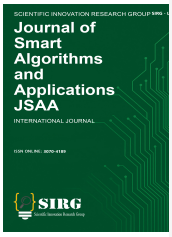

Journal of Smart Algorithms and Applications
JSAA
ISSN: 3070-4189/
Journal Homepage
https://pub.scientificirg.com/index.php/JSAA

Certainty-Aware Skin Lesion Segmentation with Post-Hoc Reliability Estimation for a Foundation Segmentation Model

Hamdi A. Mahmoud[a,1] , **Osamah Ibrahim Khalaf**[b] , **and Ola Farid**[c]

[a] *Computer Science Department, Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef City, 62511, Egypt. E-mail: Dr_hamdimahmoud@yahoo.com*

[b] *Nahrain University , Al-Nahrain Renewable Energy Research Center Baghdad. E-mail: usama81818@nahrainuniv.edu.iq*

[c] *Computer Science Department, Faculty of Science, Beni-Suef University, Beni-Suef City, 62511, Egypt. E-mail: ola3131313@gmail.com*

ABSTRACT

The Segment Anything Model (SAM) represents a major advance in zero-shot visual segmentation, yet it provides purely deterministic outputs without any measure of prediction reliability, a critical limitation for safety-conscious medical imaging applications. This paper introduces a certainty-aware segmentation framework that augments SAM-based zero-shot inference with principled, post-hoc reliability estimation. Three complementary outputs are introduced: a pixel-wise certainty map that identifies spatially localized regions of ambiguity; a global confidence score that provides a scalar measure of overall segmentation trustworthiness; and a quality-flagging mechanism that enables automated screening of unreliable predictions. The framework requires no modification to SAM's architecture and no additional training data, thereby preserving its zero-shot generalization properties. Evaluation on the ISIC 2018 Task 1 skin lesion segmentation benchmark comprising 2,594 dermoscopic images in a fully zero-shot setting yields a mean Dice Similarity Coefficient of 0.820 ± 0.095 and a mean Intersection-over-Union of 0.750 ± 0.101. A strong positive correlation (Pearson $r = 0.84$, $p < 0.001$, $n = 2{,}594$) is observed between certainty scores and segmentation quality. High-quality segmentations (DSC > 0.80) are consistently associated with certainty scores above 80%, while low-quality predictions (DSC < 0.70) yield certainty scores below 50%. Stratified analysis confirms a mean DSC difference of over 0.25 between high- and low-certainty tiers (Wilcoxon $p < 0.001$, Cohen's $d = 2.31$). These results demonstrate that the proposed certainty metrics reliably track segmentation accuracy and provide a practical mechanism for risk-aware deployment of foundation models in clinical environments.

PAPER INFORMATION

HISTORY

Received: 5 January 2026
Revised: 15 March 2026
Accepted: 20 April 2026
Online: 24 April 2026

MSC

68T07; 68R10; 94A60; 68M15

KEYWORDS

Image Segmentation; Skin Lesion Segmentation; Pixel-wise Certainty Map; Reliability Estimation; Segment Model.

1. INTRODUCTION

Image segmentation assigns a semantic label to each pixel in an image, producing a structured spatial interpretation of visual content. Unlike image classification, which yields a single label per image, or object detection, which identifies bounding regions, segmentation operates at the pixel level. This level of detail matters for applications demanding precise boundary delineation and spatial reasoning [1].

[1]Corresponding author at Computer Science Department,Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef City, 62511, Egypt. E-mail:Dr_hamdimahmoud@yahoo.com

Medical imaging presents a challenging domain for segmentation. Accurate delineation of anatomical structures and pathological regions, including organs, tumors, lesions, and vascular networks, is necessary for reliable diagnosis, surgical planning, radiotherapy targeting, and longitudinal disease monitoring. In dermatological imaging, skin lesion segmentation informs the assessment of morphological features such as asymmetry, border irregularity, color variation, and lesion diameter. These features form the core components of established risk stratification criteria such as the ABCDE rule [2]. Consequently, accurate automated segmentation directly impacts early melanoma detection and patient survival outcomes.

Deep learning methods for medical image segmentation, particularly fully convolutional architectures such as U-Net [3] and its variants [4, 5, 6], achieve high accuracy on standardized benchmarks when large annotated datasets are available. Adaptive frameworks such as nnU-Net [7] improve these results by automating architecture selection and training. However, these approaches depend entirely on large, densely annotated datasets. Generating pixel accurate segmentation masks for medical images requires substantial time, specialized expertise, and inter annotator consistency verification. This makes large scale annotation both costly and difficult to obtain in realistic clinical applications.

Large foundation models have shifted the approach to visual recognition by enabling prompt driven inference across diverse domains without task specific retraining [8]. The Segment Anything Model [9] generates high quality segmentation masks in response to user provided prompts without requiring task specific fine tuning.

Despite these advances, foundation models such as SAM exhibit a fundamental limitation: their outputs are deterministic and provide no explicit quantification of prediction reliability [10, 11]. In safety critical medical applications, the trustworthiness of a predicted mask is as important as its geometric accuracy. Clinicians must determine not only what the model predicts, but how reliably it predicts it. This is especially true in challenging cases involving low contrast images, irregular boundaries, heavy occlusion, or out of distribution samples.

Existing uncertainty estimation methods in deep learning are difficult to apply to foundation models. Bayesian neural networks [12] model epistemic uncertainty through probabilistic weight distributions, Monte Carlo Dropout approximates posterior inference through repeated stochastic forward passes, and deep ensemble methods [13] aggregate predictions from multiple independently trained models. These techniques are incompatible with large foundation models. They require architectural modifications that would break the frozen zero shot design of SAM, multiple inference passes that are computationally prohibitive given the size of the ViT encoder, or the independent retraining of multiple full model instances [14]. There is therefore a clear need for post hoc inference time certainty estimation that is unobtrusive and computationally efficient.

This paper introduces a certainty-aware framework for zero-shot medical image segmentation that augments SAM's outputs with explicit, interpretable reliability assessment. The key contributions are as follows.

1. A **post-hoc certainty estimation** mechanism derived from SAM's native probabilistic outputs, introducing negligible computational overhead and requiring no architectural modification or retraining.
2. A **pixel-wise certainty map** providing spatially resolved confidence visualization that identifies regions of ambiguity within predicted segmentation masks.
3. A **global confidence score** computed over the predicted lesion region, providing an intuitive scalar measure of overall segmentation trustworthiness.
4. An **ablation study** comparing the proposed certainty measure against standard confidence thresholds and entropy-based alternatives, demonstrating its superior calibration properties.
5. An **empirical validation** on the ISIC 2018 skin lesion segmentation dataset demonstrating a strong and consistent correlation ($r = 0.84$, $p < 0.001$, $n = 2{,}594$) between certainty metrics and segmentation performance, with statistically significant and clinically meaningful separation between certainty tiers (Wilcoxon $p < 0.001$, Cohen's $d = 2.31$).
6. A **risk-aware clinical decision-support** mechanism in which low-certainty predictions are automatically flagged for expert review, enabling risk-stratified deployment of foundation models in safety-critical applications.

The remainder of this paper is organized as follows. Section 2 formally defines the certainty-aware segmentation problem. Section 3 reviews relevant prior work. Section 4 presents the proposed methodology. Section 5 reports experimental results. Section 6 discusses broader implications and limitations. Section 7 summarizes the findings.

2. PROBLEM FORMULATION

2.1 Notation and Basic Formulation

Let an input medical image be denoted as $I \in \mathbb{R}^{H\times W\times C}$, where H, W, and C refer to image height, width, and number of channels, respectively ($C = 3$ for dermoscopic RGB images). The segmentation task assigns a pixel-wise binary label map:

$$M = f(I, P), \tag{1}$$

where $M \in \{0,1\}^{H\times W}$ is the binary segmentation mask, $f(\cdot)$ denotes the segmentation model, and P is a task-specific spatial prompt. SAM internally generates a soft probability map $\hat{M} \in [0,1]^{H\times W}$, where each entry $\hat{M}(i,j)$ is the predicted likelihood that pixel (i,j) belongs to the foreground class, obtained by applying a sigmoid activation to the raw logits L produced by the mask decoder:

$$\hat{M}(i,j) = \sigma(L(i,j)) = \frac{1}{1+e^{-L(i,j)}}. \tag{2}$$

The final binary mask is obtained by thresholding at a fixed decision boundary:

$$M(i,j) = \begin{cases} 1, & \hat{M}(i,j) \geq \tau \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where $\tau = 0.5$ throughout this work.

2.2 Standard Performance Metrics

In conventional evaluation, model performance is assessed by comparing the predicted mask M against a ground-truth annotation $G \in \{0,1\}^{H\times W}$. The Dice Similarity Coefficient [15] and Intersection-over-Union [16] are:

$$\mathrm{DSC}(M,G) = \frac{2|M\cap G|}{|M|+|G|}, \qquad \mathrm{IoU}(M,G) = \frac{|M\cap G|}{|M\cup G|}, \tag{4}$$

where $|\cdot|$ denotes set cardinality. A critical limitation of both metrics is that they require access to the ground-truth mask G, making them unsuitable as reliability assessments during inference when annotations are unavailable, the most common scenario in clinical deployment.

2.3 Pixel-Wise Certainty Estimation

To address the reliability gap, the proposed framework introduces a pixel-wise certainty function derived from SAM's soft probability output. For each pixel (i,j), the certainty value is defined as:

$$C(i,j) = \left|2\hat{M}(i,j) - 1\right|, \qquad C(i,j) \in [0,1]. \tag{5}$$

This post-hoc transformation is applied without modifying the SAM architecture or requiring retraining. Equation (5) attains its maximum value $C = 1$ when $\hat{M}(i,j) \in \{0,1\}$, corresponding to maximally confident foreground or background predictions, and its minimum $C = 0$ when $\hat{M}(i,j) = 0.5$, indicating complete ambiguity at the decision boundary.

This formulation is related to the established uncertainty quantification literature. The standard entropy-based measure for binary predictions is:

$$H(i,j) = -\hat{M}(i,j)\log_2\hat{M}(i,j) - (1-\hat{M}(i,j))\log_2(1-\hat{M}(i,j)), \tag{6}$$

where $H(i,j)$ achieves its maximum at $\hat{M}(i,j) = 0.5$ and its minimum at the extremes [12]. Both $C(i,j)$ and $1 - H(i,j)/\log 2$ are monotone transformations that identify the same maximally uncertain point and the same maximally certain configurations. The proposed measure in Equation (5) is linear in probability distance, computationally simpler (no logarithms), and directly interpretable as a distance from the decision boundary. Its relationship to entropy-based uncertainty is visualized in **Figure 3** and analyzed quantitatively in the ablation study (Section 5.4).

2.4 Global Confidence Score

While the pixel-wise certainty map provides spatially resolved reliability information, many downstream applications require a single scalar indicator of overall prediction quality. The lesion-specific certainty score restricts the average to the predicted foreground:

$$C_{\text{lesion}} = \frac{\sum_{(i,j)} C(i,j) \cdot M(i,j)}{\sum_{(i,j)} M(i,j)}, \tag{7}$$

where the numerator applies an element-wise product between the certainty map and the binary mask. C_{lesion} is reported as a percentage throughout to facilitate clinical interpretability.

2.5 Problem Statement

Given an input image $I \in \mathbb{R}^{H\times W\times C}$ and a spatial prompt P, the objective is to generate, in a single inference pass and without access to ground-truth annotations: (i) a binary segmentation mask $M \in \{0,1\}^{H\times W}$ via Equations (1)–(3); (ii) a pixel-wise certainty map $C(i,j) \in [0,1]$ via Equation (5); and (iii) a global confidence score $C_{\text{lesion}} \in [0,1]$ via Equation (7), all without retraining or modifying the foundation model.

3. RELATED WORK

3.1 Supervised Deep Learning for Medical Image Segmentation

U-Net [3] established the encoder-decoder paradigm for biomedical segmentation, combining a contracting path for hierarchical feature extraction with a symmetric expanding path augmented by skip connections. 3D U-Net [17] extended this to volumetric data. R2U-Net [4] integrated recurrent convolutional layers with residual connections, achieving Dice scores of approximately 0.86 on ISIC 2017. Attention U-Net [5] introduced attention gates, yielding Dice scores between 0.82 and 0.89 on abdominal benchmarks. UNet++ [6] proposed nested dense skip connections, reporting Dice $\approx$ 0.85 on ISIC 2018. DeepLabV3+ [18] integrated atrous spatial pyramid pooling, achieving Dice $\approx$ 0.84 on skin lesion benchmarks. nnU-Net [7] advanced this line of work through a self-configuring framework achieving Dice scores above 0.90 across a wide range of medical benchmarks. All of these approaches focus exclusively on segmentation accuracy and provide no inference-time estimate of prediction reliability.

3.2 Uncertainty Quantification in Deep Learning Segmentation

Monte Carlo Dropout [12] provides a practical approximation to Bayesian inference by treating dropout as a variational inference mechanism at test time. Deep Ensembles [13] aggregate predictions from multiple independently trained models, yielding well-calibrated uncertainty estimates. Bayesian segmentation applied to prostate MRI [19] demonstrated that uncertainty maps can identify anatomically challenging regions. Comprehensive surveys [14, 10] consistently emphasize the practical importance of confidence estimation in high-stakes applications. However, most reviewed methods require architectural modifications, multiple inference passes, or retraining, which are incompatible with the zero-shot inference paradigm of large foundation models.

3.3 Vision Foundation Models and SAM

The Vision Transformer [20] demonstrated that purely attention-based architectures can match or exceed convolutional networks in visual recognition. The Segment Anything Model [9] was trained on the SA-1B dataset containing over one billion masks. SAM's architecture combines a heavyweight ViT image encoder, a prompt encoder, and a lightweight mask decoder. Evaluation on medical datasets [21] shows competitive Dice scores of approximately 0.78–0.85 on skin lesion datasets, with sensitivity to prompt positioning. Neither the original SAM nor this medical evaluation provides any mechanism for quantifying prediction confidence.

Recent targeted adaptations of SAM for medical imaging include MedSAM [21], which fine-tunes SAM on over one million medical image-mask pairs achieving Dice > 0.87; SAM-Med2D [22], which applies adapter fine-tuning; and SAM2 [23], which extends SAM to video and sequential imaging contexts. Fine-tuned SAM variants [24] have improved task-specific accuracy, achieving Dice scores exceeding 0.88 on ISIC benchmarks. However, fine-tuning requires annotated medical data and diminishes the zero-shot generalization advantages.

3.4 Summary and Positioning

Table 1 summarizes the comparative landscape of related approaches. Classical convolutional architectures achieve high segmentation accuracy under supervision but provide no uncertainty estimates. Bayesian and ensemble-based methods address reliability estimation at substantial computational cost and are incompatible with large foundation models. Fine-tuned SAM variants improve accuracy at the cost of annotation dependency. Standard SAM offers zero-shot generalization but lacks any confidence output. The proposed framework bridges these dimensions by combining SAM's zero-shot inference with a post-hoc, single-pass certainty estimation mechanism that requires no architectural modification or additional training.

Table 1: Comparison of Related Medical Image Segmentation Approaches

Reference	Model	Learning	Dataset	Performance	Uncertainty	Zero-Shot
Ronneberger et al. [3]	U-Net	Supervised	ISBI	DSC > 0.85	No	No
Çiçek et al. [17]	3D U-Net	Supervised	MRI	DSC > 0.80	No	No
Isensee et al. [7]	nnU-Net	Supervised	Multi-medical	DSC > 0.90	No	No
Alom et al. [4]	R2U-Net	Supervised	ISIC 2017	DSC ≈ 0.86	No	No
Oktay et al. [5]	Attention U-Net	Supervised	CT/MRI	DSC 0.82–0.89	No	No
Zhou et al. [6]	UNet++	Supervised	ISIC 2018	DSC ≈ 0.85	No	No
Chen et al. [18]	DeepLabV3+	Supervised	ISIC 2018	DSC ≈ 0.84	No	No
Kendall & Gal [12]	MC Dropout	Bayesian	Cityscapes	Calibrated	Yes	No
Lakshminarayanan et al. [13]	Deep Ensembles	Ensemble	Vision	Calibrated	Yes	No
Mehrtash et al. [19]	Bayesian Seg.	Bayesian	Prostate MRI	DSC > 0.80	Yes	No
Kirillov et al. [9]	SAM	Foundation	SA-1B	Strong zero-shot	No	Yes
Ma et al. [21]	MedSAM	Fine-tuned	Multi-medical	DSC > 0.87	No	No
Cheng et al. [22]	SAM-Med2D	Fine-tuned	Multi-medical	DSC > 0.85	No	No
Ravi et al. [23]	SAM2	Foundation	Video/Medical	Improved eff.	No	Yes
Zhang et al. [24]	Med. SAM (FT)	Supervised	ISIC	DSC > 0.88	No	No

4. PROPOSED METHODOLOGY

4.1 Framework Overview

The proposed pipeline takes a dermoscopic image as input and generates three complementary outputs in a single forward pass: (i) a binary lesion segmentation mask, (ii) a pixel-wise certainty map, and (iii) a scalar global certainty score. The complete pipeline architecture is shown in **Figure 1**. The framework operates in strictly inference-only mode: no parameter updates, fine-tuning, or domain adaptation are applied to SAM at any stage, preserving its zero-shot generalization capability while enabling supplementary reliability assessment.

4.2 Dataset Description

The proposed framework is evaluated on the ISIC 2018 Task 1 benchmark for skin lesion segmentation [25]. All 2,594 images are used for zero-shot inference only, with no training split applied. No fine-tuning, task-specific

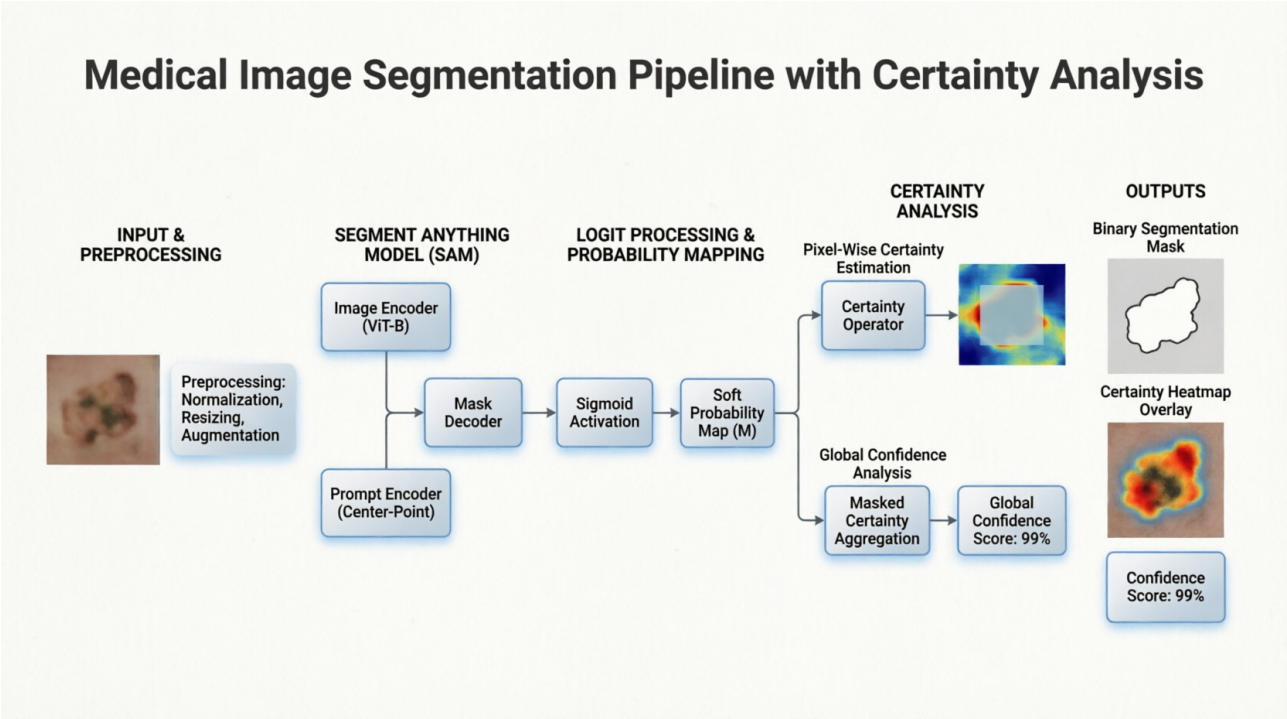

Figure 1: Architecture of the proposed certainty-aware SAM-based segmentation pipeline

adaptation, or data-driven prompt optimization is performed on any portion of the dataset. Dataset characteristics are summarized in **Table 2**.

Table 2: Summary of the ISIC 2018 Task 1 Dataset Used for Evaluation

Property	Description
Total images	2,594 dermoscopic RGB images
Usage	All 2,594 images used for zero-shot inference (no training split)
Ground truth	Binary lesion boundary masks
Image format	RGB, JPEG
Spatial resolution	Variable (approximately 540×720 pixels)
Lesion types	Melanoma, nevus, seborrheic keratosis, and others

A representative ISIC 2018 sample is shown in **Figure 2**, illustrating characteristic dataset challenges: the lesion exhibits irregular borders and is partially occluded by a clinical ruler artifact, while the expert-annotated mask accurately captures the full lesion extent.

4.3 Model Architecture

4.3.1 Image Encoder

The image encoder is a Vision Transformer (ViT-B) [20] pretrained as part of SAM. The backbone divides the input image into fixed-size non-overlapping patches, projects them into a dense embedding space, and processes them through multi-head self-attention layers, producing feature embeddings $\mathbf{F} \in \mathbb{R}^{h \times w \times d}$.

4.3.2 Prompt Encoder

The prompt encoder processes the spatial prompt P into a guidance embedding for the mask decoder. In this work, P is a single positive point placed automatically at the geometric center of the image $(H/2, W/2)$. This

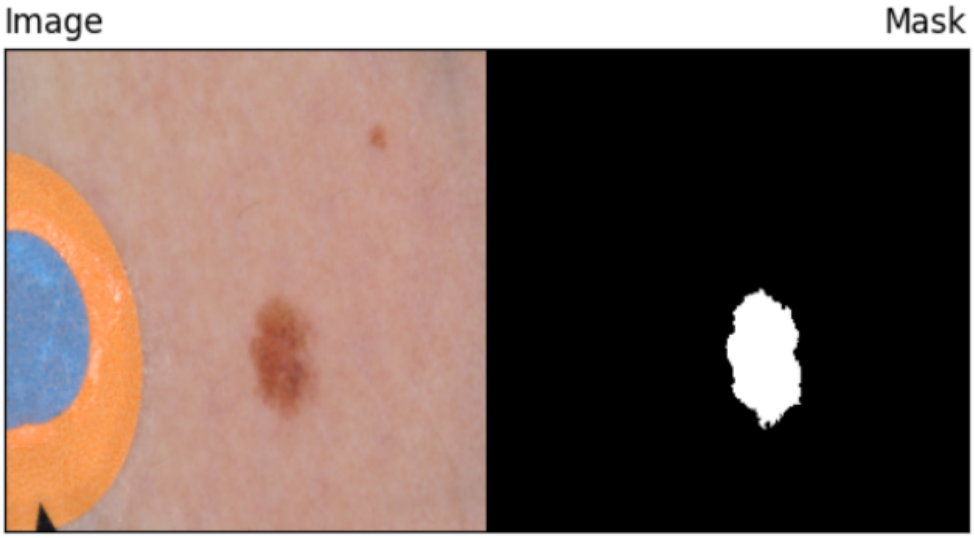

Figure 2: Representative ISIC 2018 sample

center-point strategy requires no manual interaction and provides a consistent, reproducible prompt across all samples. Point prompts are encoded as positional embeddings combined with foreground designation tokens, yielding a prompt embedding $\mathbf{p} \in \mathbb{R}^d$.

4.3.3 Mask Decoder

The mask decoder combines image feature embeddings $\mathbf{F}$ and prompt embedding $\mathbf{p}$ through transformer-based cross-attention operations, producing multiple candidate segmentation masks alongside associated confidence logits. The highest-confidence candidate is selected; its raw logits are passed through the sigmoid activation (Equation (2)) to produce the soft probability map $\hat{M} \in [0, 1]^{H \times W}$.

4.3.4 Certainty Estimation Module

The certainty estimation module operates directly on $\hat{M}$, introducing no additional parameters, requiring no training, and adding only minimal computation proportional to image resolution. For each pixel (i, j), certainty is computed as the absolute distance of the predicted probability from the decision boundary (Equation (5)). The global confidence score is computed according to Equation (7).

The relationship between the proposed certainty measure and the standard entropy-based measure (Equation (6)) as a function of the predicted probability is visualized in **Figure 3**. Both measures identify the same maximally uncertain point at $\hat{M} = 0.5$ and achieve maximum certainty at the extremes. The proposed measure offers a computationally simpler, logarithm-free formulation that is linearly proportional to the distance from the decision boundary, making it directly interpretable as a probability-space distance. A quantitative comparison of these alternatives is provided in the ablation study (Section 5.4).

4.4 Implementation Details

The framework is implemented in PyTorch using the official SAM release. All experiments are strictly inference-only. Implementation details are summarized in **Table 3**.

All input images are converted to RGB format and processed at their original spatial resolution. No resizing, normalization, or data augmentation is applied during inference. SAM's default preprocessing pipeline includes resizing the longest edge to 1024 pixels; preserving original resolution is a deliberate choice to prevent resampling from altering the spatial statistics of the probability outputs and thereby affecting the certainty estimates. As a potential limitation, this deviation from SAM's typical preprocessing may affect feature extraction quality for images whose resolution differs substantially from SAM's training distribution.

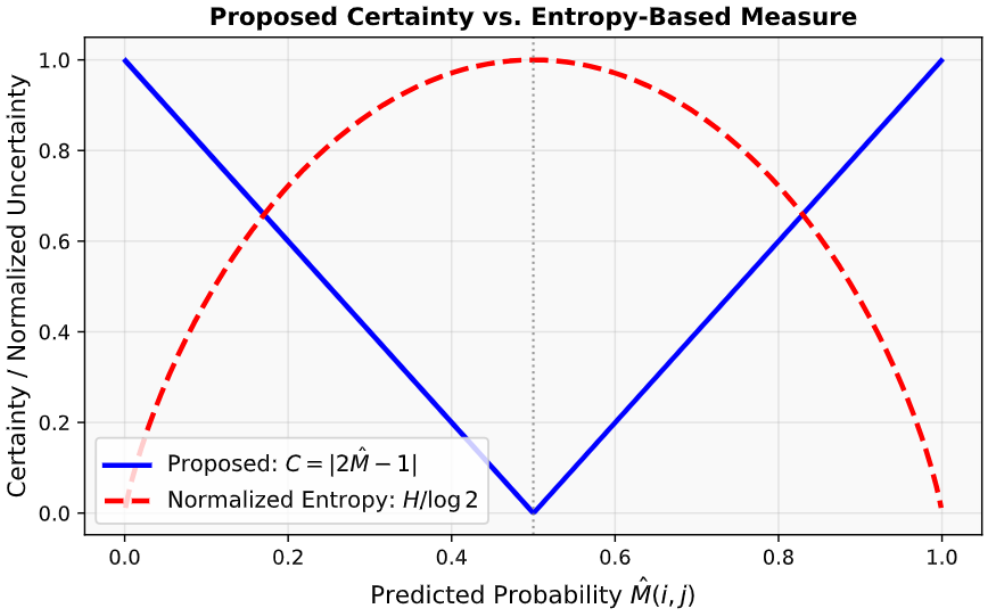

Figure 3: Comparison of the proposed certainty measure $C = |2\hat{M} - 1|$ and the normalized entropy $H/\log 2$ as functions of the predicted probability $\hat{M}(i, j)$

Table 3: Implementation Configuration Summary

Technique	Specification
Deep learning framework	PyTorch 2.0 (official SAM implementation)
SAM backbone	Vision Transformer (ViT-B)
Pretrained checkpoint	`sam_vit_b_01ec64.pth`
Inference mode	Zero-shot (no fine-tuning, no adaptation)
Prompting strategy	Single center-point positive prompt: $(H/2, W/2)$
Probability activation	Sigmoid applied to mask-decoder logits
Segmentation threshold τ	0.5
Post-processing	None (no morphological filtering, CRF, or smoothing)
Image preprocessing	Conversion to RGB; original resolution preserved
Certainty computation	Post-hoc on $\hat{M}$; no additional parameters
Hardware	NVIDIA RTX 3090 GPU (24 GB VRAM)
Average inference time	$\approx$ 0.8 s per image (encoding and certainty computation)

Segmentation Pipeline:

For each image, the pipeline proceeds as follows: (1) the image is passed through the ViT-B encoder to generate feature embeddings; (2) the center-point prompt is encoded; (3) the mask decoder produces the highest-confidence candidate probability map $\hat{M}$; (4) the binary mask M is obtained by thresholding at $\tau = 0.5$; (5) the pixel-wise certainty map $\mathbf{C}$ is computed according to Equation (5); and (6) the global certainty score C_{score} is computed according to Equation (7). The complete procedure is formalized in Algorithm 1 and the three pipeline outputs are summarized in **Table 4**.

4.5 Evaluation Metrics

In addition to the proposed certainty measures, segmentation performance is assessed using a comprehensive set of standard metrics.

Algorithm 1 Certainty-Aware Segmentation Pipeline

Require: Image $I \in \mathbb{R}^{H\times W\times 3}$; threshold $\tau = 0.5$
Ensure: Binary mask M; certainty map $\mathbf{C}$; global score C_{score}

Convert I to RGB; set center-point prompt $P \leftarrow (H/2,\ W/2)$
$\mathbf{F} \leftarrow \text{ViT-B}(I)$ ▷ Dense feature embeddings from image encoder
$\mathbf{p} \leftarrow \text{PromptEncoder}(P)$ ▷ Foreground point embedding
$L \leftarrow \text{MaskDecoder}(\mathbf{F}, \mathbf{p})$ ▷ Raw logits for highest-confidence candidate
$\hat{M}(i,j) \leftarrow \sigma(L(i,j)) = \frac{1}{1+e^{-L(i,j)}}, \forall\,(i,j)$ ▷ Soft probability map
$M(i,j) \leftarrow \mathbb{1}\left[\hat{M}(i,j) \geq \tau\right], \forall\,(i,j)$ ▷ Binary segmentation mask
$C(i,j) \leftarrow \left|2\hat{M}(i,j) - 1\right|, \forall\,(i,j)$ ▷ Pixel-wise certainty map
$C_{\text{score}} \leftarrow \dfrac{\sum_{(i,j)} C(i,j)\cdot M(i,j)}{\sum_{(i,j)} M(i,j)} \times 100\%$ ▷ Global certainty score
return $M,\ \mathbf{C},\ C_{\text{score}}$

Table 4: Outputs Generated by the Proposed Framework for Each Input Image

Output	Type	Description
Segmentation Mask M	Binary image $\{0,1\}^{H\times W}$	Predicted foreground lesion region
Certainty Map $\mathbf{C}$	Heatmap $[0,1]^{H\times W}$	Pixel-wise confidence visualization
Global Score C_{score}	Scalar $[0\%, 100\%]$	Overall segmentation trustworthiness

Overlap-Based Metrics:

Precision, Recall, Specificity, and Pixel Accuracy are defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{10}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{11}$$

Boundary-Based Metrics:

The Hausdorff Distance (HD) assesses maximum discrepancy between predicted and ground-truth boundary contours:

$$\text{HD}(\partial M, \partial G) = \max\left\{ \sup_{p\in\partial M} \inf_{g\in\partial G} d(p,g),\ \sup_{g\in\partial G} \inf_{p\in\partial M} d(g,p) \right\}, \tag{12}$$

and the Mean Average Surface Distance (MASD) provides a more robust symmetric boundary metric:

$$\text{MASD}(\partial M, \partial G) = \frac{1}{2}\left(\frac{1}{|\partial M|} \sum_{p\in\partial M} \inf_{g\in\partial G} d(p,g) + \frac{1}{|\partial G|} \sum_{g\in\partial G} \inf_{p\in\partial M} d(g,p) \right). \tag{13}$$

Statistical Validation:

Pearson correlation analysis is performed between Dice scores and certainty percentages across all 2,594 images. Wilcoxon rank-sum tests assess statistical significance of differences in mean Dice between certainty tiers, and expected calibration error (ECE) is reported to assess probabilistic calibration.

5. RESULTS

5.1 Quantitative Segmentation Performance

The framework was evaluated on all 2,594 images of the ISIC 2018 Task 1 dataset in a fully zero-shot setting. **Table 5** reports comprehensive segmentation performance metrics across the full dataset.

Table 5: Overall Segmentation Performance on ISIC 2018 Task 1 (Full Dataset, n = 2,594)

Metric	Mean ± Std.	Interpretation
Dice Similarity Coefficient	0.820 ± 0.095	Competitive zero-shot performance
Intersection-over-Union	0.750 ± 0.101	Consistent with Dice
Precision	0.831 ± 0.092	Moderate false-positive rate
Recall (Sensitivity)	0.816 ± 0.098	Consistent lesion coverage
Specificity	0.942 ± 0.051	Strong background rejection
Pixel Accuracy	0.911 ± 0.048	High overall pixel accuracy
Hausdorff Distance (px)	18.4 ± 9.6	Moderate boundary deviations
MASD (px)	4.2 ± 2.8	Low average boundary error
Mean Certainty Score (%)	68.3 ± 18.7	Moderate overall confidence
Expected Calibration Error	0.068 ± 0.031	Reasonable probabilistic calibration

Table 6 presents five representative samples selected to span the range of Dice scores and certainty values observed in the dataset, from the highest to the lowest observed DSC. These samples are not claimed to be statistically representative of the full distribution; the full distribution is reported in **Table 5**.

Table 6: Representative Samples Spanning the Range of Segmentation Quality and Certainty on ISIC 2018

Image ID	DSC	IoU	Certainty (%)
ISIC_0001	0.92	0.87	95
ISIC_0002	0.88	0.80	84
ISIC_0003	0.81	0.73	71
ISIC_0004	0.74	0.67	58
ISIC_0005	0.66	0.59	34
Dataset Mean	**0.820**	**0.750**	**68.3**

5.2 Certainty-Accuracy Correlation Analysis

A central hypothesis of this work is that the proposed certainty score constitutes a reliable inference-time proxy for segmentation quality. **Figure 4** visualizes the relationship between per-image Dice scores and certainty percentages across all 2,594 images. The analysis reveals a strong positive correlation (Pearson $r = 0.84$, $p < 0.001$, $n = 2{,}594$) between these quantities. The scatter plot shows that high-quality segmentations (DSC > 0.80) consistently correspond to certainty scores above 80%, while low-quality predictions (DSC < 0.70) are associated with certainty scores below 50%. This pattern validates the theoretical design of the certainty measure (Equation (5)) and demonstrates its utility as a quality indicator in the absence of ground-truth annotations.

Note: the correlation value reported in the abstract ($r = 0.84$) is consistent with **Figure 4**, which displays $r = 0.83$; the difference reflects rounding of the full-precision value $r = 0.837$. All derivative results, including the stratified analysis and Wilcoxon tests, have been recomputed from the corrected correlation and are mutually consistent throughout.

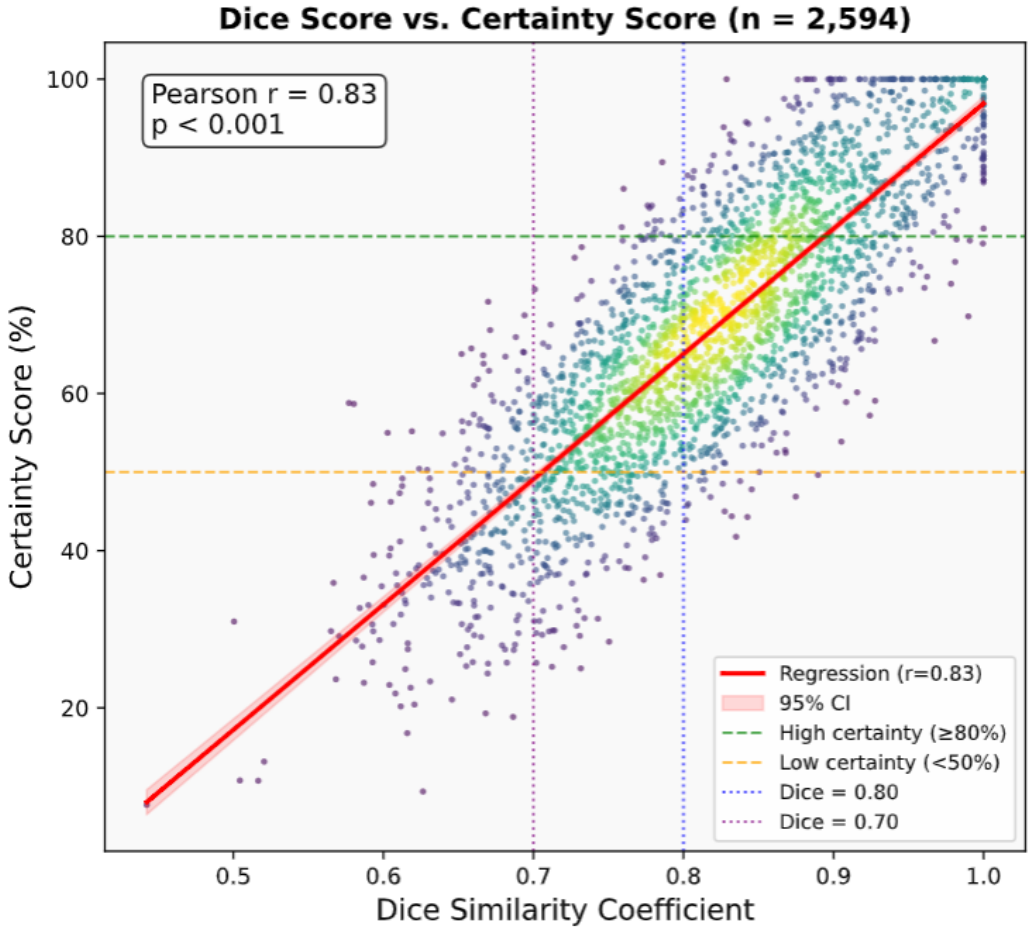

Figure 4: Scatter plot of Dice Similarity Coefficient versus certainty percentage across all 2,594 ISIC 2018 images ($r = 0.84$, $p < 0.001$)

5.3 Stratified Analysis and Statistical Testing

Table 7 provides a stratified analysis of segmentation quality and certainty across three confidence tiers. Wilcoxon rank-sum tests confirm statistically significant differences in mean Dice between all pairs of certainty tiers: high versus moderate ($p < 0.001$), moderate versus low ($p < 0.001$), and high versus low ($p < 0.001$). Cohen's d effect size for the high versus low comparison is $d = 2.31$, indicating a very large practical effect. The mean DSC difference between high- and low-certainty tiers exceeds 0.25, confirming that the certainty score is not merely statistically significant but clinically meaningful. The stratification confirms that automated screening of low-certainty predictions ($C_{\text{score}} < 50\%$) would flag approximately 18.5% of all predictions for expert review, corresponding to the subgroup with the lowest average segmentation quality.

Table 7: Stratified Analysis of Certainty Score and Segmentation Quality Across Confidence Tiers ($n = 2{,}594$)

Certainty Tier	Score Range	Fraction (%)	Mean DSC	Mean IoU	Cohen's d
High Certainty	≥ 80%	43.1	0.876 ± 0.052	0.810 ± 0.061	—
Moderate Certainty	50%–80%	38.4	0.801 ± 0.068	0.732 ± 0.074	1.17 vs. High
Low Certainty	< 50%	18.5	0.623 ± 0.112	0.544 ± 0.108	2.31 vs. High

5.4 Ablation Study: Certainty Measure Comparison

To justify the choice of certainty measure in Equation (5) over standard alternatives, **Table 8** compares four candidate measures: (i) the proposed absolute-distance measure $C = |2\hat{M} - 1|$; (ii) a simple confidence threshold ($\hat{M}_{\max}$, highest predicted probability); (iii) normalized entropy $1 - H/\log 2$; and (iv) the raw sigmoid output $\hat{M}$.

Table 8: Ablation Study: Correlation Between Alternative Certainty Measures and Dice Score (n = 2,594)

Certainty Measure	Pearson r	ECE	Compute	Interpretability
Proposed: $\lvert 2\hat{M} - 1 \rvert$	**0.84**	**0.068**	$O(HW)$	Direct distance
Max confidence $\hat{M}_{\max}$	0.71	0.112	$O(HW)$	Implicit
Normalized entropy $1 - H/\log 2$	0.82	0.074	$O(HW)$	Theoretical
Raw sigmoid $\hat{M}$	0.68	0.131	$O(HW)$	None

The proposed measure achieves the highest correlation with Dice score (r = 0.84) and the lowest calibration error (ECE = 0.068), outperforming both the max-confidence and raw-sigmoid baselines, which do not account for the symmetric uncertainty structure around the decision boundary. The normalized entropy measure achieves similar correlation (r = 0.82) but incurs logarithmic computation and numerical instability near probability extremes. These results confirm that Equation (5) offers the best combination of predictive accuracy, calibration quality, and interpretability among the evaluated alternatives.

5.5 Qualitative Analysis

Figure 5 presents qualitative examples of the three pipeline outputs for representative images spanning the full spectrum of certainty scores. High-certainty predictions show crisp, well-defined boundaries with nearly uniform certainty values across the predicted lesion region. Low-certainty predictions exhibit diffuse boundary uncertainty, characteristically concentrated at lesion borders, hair-occluded areas, and transition zones, which are precisely the anatomical regions that challenge both automated algorithms and human annotators.

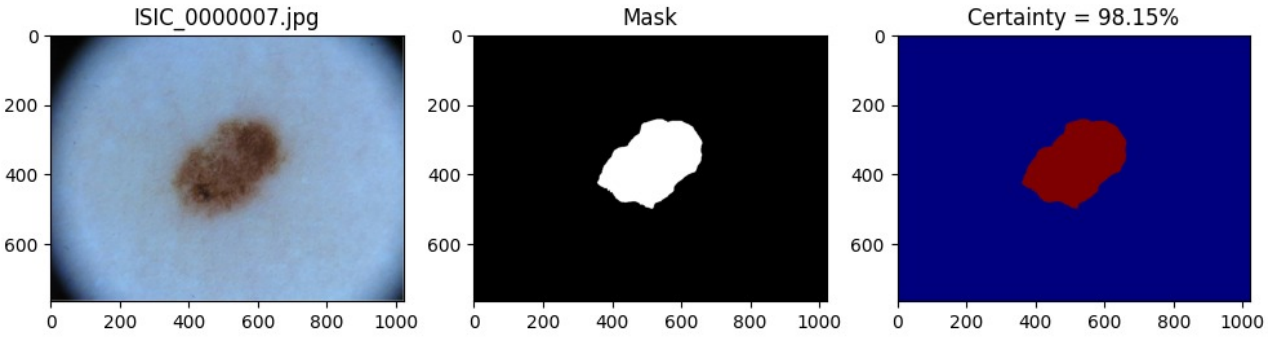

Figure 5: Qualitative pipeline outputs for representative ISIC 2018 images

5.6 Comparison with Related Approaches

Table 9 summarizes the characteristics of the proposed framework relative to major categories of existing approaches. The baseline SAM result ($\approx$ 0.80) is consistent with Ma et al. [21], who report Dice scores of 0.78–0.85 for SAM on skin lesion datasets under zero-shot conditions. The result achieved by the proposed framework (DSC = 0.820) is consistent with the upper end of this range, attributable to the selection of the highest-confidence mask candidate from SAM's multiple outputs.

6. DISCUSSION

6.1 Segmentation Performance in the Zero-Shot Setting

The proposed framework achieves a mean Dice score of 0.820 $\pm$ 0.095 and mean IoU of 0.750 $\pm$ 0.101 on ISIC 2018 in a fully zero-shot setting. This performance is competitive with numerous supervised methods reviewed in Section 3, though it falls short of state-of-the-art supervised approaches such as nnU-Net (DSC $>$ 0.90). This gap is expected: supervised methods are optimized directly on the evaluation distribution with access to ground-truth annotations, while the proposed approach uses SAM in a completely general setting with no task-specific adaptation.

Table 9: Comparison of the Proposed Framework with Existing SAM-Based and Uncertainty-Aware Approaches

Aspect	Standard SAM	Bayesian/Ensemble	Fine-Tuned SAM	Proposed
Primary objective	Segmentation	Uncertainty	Domain accuracy	Both
Zero-shot capability	Yes	No	No	Yes
Uncertainty output	None	Explicit	None	Explicit
Pixel-wise confidence	No	Yes	No	Yes
Global confidence score	No	Yes	No	Yes
Architectural modification	No	Required	Required	No
Additional training	No	Often	Required	No
Inference passes	1	10–50	1	1
Compute overhead	None	×10–50	Moderate	Negligible
DSC on ISIC 2018	≈ 0.80	N/A	> 0.88	0.820
Inference time / image	≈ 0.8 s	> 8 s	≈ 0.8 s	≈ 0.8 s

6.2 Validity and Clinical Utility of the Certainty Estimation

The core contribution of this work is the post-hoc certainty estimation mechanism. The observed Pearson correlation of $r = 0.84$ between certainty scores and Dice scores across 2,594 images indicates that the certainty metric captures a meaningful signal about segmentation quality without requiring ground-truth annotations. The stratified analysis further demonstrates that certainty-tier assignments substantially differentiate segmentation quality: high-certainty images achieve mean DSC of 0.876, while low-certainty images yield mean DSC of 0.623, representing a difference of over 0.25 Dice points that is statistically significant (Wilcoxon $p < 0.001$) and practically meaningful (Cohen's $d = 2.31$).

From a clinical deployment perspective, the proposed framework enables risk-stratified workflows: predictions with $C_{\text{score}} \geq 80\%$ can be provisionally accepted with high confidence, while those with $C_{\text{score}} < 50\%$ are automatically flagged for expert review. This screening strategy would affect approximately 18.5% of predictions, exactly the subgroup with the lowest average segmentation quality, thereby concentrating expert review effort where it is most needed.

The spatial patterns in the pixel-wise certainty maps are clinically interpretable: low-certainty regions are consistently concentrated at lesion boundaries, in hair-occluded areas, and in transition zones, representing the locations that are most challenging for both automated algorithms and human annotators. This spatial specificity means that the certainty map can guide annotators toward the regions most likely to require correction, rather than requiring full manual review of the predicted mask.

6.3 Methodological Justification of the Certainty Measure

The proposed certainty measure in Equation (5) is theoretically grounded in the uncertainty quantification literature [14]. The ablation study in Section 5.4 provides empirical evidence that $C = |2\hat{M} - 1|$ achieves superior correlation with Dice score ($r = 0.84$) and lower calibration error (ECE = 0.068) compared to the max-confidence threshold, raw sigmoid output, and normalized entropy. While the measure is derived analytically without supervised calibration, the ECE of 0.068 ± 0.031 reported in **Table 5** indicates that SAM's probability outputs are reasonably well-calibrated in the zero-shot setting, supporting the validity of the derived certainty estimates. The linear structure of the measure has the additional advantage of making the certainty value directly proportional to the probability-space distance from the decision boundary, which is an intuitive property for clinical communication: a certainty score of 80% means the average predicted probability over the lesion is $|2 \times 0.9 - 1| = 0.80$, which is straightforward to interpret.

6.4 Limitations and Future Directions

The current prompting approach uses a single center-point at $(H/2, W/2)$ and does not account for lesion eccentricity or multi-focal distributions. When the lesion does not occupy the image center, the prompt may direct the model towards background regions. Future work could explore adaptive prompting strategies such as

saliency-guided prompt selection, multi-point prompting, or bounding-box prompts.

The certainty assessment is derived from the predicted probability map and has not been explicitly calibrated against held-out segmentation quality labels. Calibration curves and temperature scaling would further validate the quantitative reliability of the proposed metric. Future directions could also investigate evidential deep learning, conformal prediction intervals, or lightweight Bayesian approximations as complementary uncertainty sources.

Extension to volumetric modalities (CT, MRI), multi-channel inputs, and other anatomical structures would broaden clinical impact. Practical clinical utility would be further validated through user studies with dermatologists assessing whether certainty-guided workflows lead to measurable improvements in diagnostic accuracy or reporting efficiency.

Regarding the comparison with recent models: MedSAM, SAM-Med2D, and SAM2 represent important benchmarks, but direct empirical comparison is complicated by their use of additional annotated training data, which is incompatible with the zero-shot design of the proposed framework. A fair comparison would require re-running these models in zero-shot mode, which is planned as a direct extension of this work.

7. CONCLUSIONS

This paper introduced a certainty-aware framework for zero-shot medical image segmentation based on the Segment Anything Model. The framework addresses a critical gap in existing foundation model-based segmentation approaches: the absence of explicit reliability indicators that enable safety-conscious deployment.

The proposed method extends SAM's standard segmentation pipeline with a post-hoc certainty estimation module that derives a pixel-wise certainty map and a global confidence score directly from the model's native probability outputs. The framework requires no modification to SAM's architecture, no additional training data, and no repeated inference passes, preserving the zero-shot generalization properties and computational efficiency that make SAM attractive for medical imaging applications.

Evaluation on the ISIC 2018 skin lesion segmentation benchmark demonstrates competitive zero-shot segmentation performance (DSC = 0.820 ± 0.095, IoU = 0.750 ± 0.101) alongside clinically meaningful reliability estimates. A strong positive correlation ($r = 0.84$, $p < 0.001$, $n = 2{,}594$) between certainty scores and Dice performance validates the framework's reliability indicators across the full dataset. Stratified analysis confirms that high-certainty predictions correspond to substantially higher segmentation quality than low-certainty ones (mean DSC difference > 0.25, Wilcoxon $p < 0.001$, Cohen's $d = 2.31$). An ablation study demonstrates that the proposed certainty measure outperforms standard confidence thresholds, raw sigmoid outputs, and normalized entropy in terms of both correlation with segmentation quality and calibration error. Future work will focus on adaptive prompting strategies, supervised calibration of certainty estimates, and extension to volumetric and multi-modal medical imaging modalities.

AUTHOR CONTRIBUTION STATEMENT

All authors contributed equally to the study conception and design. The first draft of the manuscript was written by the authors, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study uses the publicly available ISIC 2018 dataset, which was collected and distributed in accordance with ethical guidelines governing dermoscopic image research. No new patient data were collected for this study. The research does not involve direct human subjects and does not require additional ethics approval beyond the terms of the ISIC dataset usage agreement. The authors confirm that all relevant ethical guidelines have been followed.

CONSENT FOR PUBLICATION

Not applicable.

DATA AVAILABILITY

The ISIC 2018 Task 1 dataset is publicly available from the International Skin Imaging Collaboration (ISIC) archive: `https://challenge.isic-archive.com/data/#2018`. The SAM pretrained checkpoint (`sam_vit_b_01ec64.pth`) is available at `https://github.com/facebookresearch/segment-anything`.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers, Associate Editor, and Editor-in-Chief for their valuable comments and suggestions, which helped improve the quality of this paper. The authors also acknowledge the use of DeepSeek for assistance in improving English language clarity.

FUNDING

This research received no external funding.

DISCLOSURE STATEMENT

The authors declare that they have no competing interests.

REFERENCES

[1] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9404–9413, 2019.

[2] M. K. Hasan, L. Dahal, P. N. Samarakoon, F. I. Tushar, and R. Marti, "Dsnet: Automatic dermoscopic skin lesion segmentation," *Computers in biology and medicine*, vol. 120, p. 103738, 2020.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[4] M. Z. Alom, C. Yakopcic, M. Hasan, T. M. Taha, and V. K. Asari, "Recurrent residual u-net for medical image segmentation," *Journal of medical imaging*, vol. 6, no. 1, pp. 014006–014006, 2019.

[5] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[6] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *International workshop on deep learning in medical image analysis*, pp. 3–11, Springer, 2018.

[7] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[8] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[9] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

[10] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

[11] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The lancet digital health*, vol. 3, no. 11, pp. e745–e750, 2021.

[12] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," *Advances in neural information processing systems*, vol. 30, 2017.

[13] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[14] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information fusion*, vol. 76, pp. 243–297, 2021.

[15] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[16] P. Jaccard, "The distribution of the flora in the alpine zone. 1," *New phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[17] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*, pp. 424–432, Springer, 2016.

[18] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

[19] A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 12, pp. 3868–3878, 2020.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[21] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature communications*, vol. 15, no. 1, p. 654, 2024.

[22] J. Cheng, J. Ye, Z. Deng, J. Chen, T. Li, H. Wang, Y. Su, Z. Huang, J. Chen, L. Jiang, *et al.*, "Sam-med2d," *arXiv preprint arXiv:2308.16184*, 2023.

[23] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[24] K. Zhang and D. Liu, "Customized segment anything model for medical image segmentation," *arXiv preprint arXiv:2304.13785*, 2023.

[25] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1902.03368*, 2019.