This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

AffinityUMamba: Uncertainty-Aware Medical Image Segmentation via Probabilistic Weak Supervision Beyond Gold-Standard Annotations

Yukun Zhang[1], Guisheng Wang[2], William Henry Nailon[3], and Kun Cheng[1]*

[1] Beijing University of Posts and Telecommunications, Beijing, China
{zyk,kcheng}@bupt.edu.cn
[2] Department of Radiology, Third Medical Center of Chinese PLA General Hospital, Beijing, China
wangguisheng@301hospital.com.cn
[3] NHS Lothian & The University of Edinburgh, Department of Oncology Physics & The School of Engineering, Edinburgh, United Kingdom
w.nailon@ed.ac.uk

Abstract. Owing to its superior soft tissue contrast, Magnetic Resonance Imaging (MRI) has become a cornerstone modality in clinical practice. This prominence has driven extensive research on MRI-based segmentation, supported by the proliferation of publicly available benchmark datasets. Despite employing multi-expert consensus protocols to ensure annotation quality in public datasets, the inherent label noise, particularly prevalent at lesion boundary regions remains unavoidable. To address this fundamental challenge, we introduce a novel machine learning paradigm that reframes dataset annotations as probabilistic weak supervision rather than deterministic gold standards. We proposed AffinityUMamba, a novel dual-branch Unet-like framework that synergistically integrates convolutional operations with state space models, leveraging local feature coherence and global contextual agreement. And a Local Affinity-guided Label Refinement (LALR) module to identify potential noisy labels in the training data and produce refined pseudo labels. A unified uncertainty constraint paradigm combining margin-based logit smoothing with local affinity refinement, enabling simultaneous optimization of segmentation accuracy and confidence calibration. Training is stabilized through a composite objective combining topological preservation constraints with margin-aware uncertainty penalization, enabling joint optimization of structural coherence and detail fidelity. We comprehensively evaluated the proposed method on 12 public datasets spanning multiple modalities: 10 MRI, 1 Ultrasound, and 1 CT. The results of our experiments demonstrate an improved segmentation performance and reduced prediction uncertainty.

Keywords: Medical Image Segmentation · Uncertainty · Weak Supervision

* Corresponding author.

1 Introduction

Medical image analysis has evolved from rule-based workflows to deep learning paradigms, driven by architectures like nnU-Net [10]. While early research emphasized texture classification [5] and radiomics [14], contemporary efforts focus on lesion segmentation where precision impacts clinical decisions [24]. Magnetic Resonance Imaging (MRI) dominates soft tissue analysis due to unparalleled contrast resolution for neurological [6], cardiac [3], and abdominal structures [25]. However, its sensitivity to tissue heterogeneity introduces critical challenges: ambiguous lesion boundaries and intensity non-uniformities propagate annotation inconsistencies, even in multi-expert consensus protocols [12].

From our perspective, the public datasets utilize voter/averaging manner to obtain a sub-optimal gold standard, while conventional supervised learning frameworks treat annotations as deterministic gold standards, ignoring the inherent uncertainty introduced in medical image manual labeling. This approach propagates toxic training errors, particularly at boundary regions where the highest inter-observer variability is commonly discovered. Gal et al. [13] defined the aleatoric and epistemic uncertainty that were introduced in the common pipeline of deep learning segmentation models. Many researchers focus on the model architecture and training schemes to reduce epistemic (model) uncertainty, but very limited work pays attention to aleatoric (data) uncertainty in medical image segmentation.

Our philosophy here is to acknowledge the gaze annotation effort from experienced oncologists/doctors at the homogeneous interior of the region of interest (ROI), while delivering discriminative knowledge through end-to-end weakly supervised training at inhomogeneous boundary regions. Affinity has been proven beneficial for improving segmentation in weakly supervised semantic segmentation [1], which refers to the spatial and semantic relationships between neighboring pixels that share similar characteristics in medical images. A multi-task framework that uses auxiliary tasks and cross-task affinity learning to enhance weakly-supervised semantic segmentation is proposed using only image-level ground-truth labels [28]. Another end-to-end weakly-supervised semantic segmentation method based on Transformers, leverages attention mechanisms to learn affinity and refine pseudo labels for improved segmentation accuracy [23]. State Space Sequence Models (SSMs) Mamba [7], provide an effective alternative by modeling long-range dependencies and complex relationships in sequential data. When combined with CNNs, SSMs are capable of retaining local details and boundary features while simultaneously capturing global information, which proves advantageous in addressing complex boundaries and potential label noise [15].

To address these challenges, we propose a paradigm shift from deterministic annotation to probabilistic weak supervision, weakening pixel-level supervision from ground truth labels at boundary regions. The main contributions of this work are threefold: (1) We proposed AffinityUMamba, a novel dual-branch Unet-like framework that synergistically integrates convolutional operations with state space models, leveraging local feature coherence and global contextual agree-

ment; (2) We proposed the Local Affinity-guided Label Refinement (LALR) module to identify potential noisy label in the training data and produce refined pseudo labels; (3) A unified uncertainty constraint paradigm combining margin-based logit smoothing with local affinity refinement, enabling simultaneous optimization of segmentation accuracy and confidence calibration.

2 Method

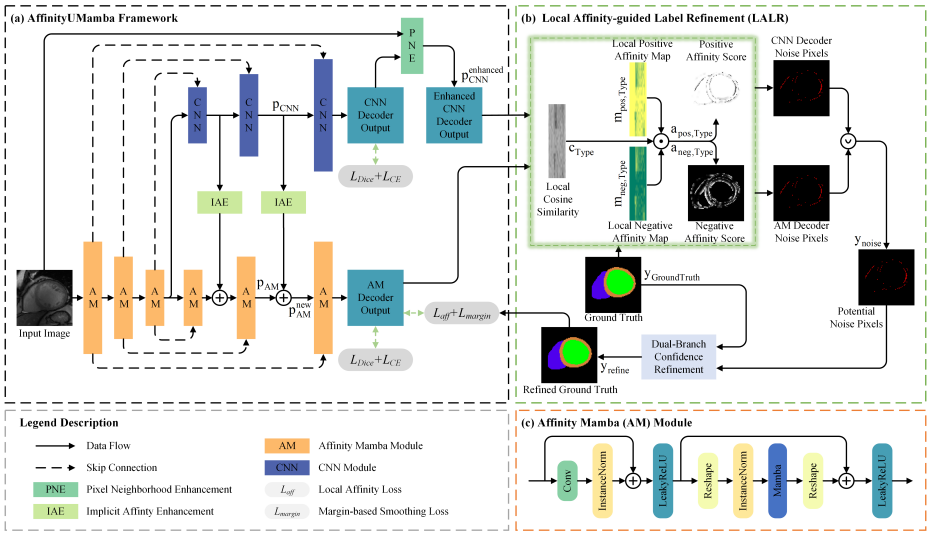

Fig. 1. Illustration of our AffinityUMamba framework.

Our framework integrates collaborative modules for probabilistic weak supervision, as shown in Fig. 1. Let $x \in \mathbb{R}^{H\times W\times D}$ be the input image, and $y \in \mathbb{R}^{H\times W\times D}$ the corresponding ground truth or predicted category map. The feature map is $p \in \mathbb{R}^{C\times H\times W\times D}$, where C is the number of classes, and H, W, and D are the height, width, and depth of the image, respectively. Other variables include the local affinity map $m \in \mathbb{R}^{27\times H\times W\times D}$, where $k = 3$ denotes the size of the local $3 \times 3 \times 3$ neighborhood, and the neighborhood $\mathcal{N}^i(k)$ refers to the region at position i. Additionally, the index j refers to a voxel within $\mathcal{N}^i(k)$, affinity scores $a \in \mathbb{R}^{H\times W\times D}$, and the local cosine similarity $c \in \mathbb{R}^{27\times H\times W\times D}$. To ensure numerical stability in the computations, a small value ϵ is added during division operations.

2.1 Implicit Affinity Enhancement (IAE)

Affinity Mamba (AM) module as shown in Fig. 1(c). The IAE module enhances the feature discriminability of the AM branch by leveraging pixel affinity re-

lationships from the CNN branch, aligning CNN's local semantics with AM's global semantics through adaptive neighborhood interactions.

$$p_{\mathrm{AM}}^{i,\mathrm{new}} = \sum_{j \in \mathcal{N}^i(k)} \mathbb{I}\left(p_{\mathrm{CNN}}^{j,i} \geq \frac{1}{|\mathcal{N}^i(k)|} \sum_{j \in \mathcal{N}^i(k)} p_{\mathrm{CNN}}^{j,i} \right) p_{\mathrm{AM}}^{j} + p_{\mathrm{AM}}^{i} \tag{1}$$

where $\mathbb{I}$ is the indicator function, returning 1 if the condition is true and 0 otherwise. The mechanism encodes semantic relationships among neighboring pixels by comparing each pixel's value to the neighborhood's average, enhancing the corresponding features in AM based on these implicit affinities. This dual-branch Unet co-regularization framework improves boundary delineation by ensuring only statistically significant correlations influence the update, ensuring reliable tissue differentiation despite intensity inhomogeneity.

2.2 Pixel Neighborhood Enhancement (PNE)

We assume that image signals remain relatively reliable despite noisy annotations, containing essential tissue boundary information. This module enhances local continuity in CNN decoder output features by directly leveraging multi-scale intensity relationships from the input image, addressing ground truth inaccuracies caused by inter-observer variability. By constructing adaptive Gaussian kernels at different scales, it explicitly captures pixel correlations based on raw intensity patterns, complementing CNN's inherent local modeling strengths. The computation involves:

$$\begin{aligned} \hat{x}^{j,i} &= \exp\left(-\frac{|x^{j,i} - x^i|^2}{\sigma + \epsilon} \right), \quad \hat{x}^{j,i} \leftarrow \frac{\hat{x}^{j,i}}{\sum_{j \in \mathcal{N}^i(k)} \hat{x}^{j,i}} \\ p_{\mathrm{CNN}}^{i,enhanced} &= \sum_{k \in \{3,5,7\}} \frac{1/\sqrt{k}}{\sum_{k'} 1/\sqrt{k'}} \sum_{j \in \mathcal{N}^i(k)} \hat{x}^{j,i} p_{\mathrm{CNN}}^{j,i} + p_{\mathrm{CNN}}^{i} \end{aligned} \tag{2}$$

It is expected to maintain structural coherence while improving incomplete boundaries which may be caused by partial volume effects.

2.3 Local Affinity-guided Label Refinement (LALR)

The LALR module derives the noise label suggestion from both CNN and AM decoders, using affinity-guided annotation inconsistencies where local feature coherence and global contextual agreement are combined.

We quantify the intrinsic affinity credibility between neighboring features through local cosine similarity, which evaluates the semantic consistency of the feature. This is formulated as:

$$c_{\mathrm{Type}}^{j,i} = \frac{p_{\mathrm{Type}}^{j,i} \cdot p_{\mathrm{Type}}^{i}}{|p_{\mathrm{Type}}^{j,i}| \cdot |p_{\mathrm{Type}}^{i}| + \epsilon}, \quad \mathrm{Type} \in \{\mathrm{CNN}, \mathrm{AM}\} \tag{3}$$

The dual-decoder verification mechanism operates through dynamically generated local affinity maps that encode the semantic relationships among neighboring pixels:

$$m^{j,i}_{\text{pos,Type}} = \mathbb{I}(y^j = y^i_{\text{Type}}), \quad m^{j,i}_{\text{neg,Type}} = 1 - m^{j,i}_{\text{pos,Type}} \tag{4}$$

where y^j denotes the category at position i from the ground truth (for LALR in section 2.3) or refined ground truth (for following Local Affinity Loss in section 2.4), and y^i_{Type} denotes the category at position i from the CNN decoder prediction, AM decoder prediction or refined ground truth.

Potential noise pixels are identified through the affinity score assessment, which detects pixels where both decoders exhibit conflicting affinity patterns—low confidence in same category feature alignment but high confidence in different category dissimilarity. The calculations are performed as follows:

$$\begin{aligned} a^i_{\text{pos,Type}} &= \frac{\sum_{j\in\mathcal{N}^i(k)} c^{j,i}_{\text{Type}} m^{j,i}_{\text{pos,Type}}}{\sum_{j\in\mathcal{N}^i(k)} m^{j,i}_{\text{pos,Type}} + \epsilon}, \; a^i_{\text{neg,Type}} = \frac{\sum_{j\in\mathcal{N}^i(k)} c^{j,i}_{\text{Type}} m^{j,i}_{\text{neg,Type}}}{\sum_{j\in\mathcal{N}^i(k)} m^{j,i}_{\text{neg,Type}} + \epsilon} \\ y^i_{\text{noise}} &= \bigcup_{\text{Type}} \left(a^i_{\text{pos,Type}} < \frac{1}{2} \right) \cap \left(a^i_{\text{neg,Type}} > \frac{1}{2} \right) \end{aligned} \tag{5}$$

The refinement process employs entropy-weighted decision fusion to integrate dual-branch confidence:

$$y^i_{\text{refine}} = y^i_{\text{noise}} \cdot \mathbb{I}(E^i_{\text{CNN}} > E^i_{\text{AM}}) \cdot y^i_{\text{AM}} + (1 - y^i_{\text{noise}}) \cdot y^i_{\text{GroundTruth}} \tag{6}$$

where $E^i_{\text{Type}} = \text{Entropy}(p^i_{\text{Type}})$ measures prediction uncertainty. When the CNN decoder branch exhibits higher entropy (greater uncertainty in local predictions), the AM decoder branch's prediction y^i_{AM} is prioritized for correction.

2.4 Loss Function

Local Affinity Loss enhances feature discriminability by enforcing semantic consistency between positive/negative samples in the AM decoder predictions using the refined ground truths. When computing the local positive and negative affinity maps, $m^{j,i}_{\text{pos,Type}}$ and $m^{j,i}_{\text{neg,Type}}$, y^j represents the label at position i from the refined ground truth y^i_{refine} (as detailed in section 2.3), ensuring semantic consistency:

$$\mathcal{L}_{\text{aff}} = \sum_i \frac{\sum_{j\in\mathcal{N}^i(k)} m^{j,i}_{\text{pos,refine}} m^{j,i}_{\text{neg,AM}}}{\sqrt{\sum_{j\in\mathcal{N}^i(k)} m^{j,i}_{\text{pos,refine}}}} + \frac{\sum_{j\in\mathcal{N}^i(k)} m^{j,i}_{\text{neg,refine}} m^{j,i}_{\text{pos,AM}}}{\sqrt{\sum_{j\in\mathcal{N}^i(k)} m^{j,i}_{\text{neg,refine}}}} \tag{7}$$

Margin-based Smoothing Loss inspired by previous work [20], we introduce dynamically adjusted, variance-aware boundary constraint, reducing ex-

treme prediction values and alleviating overconfidence in noisy boundaries:

$$
\begin{aligned}
v^i &= \sqrt{\mathrm{ReLU}\left(\sum_{j\in\mathcal{N}^i(k)} |p_{\mathrm{AM}}^{j,i}|^2 - |\sum_{j\in\mathcal{N}^i(k)} p_{\mathrm{AM}}^{j,i}|^2\right)} \\
\mathcal{L}_{\mathrm{margin}} &= \sum_i \mathrm{ReLU}\left(\max_c(p^i) - p^i - 8\cdot\frac{v^i - \min(v^i)}{\max(v^i)-\min(v^i)}\right)
\end{aligned}
\tag{8}
$$

where constant 8 is set following [20].

Overall Objective: The two losses complement the traditional Dice-CE loss:

$$
\mathcal{L} = \underbrace{\mathcal{L}_{\mathrm{Dice}} + \mathcal{L}_{\mathrm{CE}}}_{\text{Supervision}} + \underbrace{0.1(\mathcal{L}_{\mathrm{Dice}}^{\mathrm{refine}} + \mathcal{L}_{\mathrm{CE}}^{\mathrm{refine}})}_{\text{Refined Supervision}} + 0.1\mathcal{L}_{\mathrm{margin}} + 0.01\mathcal{L}_{\mathrm{aff}} \tag{9}
$$

'Supervision' uses ground truth to guide CNN and AM decoder predictions, while 'Refined Supervision' uses refined ground truth for both branches. This joint optimization balances model calibration and discrimination performance.

3 Experiments and Results

3.1 Datasets and Implementations

We validated our model on 12 public medical image segmentation datasets, including ACDC [4] (MRI: Left Ventricle, Right Ventricle, Myocardium), iSeg2017 [26] (MRI: Cerebrospinal Fluid, Gray Matter, White Matter), Brats2020 [19] (MRI: Whole Tumor, Tumor Core, Enhancing Tumor), ISLES2022 [9] (MRI: Ischemic Stroke), PROMISE2012 [17] (MRI: Prostate), MyoPS2020 [30] (MRI: Left Ventricle, Right Ventricle, etc.), MSD [2] Heart (MRI: Left Atrium), MSD [2] Hippocampus (MRI: Anterior Hippocampus, Posterior Hippocampus), AMOS2022 [11] (MRI: Liver, Right Kidney, etc.), ATLAS2022 [16] (MRI: Liver, Hepatic Tumor), CuRIOUS2022 [27] (Ultrasound: Tumor), and FLARE2021 [18] (CT: Liver, Kidney, Spleen, Pancreas). Datasets were split into 4:1:1 training, validation, and test sets. The training was performed on an NVIDIA GeForce RTX 4090 GPU using the SGD optimizer with a 0.01 learning rate for 500 epochs. During inference, the segmentation result is the AM branch output. Performance was evaluated using Dice Coefficient (DSC), 95% Hausdorff Distance (HD95) for discrimination, and Expected Calibration Error (ECE) and Classwise Expected Calibration Error (CECE) with $M = 15$ bins focusing on foreground regions to highlight method differences.

3.2 Results

As shown in Table 1, AffinityUMamba outperforms other methods. For instance, the largest DSC improvement is on the ATLAS2022 dataset (+0.030), compared to nnUNet's 0.746. In terms of HD95, the greatest reduction occurs on

Table 1. Segmentation performance comparison (Dice Coefficient [DSC] and 95% Hausdorff Distance [HD95]) across 12 datasets. **Bold**: best results; Underlined: second-best.

Dataset	nnUNet [10]		AttentionUNet [22]		UNet++ [29]		SegResNet [21]		SwinUNETR [8]		Ours	
	DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95
ACDC	0.913	1.318	0.844	2.162	0.897	1.704	0.874	2.168	0.876	1.961	**0.925**	**1.122**
iSeg2017	0.919	1.138	0.912	1.138	0.914	1.192	0.908	1.197	0.909	1.192	**0.922**	**1.137**
Brats2020	0.847	7.437	0.824	10.258	0.847	8.219	0.831	8.954	0.846	9.834	**0.852**	**5.197**
ISLES2022	0.779	4.331	0.787	**3.281**	0.765	4.109	0.760	4.126	0.768	5.516	**0.816**	3.999
PROMISE2012	0.872	3.645	0.854	4.531	0.862	4.462	0.840	6.201	0.844	5.876	**0.882**	**3.490**
MSD Hippocampus	0.886	1.413	0.883	1.474	0.881	1.439	0.872	1.499	0.883	1.468	**0.887**	**1.367**
ATLAS2022	0.746	20.156	0.701	30.817	0.725	27.663	0.694	30.147	0.700	28.665	**0.776**	**14.609**
CuRIOUS2022	0.782	36.928	0.742	56.612	0.767	46.608	0.764	35.916	0.763	39.452	**0.802**	**35.221**
MyoPS2020	0.702	11.276	0.629	16.448	0.630	15.636	0.641	15.591	0.654	14.291	**0.716**	**10.682**
AMOS2022	0.865	5.618	0.821	14.250	0.813	10.408	0.796	22.308	0.826	8.358	**0.879**	**5.087**
MSD Heart	0.930	3.106	0.931	3.782	0.933	3.606	0.923	4.405	0.927	3.859	**0.934**	**3.031**
FLARE	0.913	21.785	0.871	47.009	0.845	55.506	0.835	62.160	0.851	37.776	**0.927**	**7.727**

Table 2. Calibration performance comparison (Expected Calibration Error [ECE] and Classwise ECE [CECE]) across 12 datasets.

Dataset	nnUNet [10]		AttentionUNet [22]		UNet++ [29]		SegResNet [21]		SwinUNETR [8]		Ours	
	ECE	CECE	ECE	CECE	ECE	CECE	ECE	CECE	ECE	CECE	ECE	CECE
ACDC	0.0613	0.0314	0.0777	0.0837	0.1178	0.0816	0.1726	0.1222	0.0531	0.0753	**0.0493**	**0.0267**
iSeg2017	0.0412	0.0310	0.0542	0.0384	0.1487	0.0996	0.0404	0.0327	0.0367	0.0298	**0.0345**	**0.0284**
Brats2020	0.1682	0.3114	0.1604	0.3154	0.1491	0.3130	0.1610	0.3144	0.1669	0.3161	**0.1232**	**0.3105**
ISLES2022	0.2150	0.2270	0.1844	0.1984	0.1314	0.1972	0.1999	0.2838	0.1722	0.2695	**0.1240**	**0.1746**
PROMISE2012	0.1164	**0.1569**	0.1031	0.2802	0.1083	0.2794	0.1593	0.3406	0.1231	0.2880	**0.0753**	0.1653
MSD Hippocampus	0.0904	0.0663	0.0938	0.1202	0.0946	0.1299	0.1382	0.1432	**0.0552**	0.0943	0.0829	**0.0623**
ATLAS2022	0.1794	0.1693	0.1971	0.1866	0.1694	0.1728	0.1492	0.1633	0.1731	0.1796	**0.0810**	**0.1002**
CuRIOUS2022	0.1346	0.3875	0.1458	0.3139	0.1471	0.4051	0.1539	0.4470	0.1394	0.4056	**0.0788**	**0.2708**
MyoPS2020	0.1530	0.0545	0.0571	0.0543	0.1098	0.0539	**0.0499**	0.0542	0.1239	0.0549	0.1535	**0.0538**
AMOS2022	0.3509	0.0509	0.4016	0.0615	0.3643	0.0542	0.4190	0.0642	0.3077	0.0457	**0.0966**	**0.0126**
MSD Heart	0.0258	0.0398	0.0533	0.1223	0.0686	0.1276	0.0725	0.1537	0.0574	0.1155	**0.0251**	**0.0381**
FLARE2021	0.2117	0.1820	**0.0980**	**0.0486**	0.3450	0.2710	0.3564	0.2667	0.3177	0.1640	0.1844	0.3782

the FLARE dataset (−14.058), compared to nnUNet's 21.785. This is primarily due to our method's multi-module collaborative design: the AM module models global anatomical dependencies using Mamba with linear time complexity, capturing global topological constraints; the PNE aggregates neighborhood grayscale features adaptively with a multi-scale Gaussian kernel to improve boundary smoothness and connectivity; the LALR corrects noisy labels via a dual-branch confidence selection mechanism. Additionally, the IAE module dynamically weights neighborhood information based on feature similarity thresholds, alleviating boundary blurring in low-contrast regions.

AffinityUMamba also demonstrates significant calibration improvements across datasets, as shown in Table 2. The largest ECE reduction occurs on the AMOS2022 dataset (−0.2111), compared to SwinUNETR's 0.3077. For CECE, the largest decrease also occurs on AMOS2022 (−0.0331), compared to SwinUNETR's 0.0457. This advantage stems from key design elements: Margin-based Smoothing Loss dynamically perceives boundary constraints, suppressing overconfidence in noisy boundaries, while the Local Affinity Loss enforces semantic consistency between pseudo-labels and refined ground truths. The LALR successfully avoids overfitting to the toxic noisy labels, consequently reducing the prediction uncertainty originating from data uncertainty and improving the model segmentation performance.

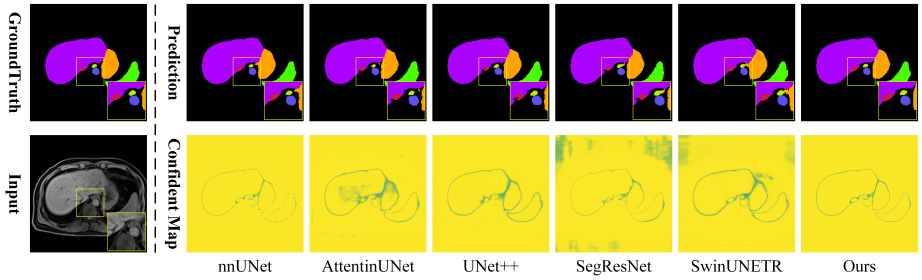

Fig. 2. Segmentation outputs and confidence maps on the AMOS2022 dataset. Confidence maps represent predictive entropy on a 0-1 scale, with yellow indicating higher confidence. Anatomical labels: liver (purple), gallbladder (orange), spleen (green), pancreas (red), kidney (blue).

Table 3. Ablation study evaluating module contributions on the ACDC dataset.

	Discrimination								Calibration	
Target	Right Ventricle		Myocardium		Left Ventricle		Avg.			
	DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95	ECE	CECE
AffinityUMamba (w/o LALR, IAE, PNE)	0.9015	1.4570	0.9043	1.1439	**0.9565**	**1.0207**	0.9214	1.2005	0.0496	0.0291
AffinityUMamba (w/o LALR)	0.9089	1.3045	0.9023	1.1398	0.9543	1.0407	0.9219	1.1617	0.0541	0.0316
AffinityUMamba$_{ours}$	**0.9143**	**1.1898**	**0.9047**	**1.1353**	0.9548	1.0414	**0.9246**	**1.1222**	**0.0493**	**0.0267**

In Fig. 2, visual predictions on the AMOS2022 dataset show that our method produced the best boundary delineation, especially for hard regions such as the gallbladder (orange) and pancreas (red), with a significant reduction in misclassified pixels at organ boundary. The significantly calibrated prediction uncertainty is demonstrated on the pixel-wise entropy confidence map, where high uncertainty is precisely concentrated on complete organ boundaries with a clean background at homogeneous regions.

Table 3 presents ablation study results. Compared to the baseline method (AffinityUMamba w/o LALR, IAE, PNE), adding IAE and PNE improves discriminative performance slightly (DSC: 0.9214 $\rightarrow$ 0.9219, HD95: 1.2005 $\rightarrow$ 1.1617), but calibration performance decreases (ECE: 0.0496 $\rightarrow$ 0.0541, CECE: 0.0291 $\rightarrow$ 0.0316), due to the absence of the LALR mechanism. Despite more ambiguous boundaries in the right ventricle and myocardium, segmentation performance still improves (DSC: Right Ventricle 0.9089 $\rightarrow$ 0.9143, Myocardium 0.9023 $\rightarrow$ 0.9047, HD95: Right Ventricle 1.3045 $\rightarrow$ 1.1898, Myocardium 1.4398 $\rightarrow$ 1.1353). In contrast, the left ventricle, with clearer boundaries, shows limited improvement (DSC: 0.9543 $\rightarrow$ 0.9548, HD95: 1.1617 $\rightarrow$ 1.1222).

4 Conclusion

This study presents AffinityUMamba, a weakly supervised framework that addresses annotation uncertainty in medical image segmentation through weak supervision. By integrating convolutional networks with state space models, our

approach effectively captures anatomical dependencies while mitigating boundary ambiguities. The synergistic architecture combines global context modeling with adaptive local refinement, demonstrating exceptional performance across 12 multi-modal datasets. Experimental results show significant improvements in both discriminative and calibration performance compared to existing state-of-the-art methods. The framework resolves annotation-caused data uncertainty, proving particularly valuable for MRI soft tissue analysis while maintaining promising generalizability across CT and ultrasound modalities. This work establishes a new insight and a reliability-aware paradigm for medical image segmentation using public datasets, bridging the gap between computational models and practical clinical decision-making needs.

Acknowledgments. This study was supported by the Special Scientific Research Program for Capital Health Development (grant number 2024-1-5041).

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

References

1. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4981–4990 (2018)
2. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)
3. 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)
4. 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)
5. Castellano, G., Bonilha, L., Li, L., Cendes, F.: Texture analysis of medical images. Clinical radiology **59**(12), 1061–1069 (2004)
6. Despotović, I., Goossens, B., Philips, W.: Mri segmentation of the human brain: challenges, methods, and applications. Computational and mathematical methods in medicine **2015**(1), 450341 (2015)
7. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
8. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
9. Hernandez Petzsche, M.R., de la Rosa, E., Hanning, U., Wiest, R., Valenzuela, W., Reyes, M., Meyer, M., et al.: Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. Scientific data **9**(1), 762 (2022)

10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
11. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in neural information processing systems **35**, 36722–36732 (2022)
12. Jungo, A., Meier, R., Ermis, E., Blatti-Moreno, M., Herrmann, E., et al.: On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I. pp. 682–690. Springer (2018)
13. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems **30** (2017)
14. Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., Van Stiphout, R.G., Granton, P., Zegers, C.M., Gillies, R., Boellard, R., Dekker, A., et al.: Radiomics: extracting more information from medical images using advanced feature analysis. European journal of cancer **48**(4), 441–446 (2012)
15. Li, H., Qin, P., Yuan, X., Chen, Z., et al.: Hcma-unet: A hybrid cnn-mamba unet with inter-slice self-attention for efficient breast cancer segmentation. arXiv preprint arXiv:2501.00751 (2025)
16. Liew, S.L., Lo, B.P., Donnelly, M.R., Zavaliangos-Petropulu, A., et al.: A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Scientific data **9**(1), 320 (2022)
17. Litjens, G., Toth, R., Van De Ven, W., Hoeks, C., Kerkstra, S., Van Ginneken, B., Vincent, G., Guillard, G., et al.: Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. Medical image analysis **18**(2), 359–373 (2014)
18. Ma, J., Zhang, Y., Gu, S., An, X., Wang, Z., Ge, C., Wang, C., Zhang, F., Wang, Y., Xu, Y., et al.: Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. Medical Image Analysis **82**, 102616 (2022)
19. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
20. Murugesan, B., Liu, B., Galdran, A., et al.: Calibrating segmentation networks with margin-based label smoothing. Medical Image Analysis **87**, 102826 (2023)
21. Myronenko, A.: 3d mri brain tumor segmentation using autoencoder regularization. In: International MICCAI brainlesion workshop. pp. 311–320. Springer (2018)
22. Oktay, O., Schlemper, J., Folgoc, L.L., et al.: Attention u-net: Learning where to look for the pancreas. ArXiv **abs/1804.03999** (2018)
23. Ru, L., Zhan, Y., Yu, B., Du, B.: Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16846–16855 (June 2022)
24. Sherer, M.V., Lin, D., Elguindi, S., Duke, S., Tan, L.T., Cacicedo, J., et al.: Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. Radiotherapy and Oncology **160**, 185–191 (2021)
25. Tian, Z., Liu, L., Zhang, Z., Fei, B.: Psnet: prostate segmentation on mri based on a convolutional neural network. Journal of medical imaging **5**(2), 021208–021208 (2018)
26. Wang, L., Nie, D., Li, G., Puybareau, É., Dolz, J., Zhang, Q., Wang, F., Xia, J., Wu, Z., Chen, J.W., et al.: Benchmark on automatic six-month-old infant brain

segmentation algorithms: the iseg-2017 challenge. IEEE transactions on medical imaging **38**(9), 2219–2230 (2019)
27. Xiao, Y., Fortin, M., Unsgård, G., Rivaz, H., Reinertsen, I.: Resect: a clinical database of pre-operative mri and intra-operative ultrasound in low-grade glioma surgeries. Med. Phys **44**(7), 3875–3882 (2017)
28. Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Sohel, F., Xu, D.: Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6984–6993 (October 2021)
29. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings 4. pp. 3–11. Springer (2018)
30. Zhuang, X.: Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE transactions on pattern analysis and machine intelligence **41**(12), 2933–2946 (2018)