Bridging Visual-Textual Modalities: Weakly Supervised Histopathology Segmentation

Sicong Gao[1†]
sicong.gao@student.unsw.edu.au
Matthew AB Baker[2]
matthew.baker@unsw.edu.au
Maurice Pagnucco[1]
morri@unsw.edu.au
Zhiwei Gao[3]
gaozw@stdu.edu.cn
Yang Song[1]
yang.song1@unsw.edu.au

[1] School of Computer Science and Engineering, University of New South Wales, Australia
[2] School of Biotechnology and Biomolecular Sciences, University of New South Wales, Australia
[3] Hebei Key Laboratory of Electromagnetic Environmental Effects and Information Processing, China

Abstract

In weakly supervised histopathology tissue segmentation, Class Activation Maps (CAMs) are commonly used to generate pseudo-masks. However, CAMs typically highlight only the most discriminative regions, leading to inaccurate tissue boundaries. Existing visual-based refinement strategies often exacerbate information loss, while text-based methods suffer from high inter-class similarity and a semantic gap between pixel-level features and text labels. To overcome these limitations, we propose a two-stage segmentation framework that jointly models a bidirectional shared latent space between visual and textual modalities to enhance pseudo-mask quality. In the segmentation (second) stage, we incorporate complex tissue textual descriptions as external discriminative knowledge to compensate for insufficient supervision. We further develop a multi-stage modality fusion strategy based on learnable query tokens and Fourier transforms. Experiments conducted on the LUAD-HistoSeg and BCSS-WSSS datasets demonstrate that our method surpasses state-of-the-art weakly supervised tissue segmentation approaches.

1 Introduction

In histopathology, tissue segmentation accelerates the identification of pathological regions and enables the extraction of key quantitative information, such as tumor size, density, and microenvironment characteristics [23, 39, 42]. However, existing segmentation methods heavily rely on pixel-level annotations, which are time-consuming and require substantial domain expertise. To alleviate the annotation burden, recent studies have explored weakly supervised semantic segmentation (WSSS) approaches. Common WSSS strategies include scribble annotations [14], image-level labels [25], point-based supervision [45], and bounding box annotations [22]. In tissue segmentation, image-level labels are relatively easier

† denotes corresponding author.

to obtain hence more practical. Most existing methods generate pseudo-masks from class activation maps (CAMs) derived from classification networks [48] and refine these masks through fully supervised training [11, 13, 46].

However, CAMs usually highlight only the most discriminative regions [1], resulting in inaccurate pseudo-mask boundaries. This issue is especially pronounced in histopathology, where subtle morphological differences and complex spatial distributions among tissues often cause partial or fragmented predictions [25]. In the natural image domain, early visual-based methods include prototype-based boundary refinement [7] and clustering-enhanced representation learning [5]. Recently, vision-language models (e.g., CLIP [37]) have been incorporated to generate pseudo-masks by introducing additional textual information [15, 29, 31], thereby introducing a new perspective.

While these methods have shown promising results in natural images, their application to tissue segmentation is still constrained by the **high inter-class similarity and intra-class variability** of tissues [8, 16], as illustrated in Figure 2. Specifically, visual-based approaches are susceptible to activating irrelevant regions with high inter-class similarity, while eliminating these activations would lead to additional information loss [13, 17], especially for tissue regions with high intra-class variability. To address this, text prompts were introduced as an auxiliary modality to complement visual signals [19, 20]. However, these methods rely heavily on pixel-text similarity, ignoring the semantic gap between image-level and pixel-level representations and high inter-class tissue similarity.

To address these issues, we propose a bidirectional pseudo-mask generation strategy that comprehensively leverages visual and textual modalities. Specifically, to reduce inter-class similarity, we enhance visual pseudo-mask generation by reactivating underutilized channels in the visual backbone. These reactivated channels encourage the network to attend to subtle, discriminative features between different tissue types, thereby facilitating the differentiation of visually similar categories. Meanwhile, to reduce intra-class variability, we construct text-driven pseudo-masks by computing pixel-text similarity guided by descriptive textual prompts. These prompts encode global semantic knowledge of tissue categories and provide consistent supervision, enabling the alignment of diverse visual instances within the same class to a shared semantic space. Unlike previous methods that use unidirectional text-to-image alignment [44, 46], our bidirectional supervision mechanism enables mutual refinement: visual pseudo-masks provide spatial priors to constrain noisy text-driven predictions, while textual semantics regularize visual activations by enforcing class-consistent responses across spatially diverse regions. This complementary design enhances the completeness and semantic fidelity of pseudo-masks under sparse supervision.

In addition, in weakly supervised settings, noisy annotations in pseudo-masks remain a major challenge. As shown in Figure 1, textual knowledge (e.g., morphology, color, interrelations) can serve as external guidance for refining pseudo-masks. However, existing methods on natural images often align modalities using cross-attention [9, 43, 44] and struggle to extract informative cues from long, complex tissue descriptions, due to the lack of an effective mechanism for selecting discriminative knowledge. In contrast, we propose a query-based knowledge extractor that employs learnable query tokens to interact with textual embeddings to enhance pixel-level discrimination. For the visual modality, tissues often exhibit subtle differences, making it difficult to generate robust representations using standard spatial-domain features. Notably, the spatial structure and semantic layout of histopathological images are better preserved in the low-frequency components of the visual spectrum [6, 27, 32, 36]. Thus, we introduce a Fourier-based visual extractor, which employs learnable filters in the Fourier domain to extract coarse-to-fine contextual cues that enhance visual discrimination. By integrating discriminative textual knowledge with frequency-domain visual semantics through a cross-attention mechanism, our method effectively refines the segmentation results. Overall, our main contributions are summarized as follows:

- To address incomplete and noisy pseudo-masks, we design a modality-complementary strategy that activates overlooked visual regions and leverages textual prompts for semantic enhancement.
- We introduce a query-fourier knowledge fusion module that combines learnable query tokens for selecting discriminative textual semantics with frequency-domain visual filters to capture structural cues, enabling semantically multimodal representations.
- Our method achieves state-of-the-art performance on LUAD-HistoSeg [17] and BCSS-WSSS [2] benchmarks, demonstrating its effectiveness under weak supervision.

2 Related Work

Weakly supervised tissue segmentation based on image labels primarily involves generating pseudo-masks using CAM, followed by refinement. HistoSegNet [4] used CRF to find more accurate tissue boundaries. OEEM [26] focused on enhancing the network's attention to reliable supervision signals. MLPS [17] enforced the consideration of global information through a progressive dropout approach. HAMIL [47] refines pseudo labels by leveraging interleaved learning among three networks with identical architectures. PCSformer [30] proposed a conflict self-elimination strategy to improve this process. HisynSeg [11] presented a data synthesis method using mosaic transformations. However, they still face the fundamental issue of insufficient information by using only image labels. The main challenge is the scattered distribution of tissue structures, the similarities between classes, and variations within classes, which are difficult to capture accurately with only image labels.

MTPTN [18] introduced progressive text-based prior prompts to generate multimodal features. LViT [28] used a language-visual loss function to supervise the training of unlabeled images with textual information. These studies have been successful in segmenting cell nuclei from the background, but the lack of consideration for multi-class segmentation has limited their methods' application in tissue segmentation. TPRO [46] attempts to apply the capabilities of multimodal pre-trained models to pseudo-mask generation by calculating the similarity between simple text features and pixel features. However, this method overlooks feature gaps and inter-class similarity issues. In contrast, our approach aligns modalities using auxiliary loss within a unified framework.

3 Methodology

This section introduces our two-stage multimodal framework for weakly supervised tissue segmentation (Figure 1). The first stage generates pseudo-masks (§ 3.1). The second stage focuses on utilizing the textual knowledge to obtain the final segmentation result (§ 3.2).

3.1 Bidirectional Pseudo Mask Generation

In the first stage, given an image $I \in \mathbb{R}^{3\times H\times W}$, visual features $F_v \in \mathbb{R}^{C_o\times H_o\times W_o}$ are extracted using the visual encoder ϕ_v, by concatenating features from its last three layers. Subsequently, F_v is processed through two parallel pipelines. The first pipeline reactivates channel

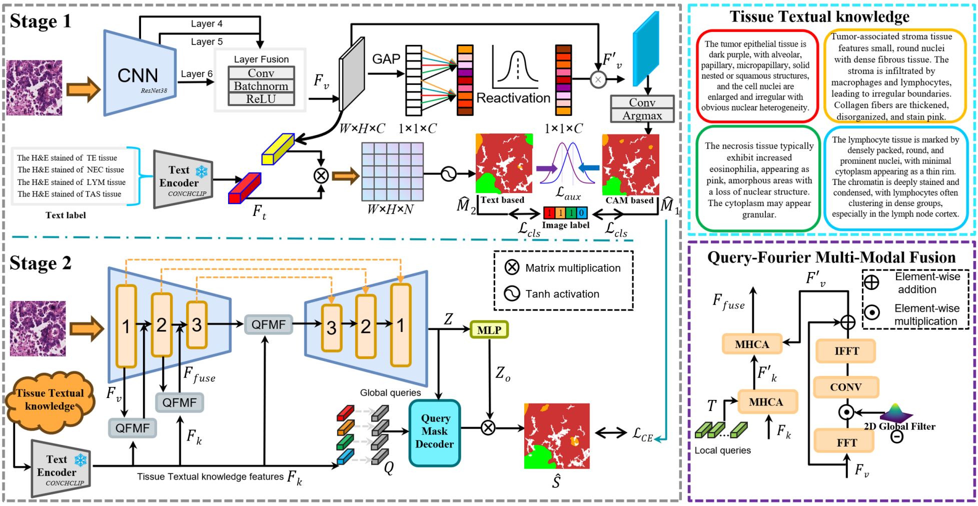


Figure 1: An overview of ours proposed method, In the first stage, two parallel pipelines generate pseudo masks $\hat{M}_1$ (via channel attention reactivation) and $\hat{M}_2$ (via text label). We then align these representations to obtain the final pseudo mask. In the second stage, a Query Fourier-based Multi-Modal Fusion (QFMF) module fuses textual knowledge features F_k with visual features F_v to produce a combined representation F_{fuse}. Finally, learnable query tokens Q in the Query Mask Decoder module yield the final mask $\hat{S}$.

attention (§ 3.1.1) to generate a vision-based pseudo-mask $\hat{M}_1$, while the second employs a text encoder ϕ_t to extract textual features $F_t \in \mathbb{R}^{C_t \times N}$ from text labels, where N represents the number of tissue classes, and C_t is the dimension of the text embeddings. This is used alongside the similarity between F_t and F_v to produce a text-based pseudo-mask $\hat{M}_2$ (§ 3.1.2). Then, Manhattan distance (§ 3.1.3) is employed to minimize the disparity between $\hat{M}_1$ and $\hat{M}_2$. Training in this stage is driven by tissue type classification using two pseudo-masks with image-level labels and the reduction of modal distances.

3.1.1 Channel Attention Reactivation

To enable the network to attend beyond only the most discriminative regions, we propose a lightweight channel attention mechanism to improve learning efficiency. Our objective is to assess the significance of each channel in generating pseudo-masks before the final classification layer and adjust their activation based on probabilistic contributions. Specifically, we employ global average pooling g to extract global features from each channel. A one-dimensional convolution δ then transforms these features into channel weights: $\omega = \delta(g(F_v)) \in \mathbb{R}^{1\times 1\times C_o}$, where C_o is the number of channels. This maps global features into channel-wise weights that control activation strength. To further refine activation, we introduce a probabilistic adjustment factor p, which is dynamically adjusted during training:

$$p_i = 1 - \frac{e^{\omega_i}}{\sum_{j=1}^{C_o} e^{\omega_j}}, \tag{1}$$

where, ω_i represents the weight of the i-th channel. The adjustment factor p_i, as a supplementary probability, prevents any channels from excessively dominating the model's response and activates underutilized channels. The new channel weights ω' are calculated as follows: $\omega' = \{\omega'_i \mid \omega'_i = p_i \cdot \omega_i, i \in \{1,2,\ldots,c_o\}\}$. Refined weights ω' enable more precise regulation of information flow, thereby facilitating the differentiation of visually similar tissue categories and further reducing inter-class similarity.

We then compute the reactivated features as $F'_v = \sigma(\omega' F_v)$, where σ denotes the sigmoid function. Next, we extract the classification weights $W_{cls} \in \mathbb{R}^{C_o \times N}$ from the final classification layer. To generate vision-based pseudo mask $\hat{M}_1$, we project the reactivated features onto the semantic classifier space, capturing the correlations between feature channels and N tissue class prototypes:

$$\hat{M}_1 = Up\left(\sum_{c=1}^{C_o} F'_{v,c} \otimes W_{cls}^{(c,:)}\right) \in \mathbb{R}^{N \times H \times W}, \tag{2}$$

where $F'_{v,c} \in \mathbb{R}^{H_o \times W_o}$ denotes the c-th channel of reactivated feature map, $\otimes$ denotes outer product, and $Up(\cdot)$ is the bilinear upsampling to restore $\hat{M}_1$ to the original image size.

3.1.2 Pixel-wise Text Supervised Module

To reduce intra-class variability and compensate for visual modalities, we incorporate class-level textual cues. Visual and textual features are projected into a shared embedding space, and the text-guided similarity between the i-th spatial location and the j-th class text-label embedding is computed as:

$$\hat{M}_{2(i,j)} = Tanh\left(\sum_{c=1}^{C_o} (W_1 F_v)_{i,c} \cdot (W_2 F_t)_{j,c}\right) \in \mathbb{R}^{1 \times 1}, \tag{3}$$

where W_1, W_2 are learnable projections that align F_v, F_t to the same dimension, and c is the channel index. A higher value of $\hat{M}_{2(i,j)}$ indicates a stronger semantic relationship between the i-th pixel and the j-th class. Based on this similarity, we construct the text-based pseudo mask $\hat{M}_2 \in \mathbb{R}^{N \times H \times W}$, which highlights regions most relevant to each tissue class.

3.1.3 Pseudo Mask Loss Design

This task is based on the multilabel soft margin loss $\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=0}^{N-1} y_i \log(\frac{1}{1+e^{-\hat{y}_i}}) + (1 - y_i)\log(\frac{e^{-\hat{y}_i}}{1+e^{-\hat{y}_i}})$, where, y_i and $\hat{y}_i$ are the image labels and predicted logits of category i. We obtain the classifications for the vision-based and the text-based pseudo-mask: $\hat{y}_1 = g(\hat{M}_1)$ and $\hat{y}_2 = g(\hat{M}_2)$ respectively, where g denotes the global average pooling. The overall classification loss is expressed as: $\mathcal{L}_c = \frac{1}{2}(\mathcal{L}_{cls}(\hat{y}_1) + \mathcal{L}_{cls}(\hat{y}_2))$. To ensure that $\hat{M}_1$ can effectively learn additional textual information and to maintain consistency between text and visual modalities while avoiding deviations in feature representation from image dimensions, we introduce an auxiliary loss $\mathcal{L}_{aux} = \|\hat{M}_1 - \hat{M}_2\|_1$. This approach aids in synchronously optimizing the model's handling of both modalities, ensuring that textual and visual information are effectively integrated. Overall, the Stage 1 loss $\mathcal{L}_{S1}$ is: $\mathcal{L}_{S1} = \mathcal{L}_c + \mathcal{L}_{aux}$.

3.2 Knowledge-enhanced Segmentation

In this stage, we utilize $\hat{M}_1$ as the supervisory signal for segmentation. However, $\hat{M}_1$ contains errors, due to the reliance solely on image labels for discrimination. To address this, we propose a Query-Fourier based Multi-Modal Fusion (QFMF) Module, which integrates class-level textual knowledge, enabling the network to learn more discriminative information from text. Specifically, we employ a vision encoder-decoder architecture comprising four encoder blocks. Each block's features are represented as F_v^i, where $i \in \{1,2,3,4\}$. We utilize a text encoder ϕ_t to extract knowledge features F_k. During encoder stage, we integrate a QFMF module (§ 3.2.1) to fuse F_v^i and F_k, yielding fused features F_{fuse}^i, which are then sent to the next block. Finally, the decoder produces the ultimate visual feature Z_o and, through the Query Mask Decoder (§ 3.2.2), obtains the final segmentation output $\hat{S}$.

3.2.1 Query-Fourier based Multi-Modal Fusion

Figure 1 illustrates the structure of the QFMF module, which consists of a Query-based textual knowledge Extractor (QKE) and a Fourier-based Visual Extractor (FVE). Initially, to extract discriminative information from textual tissues knowledge $F_k \in \mathbb{R}^{C_k \times N}$, where N and C_k represent the number of classes and the dimension of knowledge embeddings. We design a learnable query token $T \in \mathbb{R}^{L_t \times C_k}$. Here, L_t denotes the number of tokens. Next, we employ multi-head cross-attention (**MHCA**) to extract discriminative textual representations:

$$F_k' = \text{Softmax}\left(\frac{(F_k W^Q)(TW^K)^\top}{\sqrt{d}}\right)(TW^V), \tag{4}$$

The global statistical properties revealed by low-frequency components extracted via Fast Fourier Transform (FFT) [3] have been leveraged in various vision tasks, including image inpainting [10], domain generalization [24], and image super-resolution [6]. We apply the FFT transform to the input spatial domain features $F_v \in \mathbb{R}^{H \times W}$, transforming them into the corresponding frequency domain $\mathcal{F} \in \mathbb{C}^{H \times W}$. Specifically, the $\mathcal{F}$ can be described as:

$$\mathcal{F}(m,n) = \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} F_v(x,y) e^{-j2\pi(\frac{mx}{W} + \frac{ny}{H})}, \tag{5}$$

where, $F_v(x,y)$ is defined as the pixel (x,y) in the spatial domain. and $\mathcal{F}(m,n)$ denotes in the corresponding frequency domain. We employ a learnable global filter Θ for low-frequency filtering. The Θ_{init} using a 2D Gaussian filter:

$$\Theta \leftarrow \alpha \cdot \Theta_{init}(u,v); \quad \Theta_{init}(u,v) = \exp\left(-\frac{(u-u_0)^2 + (v-v_0)^2}{2\sigma^2}\right). \tag{6}$$

where, u_0 and v_0 denote the Gaussian center, σ controls the filter's spread set to 15, and α is a learnable parameter. This design enables Θ to adaptively refine frequency filtering, enhancing relevant components for global visual representation. Combined with textual cues, it helps correct pseudo-mask errors. The enhanced visual features F_v' are obtained via Θ and Fourier convolution:

$$F_{\text{fuse}} = \text{Softmax}\left(\frac{(F_v' W'^Q)(F_k' W'^K)^\top}{\sqrt{d}}\right)(F_k' W'^V); \quad F_v' = F_v + \mathcal{F}^{-1}(m,n)\left(\text{Conv}\left(\Theta \odot \mathcal{F}(m,n)(F_v)\right)\right). \tag{7}$$

3.2.2 Query Mask Decoder

We obtain the final visual feature $Z_o \in \mathbb{R}^{C_z \times H_z \times W_z}$. To enable category-specific querying, we introduce a set of learnable class-level query tokens $Q \in \mathbb{R}^{N \times C_q}$, where N is the number of categories and C_q denotes the dimension of the textual knowledge features. Each token is assigned to attend to a specific class. We initialize the query tokens with textual knowledge

Dataset		LUAD-HistoSeg [17]					BCSS-WSSS [2]				
Method	Modal	TE	NEC	LYM	TAS	mIoU	TUM	STR	LYM	NEC	mIoU
Grad-CAM [38]	Vision	70.07	66.01	70.18	64.76	67.76	65.96	56.71	43.36	30.04	49.02
SC-CAM [5]	Vision	68.29	64.28	62.07	61.79	64.10	64.28	56.16	42.87	30.14	48.36
MLPS [17]	Vision	75.68	75.59	74.01	69.97	73.81	72.98	<u>68.13</u>	56.19	55.99	63.32
OEEM [26]	Vision	69.66	72.62	72.58	66.88	70.44	66.83	58.71	49.41	51.12	56.52
HAMIL [47]	Vision	72.82	69.79	69.82	70.96	70.85	69.84	59.45	49.98	51.29	57.64
PCSformer [30]	Vision	71.82	73.55	75.31	68.98	72.42	74.54	67.36	55.29	61.47	64.67
HisynSeg [12]	Vision	<u>75.67</u>	75.83	75.10	<u>71.94</u>	74.64	**78.73**	67.41	<u>59.47</u>	<u>65.31</u>	<u>67.73</u>
CLIP-ES [29]	Vision-Language	70.12	69.34	68.25	66.58	68.57	65.83	61.29	51.32	55.27	58.43
TPRO [46]	Vision-Language	74.82	**77.55**	<u>76.40</u>	70.98	<u>74.94</u>	77.18	63.77	54.95	61.43	64.33
Ours	Vision-Language	**76.32**	<u>76.50</u>	**77.20**	**76.78**	**76.70**	<u>78.14</u>	**68.32**	**63.24**	**67.77**	**69.37**

Table 1: Comparison of the pseudo mask results.

F_k, allowing them to capture category-specific features guided by semantic context. This enhances their ability to learn in a more accurate and text-aware manner, formulated as:

$$\hat{S} = \mathbf{MHCA}(\beta(Q + F_k), Z) \otimes Z_o, \tag{8}$$

where β is a learnable parameter. This operation yields class-specific feature representations $\hat{S} \in \mathbb{R}^{N \times H \times W}$, driving the final segmentation output. We use BCE loss for optimization. $\mathcal{L}_{BCE} = -\frac{1}{P}\sum_{i=1}^{P}[S_i \log(\hat{S}_i) + (1 - S_i)\log(1 - \hat{S}_i)]$. Here, P is the number of pixels, S_i the pseudo label from $\hat{M}_1$, and $\hat{S}_i$ the predicted foreground probability for the i-th pixel.

Dataset		LUAD-HistoSeg [17]					BCSS-WSSS [2]				
Method	Modal	TE	NEC	LYM	TAS	mIoU	TUM	STR	LYM	NEC	mIoU
HistoSegNet [1]	Vision	45.59	36.30	58.28	50.82	47.75	33.14	46.46	29.05	1.91	27.64
MLPS [17]	Vision	77.70	79.32	73.41	71.98	75.60	78.84	73.16	57.30	66.39	68.92
OEEM [26]	Vision	73.81	70.49	71.89	69.48	71.42	74.86	64.68	48.91	61.03	62.37
HAMIL [47]	Vision	73.46	75.83	72.94	70.86	73.27	71.65	62.37	51.52	54.29	59.96
PCSformer [30]	Vision	72.41	74.66	75.39	70.60	73.27	74.92	69.58	59.71	62.21	66.61
HisynSeg [12]	Vision	<u>78.41</u>	77.67	76.39	<u>73.05</u>	76.38	**81.67**	<u>73.85</u>	<u>60.74</u>	<u>66.82</u>	<u>70.77</u>
LAVT [44]	Vision-Language	71.36	68.74	68.19	67.28	68.89	67.72	65.21	50.99	59.83	60.94
Bridger [43]	Vision-Language	69.79	72.33	70.76	68.49	70.34	71.25	62.14	51.73	60.82	61.49
TPRO [46]	Vision-Language	75.80	**80.56**	<u>78.14</u>	72.69	<u>76.80</u>	77.95	65.10	54.55	64.96	65.64
Ours	Vision-Language	**79.21**	<u>79.34</u>	**80.22**	**77.30**	**79.02**	<u>81.41</u>	**74.22**	**65.21**	**69.24**	**72.52**

Table 2: Comparison of the final segmentation results.

4 Experiments

4.1 Datasets and Implementation Details

Two weakly-supervised histopathology datasets are involved in our study. **LUAD-HistoSeg** [17] includes four lung tissue classes (TE, TAS, NEC, LYM) with 16,678 training, 300 validation, and 307 test patches. **BCSS-WSSS** [2] covers four breast tissue classes (TUM, STR, LYM, NEC) with 23,422 training, 3,418 validation, and 4,986 test patches. Both datasets provide image-level for training and pixel-level annotations for validation and testing. Tissue textual knowledge is collected from medical literature (details in supplementary materials).

In the first stage, we use ResNet38d [41] as the backbone for its strong localization ability, essential for generating reliable pseudo masks. It is trained for 25 epochs using SGD (initial LR = 0.01, polynomial decay) with a batch size of 20. In the second stage, we adopt Swin-B [33] to better capture long-range dependencies and align with textual knowledge for multimodal fusion. It is trained for 16 epochs using AdamW [34] (initial LR = 0.0001) with a batch size of 64. We use CONCHCLIP [35] to extract text features, as it is tailored for histopathology. IoU is used for evaluation, and all experiments were conducted on an NVIDIA RTX A6000 GPU.

4.2 Comparison with State-of-the-Art

Quality of Pseudo Masks In stage 1, we compare seven methods for the quality of pseudo-masks. In Table 1, our method exceeds the second-best method by 1.76% on LUAD-HistoSeg and by 1.64% on BCSS-WSSS. Grad-CAM and SC-CAM target natural images, while the others specifically address histopathology. On BCSS-WSSS, HisynSeg surpasses our method by 0.59% in TUM by synthesizing pseudo-masks to offset its data scarcity. However, such synthetic morphologies deviate from authentic images, leading to inferior performance in other classes. For multimodal comparisons, CLIP-ES leverages zero-shot CLIP for pseudo-mask generation on natural images. However, subtle tissue differences in histopathology hinder CLIP from relying solely on textual descriptions. TPRO, another vision–language approach, surpasses our method by 1.05% in NEC on LUAD-HistoSeg, likely benefiting from its multi-layer pseudo-mask strategy. Nonetheless, its performance on other classes declines, possibly due to residual noise across layers that remains unmitigated.

Final Segmentation Performance In stage 2, we employ six methods designed for histopathology and two multimodal segmentation methods designed for general images as benchmarks. In Table 2, our method outperforms the second-best method by 2.22% on LUAD-HistoSeg and by 1.75% on BCSS-WSSS. Due to the same reason highlighted in *Quality of Pseudo Masks*, our method underperforms HisynSeg by 0.26% in TUM and TPRO by 1.22% in NEC. However, overall, our method is the best, highlighting the effectiveness of our multi-modal approach. In addition, multimodal segmentation models such as LAVT and Bridger that focus on natural images cannot be well transferred to tissue segmentation tasks.

4.3 Ablation Studies and Analysis

Comparison of Pseudo Masks In Table 3(a), we conduct a comparison of various pseudo masks. $\hat{M}_1$ and $\hat{M}_2$ are derived through L_1 distance for consistency. The results indicate that our auxiliary loss facilitates learning between modalities, with Vision-based to $\hat{M}_1$ showing an improvement of 1.11%, and Text-based to $\hat{M}_2$ showing an enhancement of 1.55%. Notably, $\hat{M}_{1C}$ demonstrates the best performance. In addition, Table 3(b) shows ablation studies for the MLF and CAR modules affirm each component's efficacy.

Pseudo Mask	TE	NEC	LYM	TAS	mIoU
Vision-based	74.92	76.36	**77.41**	73.67	75.59
Text-based	72.57	76.21	74.09	71.41	73.57
$\hat{M}_1$	**76.32**	**76.50**	77.20	**76.78**	**76.70**
$\hat{M}_2$	73.97	75.34	76.89	74.27	75.12

(a) Ablation on different Pseudo Mask.

MLF	CAR	TE	NEC	LYM	TAS	mIoU
✗	✗	75.32	70.21	75.59	69.29	72.60
✓	✗	75.68	71.39	75.75	71.11	73.48
✗	✓	75.31	74.37	76.24	73.46	74.85
✓	✓	**76.32**	**76.50**	**77.20**	**76.78**	**76.70**

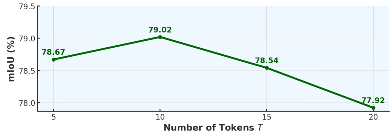
(b) Ablation on MLF and CAR in stage 1.

Table 3: Ablation study for (a) different Pseudo Masks and (b) the effects of multi-layer fusion (MLF) and Channel Attention Reactivation (CAR) in Stage 1.

Ablation on Final Segmentation In Table 4(a), the QFMF module comprises the QKE and the FVE. The experiments indicate that QKE effectively extracts discriminative information from tissue textual knowledge. FVE successfully captures global semantic information from images through low-pass Fourier convolution. Overall, our QFMF, resulted in a performance improvement of 1.94%. Additionally, the query-based mask decoder could identify class-specific features on the image, contributing to a further increase of 1.01% in performance.

QKE	FVE	QMD	TE	NEC	LYM	TAS	mIoU
✗	✗	✗	73.52	76.34	78.53	75.89	76.07
✗	✓	✗	76.41	77.47	78.15	76.57	77.15
✓	✓	✗	77.27	77.94	79.64	77.19	78.01
✓	✓	✓	**79.21**	**79.34**	**80.22**	**77.30**	**79.02**

(a) Ablation for QKE, FVE, and QMD in stage 2.



(b) Ablation on the Query T.

Table 4: (a) Ablation study for Query-based Textual Knowledge Extractor (QKE), Fourier-based Visual Extractor (FVE), and Query Mask Decoder (QMD) in stage 2. (b) Ablation on the number of tokens for Query T in QFMF module.

Table 4(b) shows that using 10 learnable query tokens in the QFMF module yields the best performance, enabling T to capture relevant knowledge most effectively.

Analysis of Different Prompts During the segmentation stage, Table 5(a) compares four text encoders: MedCLIP, PLIP, QuiltNet and CONCHCLIP. Except for MedCLIP, all others are pre-trained based on histopathology. The results show that CONCHCLIP achieves the best performance while MedCLIP performs the worst, underscoring the value of external pretraining knowledge in histopathology. In Table 5(b), we further contrast different forms of text inputs and demonstrate that textual knowledge conveys discriminative information that text labels cannot. Refer to the supplementary material for additional ablation studies.

Text encoder	TE	NEC	LYM	TAS	mIoU
MedCLIP [40]	75.24	77.39	75.27	76.77	76.17
PLIP [20]	75.27	78.14	75.29	76.88	76.40
QuiltNet [21]	**79.24**	**82.49**	74.02	77.22	78.24
CONCHCLIP [35]	79.21	79.34	**80.22**	**77.30**	**79.02**

(a) Ablation on different text encoders.

	TE	NEC	LYM	TAS	mIoU
Text Label	77.23	78.52	79.14	76.97	77.97
Textual Knowledge	**79.21**	**79.34**	**80.22**	**77.30**	**79.02**

(b) Ablation on different text input

Table 5: Ablation study for prompt engineering.

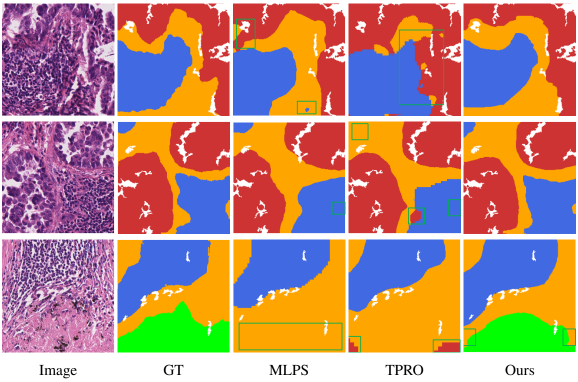


Figure 2: Visualization of segmentation results generated by MLPS, TPRO, and our method. The green boxes highlight regions where the proposed method performs better than the other methods. The tissue types include tumor epithelial (red), necrosis (green), tumor-associated stroma (orange), and lymphocyte (blue).

Qualitative Analysis Figure 2 illustrates our evaluation of two representative baselines: the vision-based MLPS [17] and the text-based TPRO [46]. MLPS generates pseudo masks by eliminating highly discriminative parts to force the network to focus on small areas, usually sacrificing key structural information. TPRO, on the other hand, depends solely on inter-modal similarity and thus struggles to achieve accurate alignment across modalities. By contrast, our method significantly outperforms both, as shown by the comparative results.

5 Conclusions

This paper presents a two-stage weakly supervised tissue segmentation framework that bridges visual and textual modalities, leveraging two types of textual inputs to guide the segmentation process. To address the issue of insufficient supervision in prior methods, the framework integrates channel attention–reactivated visual features with linguistic cues derived from simple text labels, enabling the generation of high-quality pseudo-masks. A query-based text extractor and a visual extractor using low-pass Fourier convolution are employed to capture global semantic information and facilitate complementary cross-modal interactions. Finally, a query-driven mask decoder precisely segments target tissue categories, enhancing the overall effectiveness of the segmentation process.

References

[1] Jiwoon Ahn, Suha Kwak, et al. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4981–4990, 2018.

[2] Mohamed Amgad, Habiba Elfandy, Hagar Hussein, et al. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18):3461–3467, 2019.

[3] E Oran Brigham, RE Morrow, et al. The fast fourier transform. *IEEE Spectrum*, 4(12): 63–70, 1967.

[4] Lyndon Chan, Mahdi S Hosseini, Corwyn Rowsell, et al. Histosegnet: Semantic segmentation of histological tissue type in whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10662–10671, 2019.

[5] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8991–9000, 2020.

[6] Jiadi Chen, Chunjiang Duanmu, Huanhuan Long, et al. Large kernel frequency-enhanced network for efficient single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6317–6326, 2024.

[7] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4288–4298, 2022.

[8] Zhe Chen, Zhao Chen, Jingxin Liu, Qiang Zheng, Yuang Zhu, Yanfei Zuo, Zhaoyu Wang, Xiaosong Guan, Yue Wang, and Yuan Li. Weakly supervised histopathology image segmentation with sparse point annotations. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1673–1685, 2020.

[9] Yong Xien Chng, Henry Zheng, Yizeng Han, Xuchong Qiu, and Gao Huang. Mask grounding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26573–26583, 2024.

[10] Tianyi Chu, Jiafu Chen, Jiakai Sun, Shuobin Lian, Zhizhong Wang, Zhiwen Zuo, Lei Zhao, Wei Xing, and Dongming Lu. Rethinking fast fourier convolution in image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23195–23205, 2023.

[11] Zijie Fang, Yang Chen, Yifeng Wang, et al. Weakly-supervised semantic segmentation for histopathology images based on dataset synthesis and feature consistency constraint. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 606–613, 2023.

[12] Zijie Fang, Yifeng Wang, Peizhang Xie, Zhi Wang, and Yongbing Zhang. HisynSeg: Weakly-supervised histopathological image segmentation via image-mixing synthesis and consistency regularization. *IEEE Transactions on Medical Imaging*, 2024.

[13] Siyang Feng, Jiale Chen, Zhenbing Liu, Wentao Liu, Zimin Wang, Rushi Lan, Xipeng Pan, et al. Mining gold from the sand: Weakly supervised histological tissue segmentation with activation relocalization and mutual learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 414–423. Springer, 2024.

[14] Feng Gao, Minhao Hu, Min-Er Zhong, et al. Segmentation only uses sparse annotations: Unified weakly and semi-supervised learning in medical images. *Medical Image Analysis*, 80:102515, 2022.

[15] Sicong Gao, Matthew AB Baker, Maurice Pagnucco, and Yang Song. Multimodal weakly supervised segmentation for histopathology. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2025.

[16] Ruoyu Guo, Kunzi Xie, Maurice Pagnucco, and Yang Song. Sac-net: Learning with weak and noisy labels in histopathology image segmentation. *Medical Image Analysis*, 86:102790, 2023.

[17] Chu Han, Jiatai Lin, Jinhai Mai, et al. Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. *Medical Image Analysis*, 80:102487, 2022.

[18] Xianjun Han, Qianqian Chen, Zhaoyang Xie, et al. Multiscale progressive text prompt network for medical image segmentation. *Computers & Graphics*, 116:262–274, 2023.

[19] Dingyi Hu, Zhiguo Jiang, Jun Shi, Fengying Xie, Kun Wu, Kunming Tang, Ming Cao, Jianguo Huai, and Yushan Zheng. Histopathology language-image representation learning for fine-grained digital pathology cross-modal retrieval. *Medical Image Analysis*, 95:103163, 2024.

[20] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, et al. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, 29(9): 2307–2316, 2023.

[21] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in Neural Information Processing Systems*, 36:37995–38017, 2023.

[22] Hoel Kervadec, Jose Dolz, Shanshan Wang, et al. Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision. In *Medical Imaging with Deep Learning*, pages 365–381. PMLR, 2020.

[23] Daisuke Komura and Shumpei Ishikawa. Machine Learning Methods for Histopathological Image Analysis. *Computational and Structural Biotechnology Journal*, 16:34–42, 2018. doi: 10.1016/j.csbj.2018.01.001.

[24] Sangrok Lee, Jongseong Bae, Ha Young Kim, et al. Decompose, adjust, compose: Effective normalization by playing with frequency for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11776–11785, 2023.

[25] Kailu Li, Ziniu Qian, Yingnan Han, et al. Weakly supervised histopathology image segmentation with self-attention. *Medical Image Analysis*, 86:102791, 2023.

[26] Yi Li, Yiduo Yu, Yiwen Zou, et al. Online easy example mining for weakly-supervised gland segmentation from histology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 578–587. Springer, 2022.

[27] Yingfan Li, Huaiji Zhou, Na Liu, and Yiqing Shen. Stain normalization and augmentation in frequency space for histology analysis. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2031–2035. IEEE, 2023.

[28] Zihan Li, Yunxiang Li, Qingde Li, et al. Lvit: language meets vision transformer in medical image segmentation. *IEEE Transactions on Medical Imaging*, 2023.

[29] Yuqi Lin, Minghao Chen, Wenxiao Wang, et al. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15305–15314, 2023.

[30] Chunmeng Liu, Yao Shen, Qingguo Xiao, and Guangyao Li. Pcsformer: Pair-wise cross-scale sub-prototypes mining with cnn-transformers for weakly supervised semantic segmentation. *Neurocomputing*, 593:127834, 2024.

[31] Fang Liu, Yuhao Liu, Yuqiu Kong, Ke Xu, Lihe Zhang, Baocai Yin, Gerhard Hancke, and Rynson Lau. Referring image segmentation using text supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22124–22134, 2023.

[32] Qifeng Liu, Tao Zhou, Chi Cheng, Jin Ma, and Marzia Hoque Tania. Hybrid generative adversarial network based on frequency and spatial domain for histopathological image synthesis. *BMC bioinformatics*, 26(1):29, 2025.

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[35] Ming Y Lu, Bowen Chen, Drew FK Williamson, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024.

[36] Ruidong Lu, Qiule Sun, Xueyan Ding, and Jianxin Zhang. Breast cancer histopathology image classification using frequency attention convolution network. In *International Conference on Advanced Data Mining and Applications*, pages 217–229. Springer, 2023.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[38] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 618–626, 2017.

[39] S. Wang, Donghan M. Yang, R. Rong, et al. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170–175, 2016. doi: 10.1016/j.media.2016.06.037.

[40] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, December 2022. doi: 10.18653/v1/2022.emnlp-main.256. URL https://aclanthology.org/2022.emnlp-main.256.

[41] Zifeng Wu, Chunhua Shen, Van Den Hengel, and Anton. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019.

[42] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.

[43] Zunnan Xu, Zhihong Chen, Yong Zhang, et al. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17503–17512, 2023.

[44] Zhao Yang, Jiaqi Wang, Yansong Tang, et al. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022.

[45] Hongrun Zhang, Liam Burrows, Yanda Meng, et al. Weakly supervised segmentation with point annotations for histopathology images via contrast-based variational model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15630–15640, 2023.

[46] Shaoteng Zhang, Jianpeng Zhang, Yutong Xie, and Yong Xia. Tpro: Text-prompting-based weakly supervised histopathology tissue segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 109–118. Springer, 2023.

[47] Lanfeng Zhong, Guotai Wang, Xin Liao, and Shaoting Zhang. Hamil: High-resolution activation maps and interleaved learning for weakly supervised segmentation of histopathological images. *IEEE Transactions on Medical Imaging*, 42(10): 2912–2923, 2023.

[48] Bolei Zhou, Aditya Khosla, Agata Lapedriza, et al. Learning deep features for discriminative localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, June 2016.