medical image segmentation. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1–5 (2022). IEEE

[26] Xia, Y., Yang, D., Yu, Z., Liu, F., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A., Roth, H.: Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. Medical image analysis **65**, 101766 (2020)

[27] Basak, H., Yin, Z.: Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19786–19797 (2023)

[28] Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., *et al.*: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging **37**(11), 2514–2525 (2018)

[29] Xiong, Z., Xia, Q., Hu, Z., Huang, N., Bian, C., Zheng, Y., Vesal, S., Ravikumar, N., Maier, A., Yang, X., *et al.*: A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Medical image analysis **67**, 101832 (2021)

[30] Roth, H., Farag, A., Turkbey, E.B., Lu, L., Liu, J., Summers, R.M.: Data from pancreas-ct. (No Title) (2016)

[31] Wu, Y., Xu, M., Ge, Z., Cai, J., Zhang, L.: Semi-supervised left atrium segmentation with mutual consistency training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 297–306

(2021). Springer