Article

2AM: Weakly Supervised Tumor Segmentation in Pathology via CAM and SAM Synergy

Chenyu Ren [1], Liwen Zou [2,*] and Luying Gui [1,*]

[1] School of Mathematics and Statistics, Nanjing University of Science and Technology, Nanjing 210094, China; renchyu@emdesigner.com
[2] Department of Mathematics, Nanjing University, Nanjing 210093, China
[*] Correspondence: dz20210008@smail.nju.edu.cn (L.Z.); ly.gui@njust.edu.cn (L.G.)

Abstract

Tumor microenvironment (TME) analysis plays an extremely important role in computational pathology. Deep learning shows tremendous potential for tumor tissue segmentation on pathological images, which is an essential part of TME analysis. However, fully supervised segmentation methods based on deep learning usually require a large number of manual annotations, which is time-consuming and labor-intensive. Recently, weakly supervised semantic segmentation (WSSS) works based on the Class Activation Map (CAM) have shown promising results to learn the concept of segmentation from image-level class labels but usually have imprecise boundaries due to the lack of pixel-wise supervision. On the other hand, the Segment Anything Model (SAM), a foundation model for segmentation, has shown an impressive ability for general semantic segmentation on natural images, while it suffers from the noise caused by the initial prompts. To address these problems, we propose a simple but effective weakly supervised framework, termed as **2AM**, combining **C**AM and **S**AM for tumor tissue segmentation on pathological images. Our 2AM model is composed of three modules: (1) a CAM module for generating salient regions for tumor tissues on pathological images; (2) an adaptive point selection (APS) module for providing more reliable initial prompts for the subsequent SAM by designing three priors of basic appearance, space distribution, and feature difference; and (3) a SAM module for predicting the final segmentation. Experimental results on two independent datasets show that our proposed method boosts tumor segmentation accuracy by nearly 25% compared with the baseline method, and achieves more than 15% improvement compared with previous state-of-the-art segmentation methods with WSSS settings.

Keywords: weakly supervised segmentation; Class Activation Mapping; Segment Anything Model; adaptive point selection; tumor microenvironment analysis

1. Introduction

Tumor microenvironment (TME) analysis is a prominent issue in computational pathology. However, pathologists are facing challenges in accurate assessment and quantification for TME on pathological images [1]. Adopting algorithms to construct an automatic, efficient, end-to-end pathological image feature segmentation framework greatly assists physicians in diagnosis and in the in-depth study of tumor mechanisms [2]. Segmentation is a fundamental task in pathological image analysis involving the position detection and boundary description of regions of interest (ROIs) such as tissues and nuclei. Accurate

tumor tissue segmentation is crucial for many clinical applications, including disease diagnosis, treatment planning, and monitoring disease progression.

In recent years, deep learning technology has been greatly developed for medical image analysis [3–6]. The uniqueness of deep learning lies in its ability to extract features automatically. Using deep learning methods, features, including tumor morphology, size, and density, can be automatically extracted from images in tumor image analysis. These features are crucial for diagnosis and treatment, which greatly reduce the pressure on doctors' work. Fully supervised learning takes manual annotation as the gold standard for delineating anatomical structures and pathological areas. However, it is always time-consuming and labor-intensive and requires specialized expertise [7]. Weakly supervised semantic segmentation (WSSS) aims to learn the concept of segmentation using image-level class labels and has shown great potential in pathological image analysis [8].

The Class Activation Map (CAM) [9] facilitates the identification of representative regions within images, contributing to an enhanced understanding of how neural networks respond to different categories and exposing potential mechanisms behind these activations. However, CAM usually cannot precisely delineate the boundaries of identified targets due to the lack of pixel-wise supervision. Conversely, the Segment Anything Model (SAM) [10], a foundation model for segmentation, has demonstrated impressive effectiveness and efficiency in semantic segmentation tasks on natural images. Nevertheless, it shows the weakness in robustness caused by the noise from unreliable prompts. To address these problems, we propose a simple but effective weakly supervised framework, termed as 2AM, combining CAM and SAM for tumor tissue segmentation on pathological images. Our 2AM model is composed of three modules: (1) a CAM module for generating salient regions for tumor tissues on pathological images; (2) an adaptive point selection (APS) module for providing more reliable initial prompts for the subsequent SAM by designing three priors of basic appearance, space distribution, and feature difference; and (3) a SAM module for predicting the final segmentation. Figure 1 shows two experimental examples to illustrate the benefits of the proposed 2AM method in tumor segmentation on pathological images.
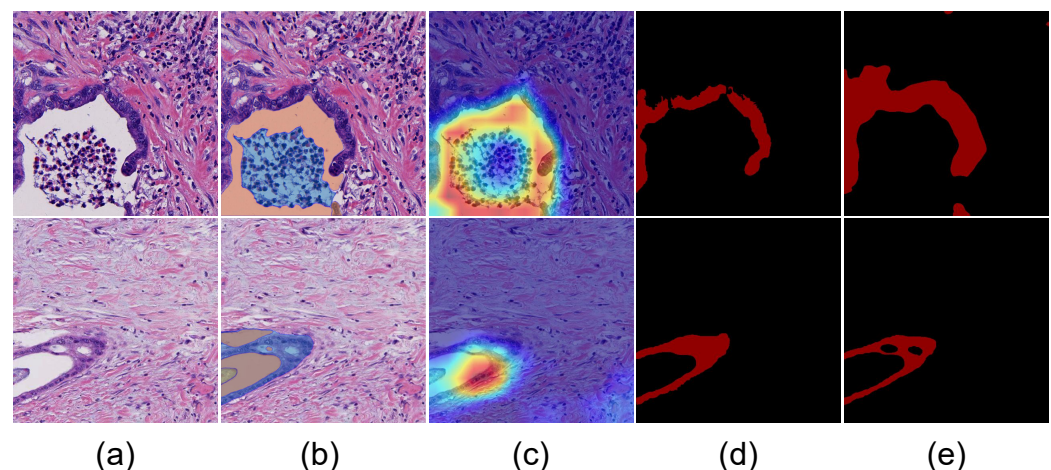

Figure 1. Example pathology images. Two examples to illustrate the benefits of the proposed 2AM method in tumor tissue segmentation on pathological images: (**a**) original pathological images, (**b**) SAM prediction, (**c**) CAM prediction, (**d**) 2AM (ours) prediction, and (**e**) ground truth.

The main contributions of our work are summarized as follows:

- To the best of our knowledge, this is the first work to combine Class Activation Mapping (CAM) and the Segment Anything Model (SAM) for pathological image analysis, and we propose a simple but effective framework, termed as 2AM, for tumor segmentation on pathological images in a weakly supervised manner.

- We present an adaptive point selection (APS) module for providing more reliable initial prompts for the subsequent SAM by designing three priors of basic appearance, space distribution, and feature difference, which achieves effective integration of CAM and SAM.
- Experimental results on two independent datasets show that our proposed method boosts tumor segmentation accuracy by nearly 25% compared with the baseline method, and achieves more than 15% improvement compared with previous state-of-the-art segmentation methods in a weakly supervised manner.

2. Related Work

2.1. Pathological Image Segmentation

Semantic segmentation has evolved substantially with the development of techniques designed initially for natural images, now effectively applied in pathological image segmentation. The seminal hierarchical encoder–decoder model, U-Net [11], established a foundational framework for medical image segmentation. Subsequent innovations have led to the development of adaptive algorithms like nnU-Net [12], tailored specifically for medical applications. A pivotal advancement was the introduction of Transformer technology [13], which not only enhanced image segmentation tasks through models such as the Swin Transformer [14], but also facilitated novel integrations with U-Net, creating hybrid models like Swin-U-Net [15]. The recent emergence of the ConvNeXt model [16] underscored the capability of pure convolutional neural networks to outperform Swin Transformers in accuracy and inference speed with comparable parameter counts.

Moreover, significant strides have been made in applying these AI models to practical challenges in pathological image segmentation. Breakthroughs in delineating fine structures, such as tertiary lymphoid structures [8] and cell nuclei [17], have been reported. In 3D image processing, masked image modeling (MIM) techniques [18] have been adapted for self-supervised learning in medical imaging. High-quality original images essential for accurate segmentation are now more attainable using generative adversarial networks (GANs), which generate high-resolution, large-scale histopathological images for segmentation purposes [19].

2.2. Class Activation Mapping

Class Activation Mapping (CAM) offers a viable approach for weakly supervised image segmentation by leveraging only image-level labels, thus circumventing the need for exhaustive pixel-level annotations. This technique utilizes convolutional neural networks to derive pixel-level predictions from feature maps. The core methodology transforms feature maps into a comprehensive Class Activation Map via global average pooling (GAP) at the final convolutional layer, identifying the activation level of each pixel relative to specific image classes. Critical developments in CAM include grad-CAM [9], the HistoSegNet framework for pixel-level segmentation of whole slide images (WSIs) [20], the SEAM method integrating attention mechanisms with self-supervised loss [21], and MCTformer, which combines Transformer technology with CAM for enhanced accuracy [22].

2.3. Segment Anything Model

The Segment Anything Model (SAM) represents a frontier in segmentation technology, capable of producing high-quality object masks based on precise input prompts, such as points or boxes, to encapsulate all objects within an image. Developed from a training dataset encompassing 11 million images and 11 billion masks, SAM has demonstrated exceptional zero-shot transfer capabilities across new image distributions and tasks, achieving remarkable segmentation results. The specialized version for medical applications,

MedSAM [23], utilizes a meticulously curated dataset of over 1 million images, setting new benchmarks in segmenting essential biomarkers critical for tumor burden quantification and advancing diagnostic tools and personalized treatment strategies. While several iterations based on SAM have been explored in the medical domain [24–26], the relevance of these advances remains a focal point of ongoing research. To underscore the distinct advantages of our method, we have employed the original SAM model, which is uniquely augmented with negative point annotations to enhance the efficacy and precision of segmentation.

3. Methods

3.1. Pipeline

As illustrated in Figure 2, our proposed 2AM method is composed of three modules: (1) a CAM module for generating salient regions for tumor tissues on pathological images; (2) an adaptive point selection (APS) module for providing more reliable initial prompts for the subsequent SAM by designing three priors of basic appearance, space distribution, and feature difference; and (3) a SAM module for predicting the final segmentation.
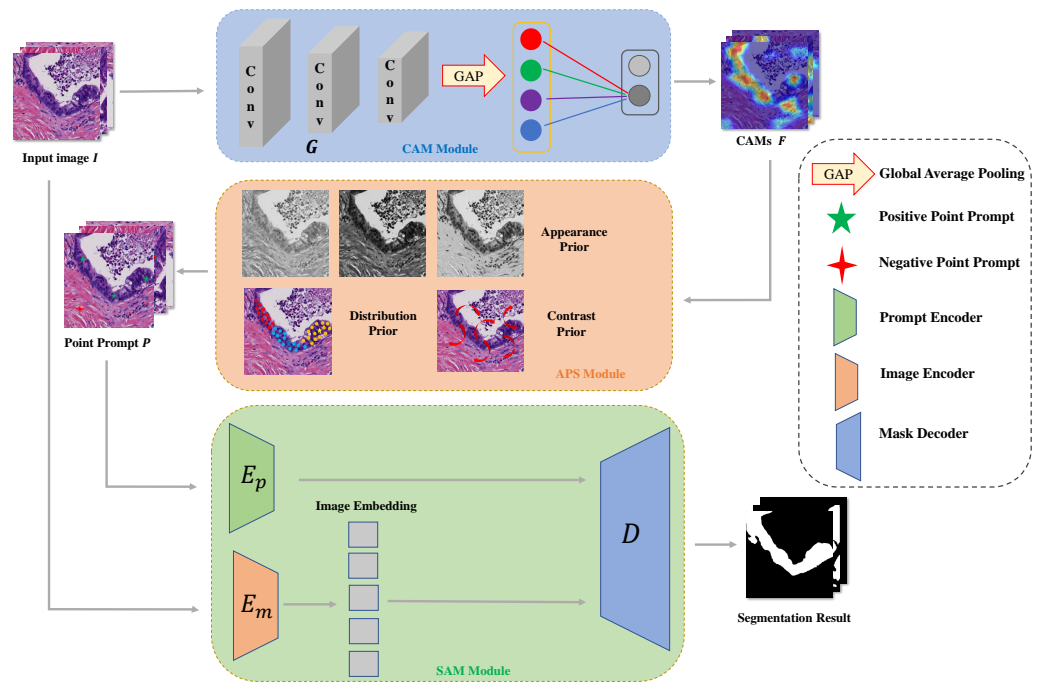

Figure 2. Overview of 2AM pipeline. Pipeline of our proposed 2AM model, which is composed of three modules: (1) a CAM module for generating salient regions for tumor tissues on pathological images; (2) an adaptive point selection (APS) module for providing more reliable initial prompts for the subsequent SAM by designing three priors of basic appearance, space distribution, and feature difference, respectively; and (3) a SAM module for predicting the final segmentation.

3.2. CAM Module

First, we define an image classifier G for tumor tissue patch identification (nerve tissue patch as the negative class) as a Class Activation Maps generator. The encoder G is designed for extracting the CAMs $F \in \mathbb{R}^{W \times H \times C}$ from an input image $I \in \mathbb{R}^{W \times H \times 3}$ as follows:

$$F = G(I), \tag{1}$$

where $C = 2$ is the number of classes. Then, we obtain an image-level class prediction logit $\mathbf{y} \in \mathbb{R}^C$ as follows:

$$y = GAP(F), \tag{2}$$

where GAP is a global average pooling layer along the spatial axes. We adopt the binary cross-entropy loss to train the encoder G, as follows:

$$L_{BCE}(p, \hat{y}) = -(\hat{y} \log p + (1 - \hat{y}) \log(1 - p)), \tag{3}$$

where $p = argmax(sigmoid(\mathbf{y}))$ is the predicted confidence for tumor identification and $\hat{y}$ is the image-level classification label.

3.3. Adaptive Point Selection (APS) Module

It is known that SAM always suffers from a noise prompt, such as an unreliable point prompt, leading to failure segmentation. How to obtain reasonable prompts is the devil in foundation segmentation models. In this section, we introduce an adaptive point selection (APS) module to efficiently combine CAM and SAM for tumor segmentation on pathological images. The APS module consists of three designed priors of basic appearance, space distribution, and feature difference.

The appearance prior calculates the mean value of tumor/background color from the image-level labels in the training set, while the contrast prior uses algebraic operations. The computational complexity of both is O(1). The distribution prior adopts k-means++ (with a complexity of O(n)), which is independent of training. This design ensures the robustness of the APS module when facing different datasets.

Appearance prior. This prior ensures that prompts align with tumor-specific visual features (e.g., color and texture) derived from H&E staining, such as darker nuclei in tumor cells. By anchoring prompts to these characteristics, we reduce SAM's tendency to respond to irrelevant cues. Pathological images exhibit distinct appearance features between tumors and the surrounding areas, such as the intensity and the texture. Therefore, a reliable positive point prompt should be compatible with a certain appearance requirement. Mathematically, for testing an image set $\{I_i^s; M_i^s\}_{i=1}^L$, where $I_i^s \in \mathbb{R}^{W \times H \times 3}$ denotes the testing image, $M_i^s \in \{0,1\}^{W \times H}$ denotes the ground truth mask, and L denotes the number of testing images, we calculate the average appearance representation $T \in \mathbb{R}^3$ and $B \in \mathbb{R}^3$ for the tumor tissue and background, respectively, as follows:

$$T = \frac{\sum_{i=1}^L \sum_{x,y \in I_i^s} I_i^s(x,y) M_i^s(x,y)}{\sum_{i=1}^L \sum_{x,y \in I_i^s} M_i^s(x,y)}, \tag{4}$$

$$B = \frac{\sum_{i=1}^L \sum_{x,y \in I_i^s} I_i^s(x,y)(1 - M_i^s(x,y))}{\sum_{i=1}^L \sum_{x,y \in I_i^s} (1 - M_i^s(x,y))}, \tag{5}$$

For any positive point prompt candidate $Q = (x,y) \in \mathbb{R}^2$, it should be in the neighbor appearance field of the mean tumor representation T and as far away from B as possible. Therefore, we define the appearance prior (AP) as follows:

$$I(x,y) \in U_{ap}(T, R_{ap}) \in \mathbb{R}^3, \tag{6}$$

$$R_{ap} = \alpha \min\{|I_m - T|, |T - B|, |B|\}, \tag{7}$$

where $U_{ap}(T, R_{ap})$ denotes the neighborhood with T as the center and R_{ap} as the radius, α denotes the parameter for radius controlling, and I_m denotes the max intensity of the pathological images.

Distribution prior. Tumors on pathological images are often spatially dispersed and heterogeneous. By using clustering, this prior places prompts in distinct tumor regions,

ensuring comprehensive coverage of the tumor area, thereby enhancing SAM's segmentation of the tumor region under the given prompts. In order to select representative positive point prompts as much as possible, we also take the distribution requirements into account. Mathematically, an input image $I \in \mathbb{R}^{W \times H \times 3}$, and its CAM-based segmentation $S \in \{0,1\}^{W \times H}$, which is calculated as follows:

$$S(x,y) = \begin{cases} 1, & F(x,y) > \gamma \\ 0, & otherwise \end{cases} \tag{8}$$

where γ is the confidence threshold of the predicted CAMs. Then, we obtain the point sets J of the predicted foreground class, as follows:

$$J = \{(x,y)|S(x,y) = 1\}. \tag{9}$$

The K-means++ (KMPP) algorithm is adopted for generating the cluster centers of these foreground points. In this work, we set the number of clusters to 3. Therefore, we can obtain three centers of spatial distribution, as follows:

$$C_1, C_2, C_3 = KMPP(J). \tag{10}$$

For any positive point prompt candidate $Q_i = (x,y) \in \mathbb{R}^2, i = 1,2,3$, in addition to meeting the AP requirement defined above, it should also be in the neighbor spatial field of the cluster centers. Therefore, we define the distribution prior (DP) as follows:

$$Q_i = (x,y) \in U_{dp}(C_i, R_{dp}^i) \in \mathbb{R}^2, i = 1,2,3 \tag{11}$$

$$R_{dp}^i = \beta \max_{p,q \in C_i} d(p,q), \tag{12}$$

where $U_{dp}(C_i, R_{dp}^i)$ denotes the neighborhood with C_i as the center and R_{dp}^i as the radius, β denotes the parameter for radius controlling, and $d(p,q)$ is the Euclidean distance between p and q. Selecting points adjacent to these centers as prompts further enhances their representativeness and ensures a geographically distributed coverage. The inherent separation between the centers precludes the aggregation of prompt points, thereby adhering to the desired distribution prior. This methodology not only supports a more systematic exploration of the spatial features but also optimizes the segmentation process by ensuring a diverse yet specific sampling of the area under study.

Contrast prior. Leveraging SAM's ability to refine segmentation with negative points, this prior places points in non-tumor regions to sharpen boundaries, countering SAM's sensitivity to ambiguous edges in low-contrast images. Therefore, we adopt the negative point strategy for SAM prompt generation. We first randomly select three positive point prompts that meet the AP and DP requirements, as follows:

$$Q_1, Q_2, Q_3 \sim AP \cap DP; \tag{13}$$

then we sample a negative point Q_0 to construct the contrast prior (CP), as follows:

$$Q_0 \sim \{(x,y)|S(x,y) = 0\} \tag{14}$$

Based on the three designed priors above, we can obtain a more reliable point prompt for the subsequent SAM module.

The following Algorithm 1 integrates these priors to select three positive points and one negative point, ensuring robust prompt generation for SAM.

Algorithm 1 Adaptive Point Selection (APS) for Tumor Segmentation

---

1: **Input:** CAM heatmap, Image
2: **Output:** Positive points P_1, P_2, P_3, Negative point N
3: **Compute color prior:** Calculate mean tumor color (R_t, G_t, B_t) from training images
4: **Select positive points:**
    Extract high-response region A (CAM score > 0.95)
    Apply K-means++ clustering on A ($k = 3$, 10 iterations) to select P_1, P_2, P_3
    Ensure minimum distance d between points using distance prior

5: **Select negative point:** Randomly choose N from region with CAM score < 0.1

3.4. SAM Module

After obtaining the positive points Q_1, Q_2, Q_3 and the negative point Q_0 through the above methods, we can input them into SAM [10] as point prompts. The process is as follows:

$$M^c = D(E_m(I), E_p(P)), \tag{15}$$

where I represents the input image; $P = \{Q_1, Q_2, Q_3, Q_0\}$ represents the point prompts; M^c represents the segmentation results; and E_m, E_p, and D denote the image encoder, prompt encoder, and mask decoder of the SAM module [10].

4. Experiments

4.1. Datasets

We evaluate our proposed method on three independent datasets for weakly supervised tumor segmentation on pathological images. It should be noted that our division of the dataset is based on patients, which can effectively prevent data leakage caused by overlapping slides.

Training set. We collected 1000 pathological H&E images from the public PAIP dataset [27] to train the CAM module, each of which is with a dimension of 512×512. Among them, 800 were used for training, approximately 120 patients, and 200 were used for validation, approximately 30 patients. There are 500 images containing tumor tissues with the positive class labels and 500 images containing nerve tissues with the negative class labels.

Testing set. We evaluated our proposed method on two independent datasets collected from Nanjing Drum Tower Hospital (NDTH) and Jiangsu Province Hospital of Chinese Medicine (JHCM). The NDTH dataset consists of 200 H&E images (approximately 30 patients) with a 512×512 size and pixel-wise tumor tissue annotations. The JHCM dataset consists of 100 H&E images (approximately 20 patients) with a 512×512 size and pixel-wise tumor tissue annotations. All the manual annotations of tumor tissues were performed by two experienced pancreatic pathologists.

4.2. Metrics

For the quantitative evaluation of the segmentation performance of tumor tissues on pathological images, we used two commonly used evaluation metrics, namely, Dice similarity coefficient (Dice) and Hausdorff Distance (HD). The Dice metric is used to measure the degree of overlap between the segmentation results and the ground truth labels. The HD metric is used to measure the accuracy of boundary segmentation. They can be calculated as follows:

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN} = \frac{2 \cdot |M \cap N|}{|M| + |N|}, \tag{16}$$

$$HD(M, N) = \max(hd(M, N), hd(N, M)), \tag{17}$$

$$hd(M, N) = \max_{m \in M} \min_{n \in N} d(m, n), \tag{18}$$

where *M* represents the prediction and *N* represents the ground truth. Additionally, we also adopted two metrics, namely, False-Positive Segmentation Rate (FPSR) and False-Negative Segmentation Rate (FNSR), to measure the discrepancy between the prediction and ground truth [28], which can be calculated as follows:

$$FPSR(M, N) = \frac{|M - N|}{|M| + |N|} \tag{19}$$

$$FNSR(M, N) = \frac{|N - M|}{|M| + |N|} \tag{20}$$

4.3. Implementation Details

The proposed scheme was developed using Python 3.8.0, with PyTorch 2.1.0. The machine on which the experiments were conducted has an GeForce RTX 4090 GPU (produced by Inspur Information Company in Shandong, China) with 24 GB memory. All the learning-based models were trained for 200 epochs with a batch size of 4. We adopted ResNet-18 as the backbone of the image classifier *G*. For SAM, we utilized the officially released version with the ViT-H model (SAM-ViT-H-4b8939) (https://github.com/facebookresearch/segment-anything, accessed on 12 April 2024). We set the parameters of radius controlling α and β to 0.5 in our experiments. We also set the confidence threshold γ of the CAM-based methods to 0.5 for obtaining the binary segmentation masks. The Adam optimizer was used in our training, with a learning rate $= 1 \times 10^{-4}$ and a batch size = 4.

4.4. Computational Efficiency Analysis

To assess the computational requirements of the proposed 2AM framework, we theoretically estimate the number of parameters and floating point operations (FLOPs) for each component. The parameter counts are calculated by summing all trainable weights in the models. The FLOPs are computed for an input size of 512×512 pixels.

As shown in Table 1, the CAM module (based on ResNet-18) contains 12M parameters and requires 4 GFLOPs. The SAM module (ViT-H) is significantly larger, with 636M parameters and 256 GFLOPs. Therefore, the entire 2AM framework has approximately 648M parameters and 260 GFLOPs. For comparison, we also include the parameter counts and FLOPs of several state-of-the-art segmentation networks used in our weakly supervised experiments. U-Net [11] has 34M parameters and 180 GFLOPs, while DeepLabV3+ [29] (with ResNet-101 backbone) requires 58M parameters and 350 GFLOPs. Swin Transformer (Swin-T) [14] has 60M parameters and 250 GFLOPs for a 512×512 input.

Table 1. Theoretical computational efficiency comparison.

Method	Parameters (M)	FLOPs (G)
U-Net	34	180
DeepLabV3+	58	350
Swin Transformer	60	250
2AM (Ours)	648	260

4.5. Ablation Study

Our main contribution lies in designing an adaptive point selection (APS) module to effectively combine CAM and SAM for weakly supervised tumor tissue segmentation on

pathological images. In this section, we present the ablation study to verify the effectiveness of the three priors in our proposed APS module. We show the ablation study results in Table 2 to investigate the individual impact of each prior; the best performance is shown in bold. In Table 2, the without CP, the AP and DP experiment (Dice 55%) approximates a no-SAM baseline, where CAM masks are refined by morphological operations, revealing coarse boundaries and limited handling of tumor heterogeneity [9].

Table 2. Experimental results of the ablation study on two testing datasets. CP, AP, and DP denote contrast prior, appearance prior and distribution prior, respectively. A tick indicates the use of this prior. Arrow indicates the direction towards better performance. Bold denotes the best performance.

Datasets	CP	AP	DP	HD ↓	FPSR ↓	FNSR ↓	Dice ↑
NDTH				215.29 ± 47.76	33.24 ± 11.48	12.14 ± 5.77	54.62 ± 11.86
	✓			211.03 ± 45.58	17.60 ± 11.81	25.75 ± 10.29	56.65 ± 14.47
		✓		221.61 ± 46.16	25.06 ± 12.57	19.68 ± 13.20	55.26 ± 11.04
			✓	238.61 ± 31.13	28.38 ± 11.38	12.77 ± 6.78	58.85 ± 12.07
	✓	✓		185.48 ± 54.87	$\mathbf{6.53 \pm 2.93}$	23.15 ± 12.57	66.32 ± 11.33
	✓		✓	197.54 ± 51.31	22.54 ± 12.38	14.51 ± 4.81	62.96 ± 14.67
		✓	✓	197.91 ± 37.83	25.44 ± 13.48	$\mathbf{11.78 \pm 5.78}$	62.78 ± 10.87
	✓	✓	✓	$\mathbf{176.92 \pm 47.77}$	6.59 ± 2.96	17.47 ± 5.35	$\mathbf{70.94 \pm 9.64}$
JHCM				170.10 ± 93.53	30.23 ± 13.50	11.72 ± 5.18	58.05 ± 15.72
	✓			143.90 ± 58.09	27.23 ± 11.96	6.77 ± 2.75	65.99 ± 12.91
		✓		180.99 ± 93.86	$\mathbf{22.45 \pm 13.91}$	14.70 ± 9.94	62.85 ± 16.66
			✓	183.58 ± 95.42	32.66 ± 12.15	6.78 ± 4.91	63.56 ± 10.57
	✓	✓		139.06 ± 59.96	19.44 ± 11.24	9.20 ± 4.43	65.36 ± 12.33
	✓		✓	$\mathbf{121.47 \pm 60.65}$	32.19 ± 9.70	4.49 ± 2.45	66.02 ± 9.86
		✓	✓	190.41 ± 97.00	35.22 ± 11.50	4.74 ± 2.49	63.03 ± 11.04
	✓	✓	✓	121.65 ± 65.29	29.95 ± 9.66	$\mathbf{3.64 \pm 2.52}$	$\mathbf{68.41 \pm 9.29}$

Effectiveness of Appearance Prior (AP). Without using AP, the performance significantly drops on our collected datasets, leading to a decrease of 2.39–7.98% in Dice and 2.24–15.95% in FPSR.

Effectiveness of Distribution Prior (DP). Without using DP, the performance significantly drops on our collected datasets, leading to a decrease of 3.05–4.62% in Dice, 8.56–17.41 in HD, and 5.56–5.68% in FNSR.

Effectiveness of Contrast Prior (CP). Without using CP, the performance significantly drops on our collected datasets, leading to a decrease of 5.38–8.16% in Dice and 20.99–68.76 in HD.

Figure 3 shows a visual comparison of the ablation study for the proposed three priors in the APS module. It can be found that our approach not only excelled in the overall segmentation accuracy, but also closely approached the optimal metrics in other areas, demonstrating minor deviations compared with more significant variances observed in alternative methods. Meanwhile, although our dataset introduces variations in color and tumor heterogeneity, the appearance prior detects darker tumor regions that are not affected by staining changes, which emphasizes the effectiveness and robustness of the prior and model configuration we have selected. In Figure 3, failure modes are image-specific: missing the appearance prior causes multi-focal lesion misses (Figure 3b), lacking the distribution prior leads to clustered prompts (Figure 3c), and omitting the contrast prior results in stroma over-segmentation (Figure 3d).
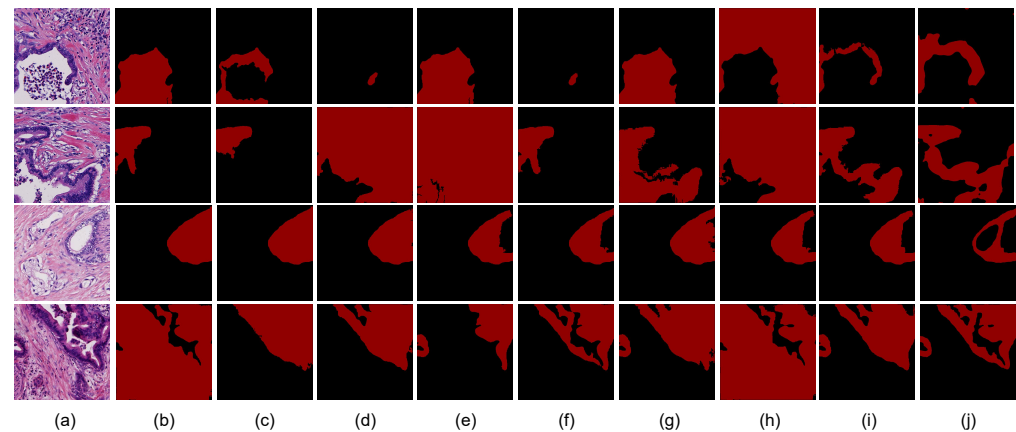
Figure 3. Ablation study on APS priors. Visual comparison of the ablation study in the APS module. The four testing images above are from the NDTH dataset, and the four testing images below are from the JHCM dataset. CP, AP, and DP denote contrast prior, appearance prior, and distribution prior, respectively. (**a**) Image; (**b**) w/o CP, AP, and DP; (**c**) w/o AP and DP; (**d**) w/o CP and DP; (**e**) w/o CP and AP; (**f**) w/o DP; (**g**) w/o AP; (**h**) w/o CP; (**i**) w/ CP, AP, and DP; (**j**) ground truth.

4.6. Comparison with Other Weakly Supervised Segmentation Methods

In this section, we compare our proposed method with other weakly supervised segmentation methods for tumor tissue segmentation on pathological images. We compare with the following two types of methods.

CAM-based methods. In the field of weakly supervised learning, traditional CAM methods and their derivative models, such as Grad-CAM [30] and Grad-CAM++ [31], typically enhance target segmentation accuracy by refining salient feature areas to achieve semantic segmentation with image-level labels. However, our approach adopts an innovative strategy that does not focus on improving the accuracy of these feature areas. Instead, it transforms these regions into high-precision prompt points and leverages the robust segmentation capabilities of SAM to generate accurate segmentation regions from these points. This essentially represents an effective enhancement of the CAM technology. In our experiments, we compare our method with CAM and its enhanced algorithms, Grad-CAM and Grad-CAM++, for benchmarking. Experimental results on two independent testing datasets are illustrated in Table 3. It can be found that our method achieves more than 20% improvement for tumor segmentation compared with these CAM-based methods. Figure 4 presents a visual comparison of the segmentation performances between our method and other CAM-based methods.

Table 3. Quantitative comparison results between CAM-based methods and our proposed method for tumor segmentation on two independent testing datasets. Arrow indicates the direction towards better performance. Bold denotes the best performance. * and $^+$ denotes that our method achieves statistically significant $p < 0.01$ and $p < 0.05$ performance improvement compared with other methods using the paired t-test.

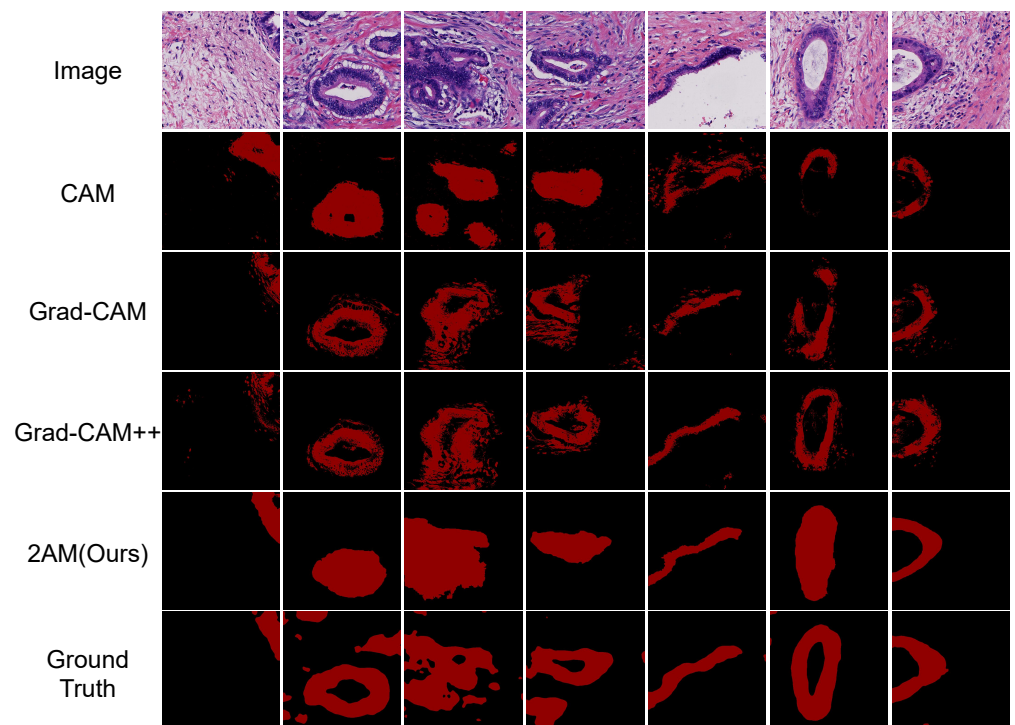| Datasets	Methods	HD ↓	FPSR ↓	FNSR ↓	Dice ↑
NDTH	CAM	257.61 ± 62.01 *	12.99 ± 6.38 *	32.52 ± 11.06 *	46.50 ± 8.58 *
	Grad-CAM	219.80 ± 36.44 *	9.99 ± 2.98 $^+$	39.50 ± 4.09 *	47.75 ± 6.15 *
	Grad-CAM++	226.79 ± 46.08 *	12.91 ± 4.43 *	39.34 ± 5.54 *	50.51 ± 5.06 *
	2AM(Ours)	**176.92 ± 47.77**	**6.59 ± 2.96**	**17.47 ± 5.35**	**70.94 ± 9.64**
JHCM	CAM	212.97 ± 84.14 *	**15.24 ± 6.51**	17.10 ± 8.72 *	42.66 ± 6.93 *
	Grad-CAM	222.41 ± 79.88 *	16.69 ± 6.42	29.98 ± 12.42 *	43.33 ± 8.37 *
	Grad-CAM++	218.27 ± 75.48 *	16.40 ± 7.03	25.26 ± 12.41 *	48.34 ± 8.75 *
	2AM(Ours)	**121.65 ± 65.29**	29.95 ± 9.66	**3.64 ± 2.52**	**68.41 ± 9.29**

Figure 4. CAM comparison results. Visual comparison results between CAM-based methods and our proposed method for tumor segmentation on the NDTH dataset.

Segmentation networks with a weakly supervised manner. We also compare with other state-of-the-art (SOTA) convolutional neural networks (CNNs) or Transformer-based networks in a weakly supervised manner. Specifically, we utilize the generated CAM-based binary masks to train the networks with Dice and cross-entropy loss. In our experiments, we compared our method with U-Net [11], Swin Transformer [14], and DeepLabV3+ [29]. Experimental results on two independent testing datasets are illustrated in Table 4. It can be found that our method achieves more than 13%–19% improvement for tumor segmentation compared with these SOTA methods in a weakly supervised manner. Figure 5 presents a visual comparison of the segmentation performance between our method and other CAM-based methods.

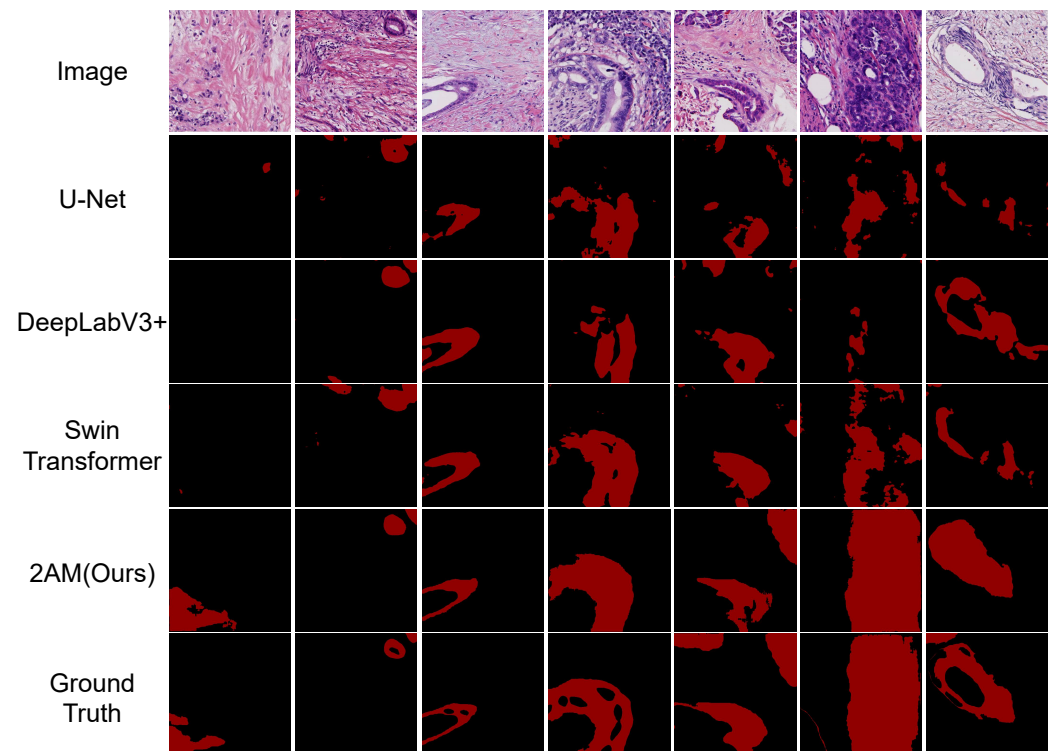**Table 4.** Quantitative comparison results between weakly supervised segmentation networks and our proposed method for tumor segmentation on two independent testing datasets. Arrow indicates the direction towards better performance. Bold denotes the best performance. * denotes that our method achieves statistically significant $p < 0.01$ performance improvement compared with other methods using the paired t-test.

Datasets	Methods	HD ↓	FPSR ↓	FNSR ↓	Dice ↑
NDTH	DeepLabV3+	169.13 ± 59.79	**1.04 ± 0.48**	50.34 ± 14.67 *	48.42 ± 14.41 *
	Swin Transformer	167.51 ± 47.72	2.41 ± 1.18	32.33 ± 11.85 *	55.26 ± 10.92 *
	U-Net	**140.37 ± 40.58**	1.79 ± 0.69	31.13 ± 9.83 *	57.09 ± 9.65 *
	2AM(Ours)	176.92 ± 47.77	6.59 ± 2.96	**17.47 ± 5.35**	**70.94 ± 9.64**
JHCM	DeepLabV3+	198.67 ± 90.70 *	**17.42 ± 6.26**	36.39 ± 19.15 *	46.19 ± 14.52 *
	Swin Transformer	249.00 ± 73.80 *	25.78 ± 6.72	24.72 ± 14.83 *	49.50 ± 11.28 *
	U-Net	258.79 ± 80.68 *	23.71 ± 7.73	31.56 ± 14.11 *	44.73 ± 11.77 *
	2AM(Ours)	**121.65 ± 65.29**	29.95 ± 9.66	**3.64 ± 2.52**	**68.41 ± 9.29**

It should be noted that, in order to fairly evaluate the contribution of SAM in our proposed 2AM method, we trained all models, including the baselines, using the same

CAM-generated annotations. While these annotations are less accurate by design, they ensure a consistent comparison across methods. We performed 200,000 training iterations for the baseline models (e.g., DeepLabV3+, Swin Transformer, and U-Net), ensuring sufficient training to guarantee the accuracy and effectiveness of the models; however, the limited accuracy of the CAM annotations constrained their Dice scores and Hausdorff Distance (HD) values, as observed in Table 4.

The paired t-test experiments are conducted to illustrate the statistical significance of the performance comparison. Figure 6 illustrates the performance distributions for the comparison experiment. It can be found that our proposed method has a more compact distribution with significant advantage.

Figure 5. WSSS comparison results. Visual comparison results between weakly supervised segmentation networks and our proposed method for tumor segmentation on the JHCM dataset.
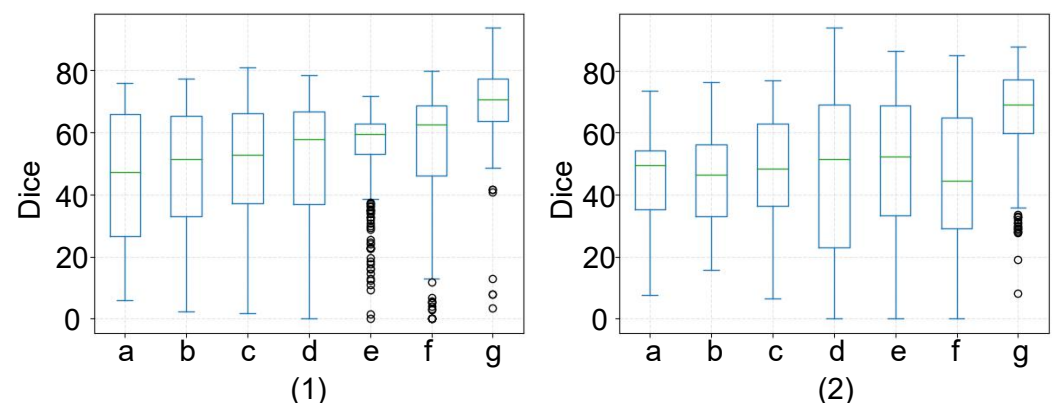

Figure 6. Performance distribution across datasets. Performance distribution of different methods for tumor segmentation on (**1**) the NDTH testing dataset and (**2**) the JHCM dataset. (a) CAM, (b) Grad-CAM, (c) Grad-CAM++, (d) DeepLabV3+, (e) Swin Transformer, (f) U-Net, and (g) 2AM (ours).

5. Discussion

5.1. Failure Case Analysis and Limitations

In natural images, an object typically exhibits high contrast, well-defined edges, and consistent textures, which allow segmentation models like SAM to accurately interpret prompts and delineate boundaries. In contrast, pathological images, such as H&E-stained histological slides, present unique challenges due to low contrast, irregular textures, and blurred boundaries. These issues stem from factors such as staining variability, tissue preparation artifacts, and inherent biological complexity. For example, uneven staining can cause different tissue regions or tissue-background interfaces to appear visually similar, creating ambiguous boundaries or false gradients (e.g., pseudo-edges from staining inconsistencies). These noisy cues often mislead SAM into misinterpreting object boundaries, increasing its sensitivity to inaccurate or ambiguous prompts and leading to segmentation errors. Our proposed framework integrates Class Activation Mapping (CAM) and the Segment Anything Model (SAM). It leverages its strengths to transition from image-level classification labels to precise pixel-level semantic segmentation. Initially, the CAM module efficiently identifies salient regions of tumor tissues in pathological images; however, it does not precisely delineate the target boundaries. To overcome this, we transform these initial coarse maps into reliable prompt points for SAM, employing robust segmentation capabilities to achieve accurate final segmentations. This strategy effectively bridges the coarse localization by CAM and the high-resolution delineation by SAM, addressing the challenge posed by SAM's requirement for precise initial prompts.

Despite its successes, our approach encounters limitations when CAM fails to recognize targets, a situation exacerbated by the low contrast in tissue differentiation typical in pathological images. For instance, as illustrated in Figure 7, CAM may only identify some targets when multiple are present or may incorrectly localize tumor areas. These scenarios result in suboptimal segmentation outcomes. Examples of such failures are due to the fact that the staining of the tumor is close to the surrounding color, and it is difficult to distinguish the failure, which may be caused by factors such as uneven staining, although it is inevitable, but the frequency is not high. Thus, while our framework advances the field of weakly supervised segmentation and enhances accuracy and efficiency, it also highlights the necessity for future improvements in CAM's target recognition capabilities to bolster the overall effectiveness of the segmentation process.
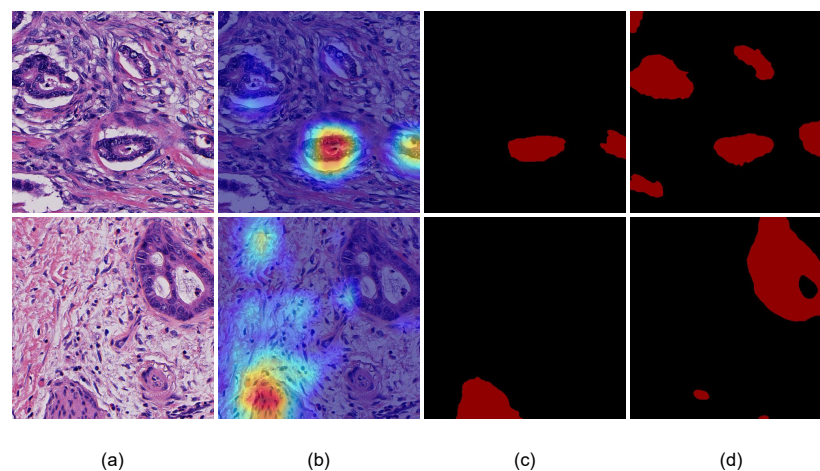

Figure 7. Two failure cases. (**a**) Original pathological images, (**b**) CAM prediction, (**c**) 2AM (ours) prediction, and (**d**) ground truth.

5.2. Future Directions

In this study, we enhanced tumor segmentation by integrating the functionalities of CAM and SAM, proposing a novel approach that employs the lowest-indexed points identified by CAM as negative points within the SAM framework. This strategy aims to refine SAM's prediction boundaries, leading to more precise segmentation. However, the effectiveness of this method is currently constrained by several intrinsic limitations. These practical limitations also arise from our method's reliance on image color attributes to determine negative points, potentially reducing its applicability across different imaging conditions and modalities.

Future research will focus on adapting and enhancing the model to improve its predictive capabilities for diverse anatomical structures, including nerves and blood vessels. This direction aims to expand the clinical applicability of the model and increase its utility in improving diagnostic precision. Perhaps conducting generalization validation on TCGA would be a good option. Meanwhile, we are also considering combining features other than color or adopting dyeing normalization techniques to improve or even replace our point selection scheme, which might enable it to be applied to other image processing problems. For example, when facing the re-recognition of occluded people, the APS module can be replaced with THCB-Net [32]; when dealing with the task of classifying social networks, it can be changed to TSNet [33]. Furthermore, since ViT-H is a very large model, we will consider optimizing the SAM module in the future (for example, exploring a lightweight module like Attention GhostUNet++ [34] or fabSAM [35]), while exploring the attention mechanism to balance the performance and efficiency of the model, in order to enhance the model's robustness and reduce computational overhead.

In terms of the completeness of the experiment, due to the limited time, we only used the t-test, which was not rigorous enough. Therefore, we can perform rigorous statistical validation like Bootstrap confidence intervals or analyze threshold robustness via ROC curves in future work. In addition, we have not conducted blind evaluations by pathologists. Therefore, in the future, we may invite experts to assess the segmentation quality and practicality to verify the value of 2AM in clinical settings. By addressing these specific challenges, we aim to pave the way for the broader adoption and deeper integration of advanced artificial intelligence techniques in medical image analysis. We will consider supplementing relevant ablation experimental analyses in the future and taking dynamic point selection into account.

6. Conclusions

In this paper, we propose an effective weakly supervised framework called **2AM**, which combines Class Activation Mapping (C**AM**) and the Segment Anything Model (S**AM**) for tumor tissue segmentation on pathological images. Our 2AM model comprises three modules: CAM, adaptive point selection (APS), and SAM. The CAM module is adapted to generate salient regions for tumor tissues in pathological images. The APS module is proposed to provide more reliable initial prompts for the subsequent SAM by incorporating three priors: basic appearance, space distribution, and feature difference. The SAM module is utilized to predict the final segmentation. Our experimental results, based on two independent datasets, demonstrate that the proposed method enhances tumor segmentation accuracy by nearly 25% compared with the baseline method and achieves over 15% improvement compared with previous state-of-the-art segmentation methods in a weakly supervised manner. Moreover, our method can be extended to other partially supervised pathological image analyses, such as tissue or nuclei segmentation.

Of course, 2AM also has the following limitations: (1) the computational complexity of the SAM module is relatively high, which limits its application in resource-constrained

environments; (2) the randomness of negative point selection may lead to unstable segmentation; and (3) the current methods mainly target pathological images, and their cross-domain applicability needs further verification. In response to these limitations, we have some future improvement directions, including exploring lightweight SAM models, developing more robust negative point selection strategies, and verifying the performance of 2AM in other image tasks. We hope that these changes will make 2AM more powerful and versatile.

Author Contributions: Methodology: C.R.; Validation: C.R.; Formal analysis: C.R.; Writing—original draft: C.R.; Writing—review and editing: L.Z. and L.G.; Visualization: L.Z.; Supervision: L.G. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: All subjects provided informed consent for inclusion before participating in this study. Ethics approval is not required for this type of study. This study has been granted exemption by the Ethics Committees of Nanjing Drum Tower Hospital and Jiangsu Province Hospital of Chinese Medicine.

Data Availability Statement: Due to hospital ethical restrictions, NDTH and JHCM datasets cannot be disclosed.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kiemen, A.; Braxton, A.M.; Grahn, M.P.; Han, K.S.; Babu, J.M.; Reichel, R.; Amoa, F.; Hong, S.-M.; Cornish, T.C.; Thompson, E.D.; et al. In situ characterization of the 3d microanatomy of the pancreas and pancreatic cancer at single cell resolution. *bioRxiv* **2020**. [CrossRef]
2. Jiao, Y.; Li, J.; Qian, C.; Fei, S. Deep learning-based tumor microenvironment analysis in colon adenocarcinoma histopathological whole-slide images. *Comput. Methods Programs Biomed.* **2021**, *204*, 106047. [CrossRef]
3. Li, Y.; Ping, W. Cancer metastasis detection with neural conditional random field. *arXiv* **2018**, arXiv:1806.07064. [CrossRef]
4. Liu, Y.; Gadepalli, K.; Norouzi, M.; Dahl, G.E.; Kohlberger, T.; Boyko, A.; Venugopalan, S.; Timofeev, A.; Nelson, P.Q.; Corrado, G.S.; et al. Detecting cancer metastases on gigapixel pathology images. *arXiv* **2017**, arXiv:1703.02442. [CrossRef]
5. Madabhushi, A.; Lee, G. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Med. Image Anal.* **2016**, *33*, 170–175. [CrossRef] [PubMed]
6. Zou, L.; Cai, Z.; Mao, L.; Nie, Z.; Qiu, Y.; Yang, X. Automated peripancreatic vessel segmentation and labeling based on iterative trunk growth and weakly supervised mechanism. *Artif. Intell. Med.* **2024**, *150*, 102825. [CrossRef]
7. Randell, R.; Ruddle, R.A.; Thomas, R.; Treanor, D. Diagnosis at the microscope: A workplace study of histopathology. *Cogn. Technol. Work* **2012**, *14*, 319–335. [CrossRef]
8. Wang, B.; Zou, L.; Chen, J.; Cao, Y.; Cai, Z.; Qiu, Y.; Mao, L.; Wang, Z.; Chen, J.; Gui, L.; et al. A weakly supervised segmentation network embedding cross-scale attention guidance and noise-sensitive constraint for detecting tertiary lymphoid structures of pancreatic tumors. *J. Biomed. Health Inform. (J-BHI)* **2024**, *28*, 12. [CrossRef]
9. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
10. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 4015–4026.
11. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
12. Isensee, F.; Petersen, J.; Klein, A.; Zimmerer, D.; Jaeger, P.F.; Kohl, S.; Wasserthal, J.; Koehler, G.; Norajitra, T.; Wirkert, S.; et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv* **2018**, arXiv:1809.10486.

13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.

14. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.

15. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer Nature Switzerland: Cham, Switzerland, 2022; pp. 205–218.

16. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.

17. Lou, W.; Li, H.; Li, G.; Han, X.; Wan, X. Which pixel to annotate: A label-efficient nuclei segmentation framework. *IEEE Trans. Med Imaging* **2022**, *42*, 947–958. [CrossRef] [PubMed]

18. Chen, Z.; Agarwal, D.; Aggarwal, K.; Safta, W.; Balan, M.M.; Brown, K. Masked image modeling advances 3d medical image analysis. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 1970–1980.

19. Li, W.; Li, J.; Polson, J.; Wang, Z.; Speier, W.; Arnold, C. High resolution histopathology image generation and segmentation through adversarial training. *Med. Image Anal.* **2022**, *75*, 102251. [CrossRef]

20. Chan, L.; Hosseini, M.S.; Rowsell, C.; Plataniotis, K.N.; Damaskinos, S. Histosegnet: Semantic segmentation of histological tissue type in whole slide images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10662–10671.

21. Wang, Y.; Zhang, J.; Kan, M.; Shan, S.; Chen, X. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 12275–12284.

22. Xu, L.; Ouyang, W.; Bennamoun, M.; Boussaid, F.; Xu, D. Multi-class token transformer for weakly supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4310–4319.

23. Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; Wang, B. Segment anything in medical images. *Nat. Commun.* **2024**, *15*, 654. [CrossRef]

24. Cui, C.; Deng, R.; Liu, Q.; Yao, T.; Bao, S.; Remedios, L.W.; Landman, B.A.; Tang, Y.; Huo, Y. All-in-sam: From weak annotation to pixel-wise nuclei segmentation with prompt-based finetuning. *J. Phys. Conf. Ser.* **2024**, 2722, 012012. [CrossRef]

25. Ke, L.; Ye, M.; Danelljan, M.; Tai, Y.-W.; Tang, C.-K.; Yu, F. Segment anything in high quality. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 29914–29934.

26. Zhang, J.; Ma, K.; Kapse, S.; Saltz, J.; Vakalopoulou, M.; Prasanna, P.; Samaras, D. Sam-path: A segment anything model for semantic segmentation in digital pathology. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Vancouver, BC, Canada, 8–12 October 2023; Springer Nature Switzerland: Cham, Switzerland, 2023; pp. 161–170.

27. Nateghi, R.; Pourakpour, F. Perineural invasion detection in multiple organ cancer based on deep convolutional neural network. *arXiv* **2021**, arXiv:2110.12283. [CrossRef]

28. Zou, L.; Cai, Z.; Qiu, Y.; Gui, L.; Mao, L.; Yang, X. Ctg-net: An efficient cascaded framework driven by terminal guidance mechanism for dilated pancreatic duct segmentation. *Phys. Med. Biol.* **2023**, *68*, 215006. [CrossRef]

29. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.

30. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.

31. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 839–847.

32. Wang, C.; He, S.; Wu, M.; Lam, S.-K.; Tiwari, P.; Gao, X. Looking clearer with text: A hierarchical context blending network for occluded person re-identification. *IEEE Trans. Inf. Forensics and Secur.* **2025**, *20*, 4296–4307. [CrossRef]

33. Wang, C.; Cao, R.; Wang, R. Learning discriminative topological structure information representation for 2d shape and social network classification via persistent homology. *Knowl.-Based Syst.* **2025**, *311*, 113125. [CrossRef]

34. Hayat, M.; Aramvith, S.; Bhattacharjee, S.; Ahmad, N. Attention ghostunet++: Enhanced segmentation of adipose tissue and liver in ct images. *arXiv* **2025**, arXiv:2504.11491.

35. Xie, Y.; Wu, H.; Tong, H.; Xiao, L.; Zhou, W.; Li, L.; Wanger, T.C. fabsam: A farmland boundary delineation method based on the segment anything model. *arXiv* **2025**, arXiv:2501.12487.