This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

IVAAN: Instance-Level Vision-Language Alignment via Attribute-Guided Text Prompt Generation for Nuclei Analysis

Jaehoon Jeong[1,2] Yi Hu[2] Soopil Kim[1] Jongseong Jang[2] Soonyoung Lee[2] Sang Hyun Park[1]

[1]DGIST, Daegu, Republic of Korea [2]LG AI Research, Seoul, Republic of Korea

{j.hoon, soopilkim, shpark13135}@dgist.ac.kr, {j.jang, soonyoung.lee}@lgresearch.ai

Abstract

Nuclei instance segmentation and classification are fundamental but remain challenging in pathology due to severe class imbalance and organ- and stain-induced variability. While vision–language approaches can inject explicit semantic cues that reduce spurious contextual bias under imbalance, the absence of instance level textual annotations has limited their utility for nucleus-level analysis. We introduce an instance-level vision–language framework that derives attribute-guided textual descriptions from ground-truth masks. We then align visual representations with these semantic text anchors via contrastive learning, coupling morphology with semantics at the instance level. To capture intra-class variations while maintaining organ-consistent class semantics, we learn multiple class-specific tokens that act as prototypes representing diverse submodes within a class, summarizing morphologically similar nuclei. Our approach improves both segmentation and classification without manual text labels, indicating that language-guided instance alignment combined with prototype-based semantic feedback yields more discriminative and generalizable nuclei representations.

1. Introduction

Nuclei instance segmentation and classification are fundamental tasks in computational pathology, playing a crucial role in cancer diagnosis and prognosis prediction. Recent nuclei analysis models [12, 34, 40] have achieved remarkable progress in segmentation performance, yet their classification ability remains limited. This limitation is largely due to severe class imbalance and organ-specific biases in existing nuclei datasets: certain phenotypes appear mainly in a few organs and class distributions differ widely across images. Under these conditions, models trained only with class labels must implicitly infer morphology, which encourages them to rely on contextual cues like background coloration, staining intensity or organ-dependent features instead of the intrinsic characteristics of nuclei. As a result, nuclei belonging to the same biological class but originating from different organs or staining protocols are frequently misclassified despite having similar morphology. Because these implicit learning schemes cannot effectively capture variations in shape, color and texture induced by differences in organ or staining conditions, it remains challenging to learn representations that are simultaneously class discriminative and robust to the broad visual variability across organs and stains.

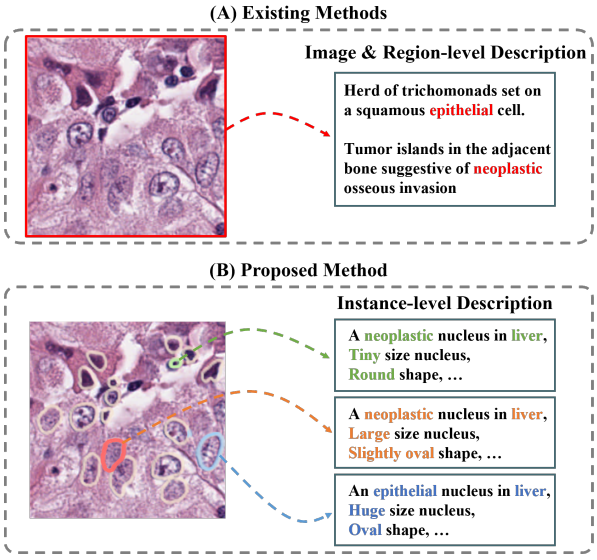


Figure 1. Comparison with existing pathology vision-language models that only use the image- or region-level text prompts.

Recently, vision–language (VL) learning frameworks have emerged as a powerful paradigm to align visual and textual representations in a shared semantic space. By incorporating explicit textual cues, these methods seek to enhance semantic consistency across diverse domains and mitigate the implicit learning issues of purely vision-based

models. Representative models such as CLIP [29], ALIGN [14], and BLIP [17] learn shared image–text embeddings and have demonstrated remarkable transferability across various vision tasks. Following this trend, several VL models tailored for pathology, including PLIP [13] and CONCH [21], have extended CLIP-style alignment to histopathology image–report pairs, achieving impressive domain-level generalization. However, these frameworks are still trained at the image level, making them unsuitable for analyzing individual nuclei with diverse morphological and semantic characteristics as shown in Fig. 1(A). Recent region-level approaches such as GLIP [18], RegionCLIP [43], and GroundingDINO [19] have attempted to align textual descriptions with bounding-box regions, but this formulation cannot be directly applied to nucleus-level instance classification, where numerous small and overlapping nuclei are densely distributed, often leading to mixed or contaminated representations within a single bounding box or region. Moreover, research that analyzes nuclei through textual descriptions has been extremely limited, as constructing instance-level text annotations is prohibitively difficult and labor-intensive. As a result, existing vision-language frameworks face both structural and data limitations in achieving the instance-level alignment required for nuclear analysis.

To enable instance-level visual-language alignment in pathology datasets, we design a process that automatically generates text prompts for each nucleus from ground-truth masks and associated image features, as shown in Fig. 1(B). We quantify morphological attributes that pathologists commonly refer to (*e.g*. size, shape, staining intensity, and boundary irregularity [2, 35]) and convert these measurements into semantically meaningful descriptors that serve as attribute-specific prompts. We use these prompts to perform instance-level contrastive learning that aligns nucleus features with their corresponding semantic attributes, effectively coupling morphology with semantics at the instance level. Unlike prior works, our approach aligns each nucleus with its own attribute prompts, encouraging the model to associate each nucleus with both its morphological appearance and its semantic description. Consequently, the learned representations become class-consistent across organs, improving classification and cross-organ generalization while maintaining strong instance-segmentation performance and reducing the semantic ambiguity observed in earlier models.

Even after aligning each nucleus with attribute-guided prompts, substantial intra-class variability remains because nuclei of the same biological class can have different shapes, colors and textures depending on the organ or tissue. These differences cause nucleus representations to split into multiple submodes, undermining the consistency of class embeddings. To model such diversity and still maintain coherent class-level semantics, we introduce learnable class tokens and a Semantic Interaction Module. These tokens act as class-specific prototypes that represent the diverse visual sub-distributions within each class, each summarizing a subset of morphologically similar nuclei. The Semantic Interaction Module performs cross-attention between the class tokens and the visual features. This bidirectional aggregation encourages nuclei within the same class to converge toward shared prototypes while still preserving legitimate submodes, thereby reducing intra-class variance and accommodating organ-specific distribution differences.

Our contributions are summarized as:

- We present an instance-level VL framework that generates attribute-guided pseudo text prompts and aligns these semantic text anchors with visual features through contrastive learning.
- We introduce learnable class tokens that act as class-specific prototypes, and a Semantic Interaction Module that aggregates instance evidence around these prototypes and propagates class context back to instances.
- Our method achieves superior segmentation and classification performance on multiple datasets.

2. Related Work

2.1. Nuclei Segmentation And Classification

Early approaches for nuclei segmentation improved instance separation by leveraging geometric or distance-based representations [10, 25–27]. DIST [27] regressed a distance map to enhance boundary delineation between adjacent nuclei, while HoVer-Net [10] predicted horizontal and vertical maps to jointly perform instance segmentation and classification. Recent studies extended these ideas with keypoint- or transformer-based modeling. PointNu-Net [40] incorporated auxiliary keypoint predictions to improve localization accuracy within densely packed regions, and CellViT [12] demonstrated that vision transformers can effectively unify segmentation and classification through global attention. More recently, PromptNucSeg [34] employed spatial prompts such as points or masks to guide instance discovery and refinement. However, most frameworks supervise nuclei with only a single class label per instance, without instance-level morphological attributes or organ context. This label-only design implicitly encourages an organ-invariant, unimodal view of each class, making it difficult to represent organ-specific submodes and often leading to ambiguous boundaries in the embedding space.

2.2. VLMs in Nuclei Analysis

Vision-language models (VLMs) such as CLIP [29], ALIGN [14], and BLIP [17] align visual and textual representations within a shared embedding space through large-scale contrastive pretraining. These models have demon-

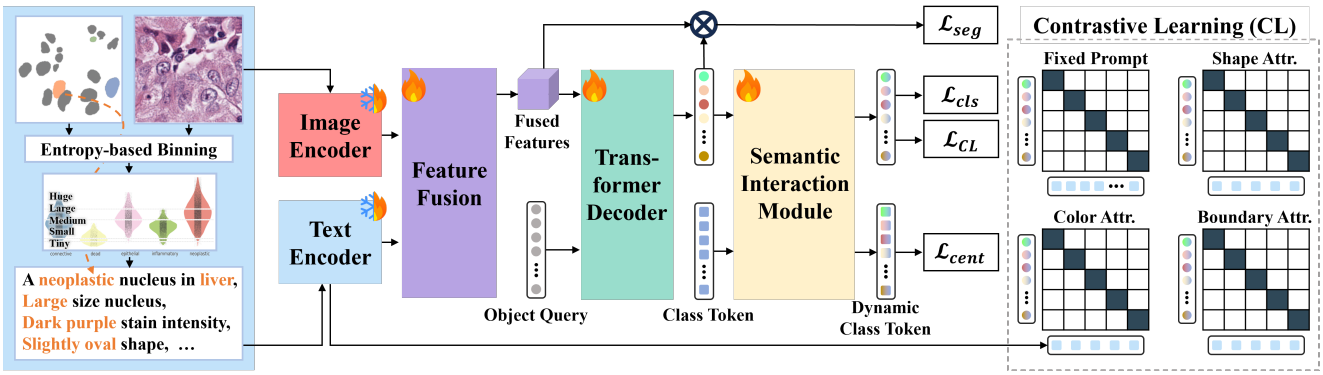


Figure 2. Overview of the proposed framework. The last layers of the text and image encoders are unfrozen for joint optimization.

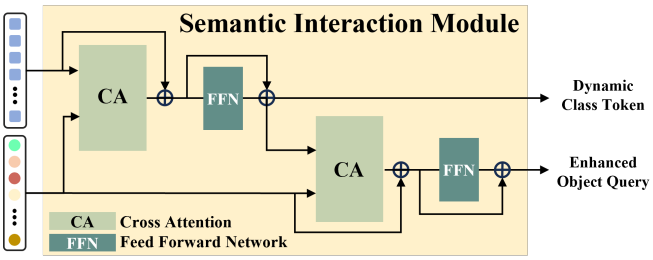


Figure 3. Structure of the Semantic Interaction Module (SIM).

strated strong zero-shot generalization and semantic retrieval capabilities across diverse natural image domains. Following this paradigm, several studies have adapted VLMs to the pathology domain. PLIP [13] extended the CLIP framework by learning from slide–report pairs, achieving domain-level transferability for slide retrieval and zero-shot classification. CONCH [21] scaled up multimodal pretraining using millions of pathology–text pairs, further improving cross-domain generalization. Similarly, PathAlign [1], and ViLa-MIL [33] leveraged slide-level contrastive learning to align diagnostic reports with pathology images. However, these *image-level* VLMs focus on global image–text alignment, lacking the ability to capture fine-grained, instance-level semantics.

To overcome the limitations of image-level alignment, several *region-level* grounding frameworks have been proposed, including GLIP [18], OWL-ViT [24], and GroundingDINO [19]. These models associate textual phrases with bounding-box regions, enabling localized alignment at the object level. However, in histopathology, nuclei are densely distributed and frequently overlap, causing a single bounding box to include multiple nuclei or incomplete boundaries, which leads to mixed or contaminated regional embeddings. Bridging this gap requires an instance-level VLM that can capture localized semantics essential for nucleus-level understanding.

3. Method

3.1. Framework Overview

Fig. 2 shows an overview of our framework. We build upon the transformer-based Mask2Former [6] architecture, which consists of three parts 1) GT-Driven Attribute Prompt Generation, 2) Instance-Level VL Alignment, and 3) Semantic Interaction Module.

We begin by quantifying clinically relevant attributes and discretizing them into semantically meaningful bins; each attribute combination is then converted into a pseudo textual description that provides explicit morphological cues. In parallel, an image encoder extracts multi-scale visual features. These visual features are fused with the text embeddings and then with object queries to produce instance-specific representations that couple visual and linguistic information.

To accommodate intra-class variation, we learn multiple class tokens per class. Acting as local prototypes, these tokens interact bidirectionally with the object query features via the Semantic Interaction Module (Fig. 3), which aggregates visual evidence into the prototypes and propagates global context back to each instance. The enhanced query features resulting from this module are then fed to a mask decoder for segmentation and to a classification head for class prediction. During training, we enforce instance-level vision–language alignment through contrastive learning between the enhanced query features and their corresponding text embeddings.

3.2. GT-Driven Attribute Prompt Generation

To enable instance-level textual supervision in pathology images, 1) we quantitatively analyze nuclear properties that pathologists commonly use for diagnosis, 2) discretize their distributions based on information gain, and 3) convert the resulting attribute ranges into textual prompts. This process allows the model to learn visual–language alignment using interpretable, data-driven linguistic descriptions.

Attribute Definition and Selection. As illustrated in Fig. 4, the distributions of key nuclear attributes differ across semantic classes, indicating that these quantitative descriptors carry class-discriminative information. This observation motivates the use of attribute-guided textual representations for instance-level alignment.

To use attributes, we first define quantitative indicators that correspond to clinically relevant nuclear properties such as pleomorphism (shape irregularity), hyperchromasia (color intensity), membrane irregularity, and chromatin texture [2, 35]. To quantitatively describe these indicators, we extract a total of eleven morphological and intensity-based features from each nucleus: equivalent diameter, eccentricity, aspect ratio, solidity, and extent for shape; normalized perimeter ratio and boundary gradient mean for membrane irregularity; and mean hematoxylin intensity [22], variance of the hematoxylin channel, and core–rim intensity difference for chromatin texture. All chromatin-related features are standardized using per-image z-scores.

To identify features that are both discriminative across classes and nonredundant within the feature set, we evaluate their statistical separability and correlation. Specifically, we compute the effect size (Cohen's d [7]) for every pair of classes and define a separability score for each metric as $S(m) = \text{median}(c_i, c_j)|dij|$, where d_{ij} denotes Cohen's d effect size between classes c_i and c_j. Cohen's d is used to measure the strength of inter-class separation, allowing us to quantify how well a feature distinguishes nuclei of different classes. Features with high redundancy (Spearman's ρ [36] > 0.7) are filtered by retaining the one with the higher separability score. Spearman's correlation is adopted to identify features that convey similar information, thereby preventing overrepresentation of correlated metrics. As a result, we select four representative quantitative attributes: size (*equiv-diam-px*), eccentricity, color intensity (*color-z*), and boundary irregularity (*perim-ratio*), which best capture the four clinical indicators above. After selecting the representative attributes, we transform their continuous values into discrete, semantically meaningful categories to enable textual prompt generation.

Entropy-based Binning. For each of the four selected attributes, we perform a supervised discretization process to obtain class-informative bins instead of relying on fixed thresholds. Given attribute values x and class labels y, we evaluate candidate thresholds θ by computing the information gain between parent and child partitions. The entropy within a segment is defined as $H = -\sum_c p(c) \log_2 p(c)$, where $p(c)$ denotes the class probability within the segment. For a threshold θ, we compute the total entropy before splitting (H_{total}) and the entropies of the left and right partitions (H_L and H_R). The information gain is then formulated as

$$\text{Gain}(\theta) = H_{\text{total}} - \frac{n_L}{n} H_L - \frac{n_R}{n} H_R, \tag{1}$$

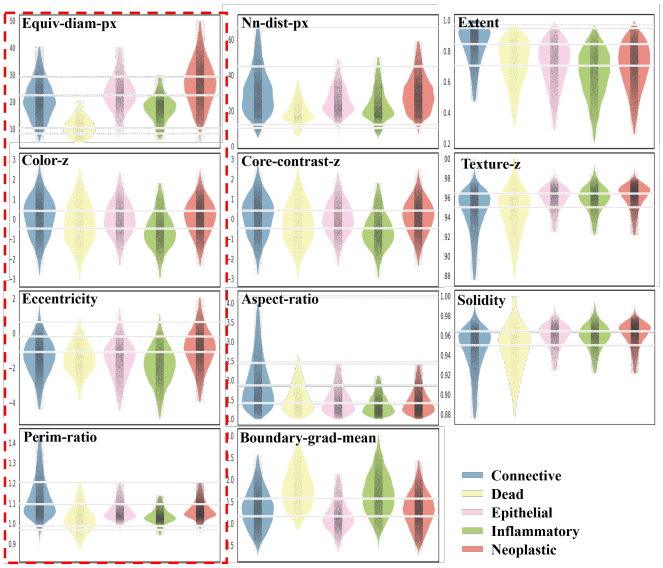


Figure 4. Violin distributions for eleven representative attributes. A red dotted box indicates the selected attributes. A black dotted lines in the graphs marks the partition boundary. To ensure a clear visualization, we removed statistical outliers using Tukey's fences [37].

where n_L and n_R denote the number of samples in the left and right partitions, respectively, and $n = n_L + n_R$. The optimal threshold θ^* is selected as the one that maximizes $\text{Gain}(\theta)$: $\theta^* = \arg\max_\theta \text{Gain}(\theta)$.

This process is recursively applied up to a maximum depth of four, producing up to five attribute intervals. To ensure stability, the discretization is repeated across five folds of the training set, and thresholds appearing in at least 60% of the iterations are retained through clustering.

Finally, each interval is assigned a human-readable text label (e.g., *tiny*, *small*, *medium*, *large*, *huge*), forming attribute-aware textual prompts for each instance. These instance-level prompts are later used as linguistic supervision to facilitate vision-language contrastive learning.

3.3. Instance-Level Vision-Language Alignment

Vision-Language Alignment. With the text prompts we generated, we perform contrastive learning using matched query–instance pairs obtained from Hungarian matching following [6]. For each matched enhanced object query feature f_{enhq}, produced by the Semantic Interaction Module, we apply a linear projection W_v followed by ℓ_2 normalization to obtain the visual embedding $v = \text{normalize}(W_v f_{enhq}) \in \mathbb{R}^D$. Each text prompt is encoded by a pretrained CLIP text encoder whose final layer is set trainable to adapt linguistic representations to the pathology domain. The resulting text feature f_{text} is projected through a learnable linear layer W_t and ℓ_2-normalized to

produce the text embedding $t = \text{normalize}(W_t f_{text})$.

We employ two types of textual prompts to provide complementary supervision. The first, referred to as *fixed prompts*, conveys class-level context and follows the template "a {class} nucleus in {organ}," serving as semantic anchors for the class-level contrastive objective:

$$L_{\text{fix}} = -\frac{1}{N}\sum_i \log \frac{\exp(v_i^\top T_{y_i}/\tau)}{\sum_j \exp(v_i^\top T_j/\tau)}, \tag{2}$$

where T_{y_i} denotes the text embedding of the ground-truth class y_i, τ is the temperature parameter controlling the sharpness of alignment, and j indexes all possible classes in the text embedding set $T = \{T_1, \ldots, T_C\}$.

In addition to the fixed prompts, we use *attribute prompts* that describe fine-grained morphology such as size, color, shape, or boundary irregularity. For each attribute $a \in A$ and instance i, we define one positive attribute text $t^+_{i,a}$ corresponding to the ground-truth attribute and consider all other attribute texts as negatives. Each text is encoded and projected as above to obtain normalized embeddings. The contrastive loss for attribute a is formulated as

$$\mathcal{L}^a_{\text{attr}} = -\frac{1}{N_a}\sum_i \log \frac{\exp(v_i^\top t^+_{i,a}/\tau_{\text{attr}})}{\sum_m \exp(v_i^\top t^m_{i,a}/\tau_{\text{attr}})}, \tag{3}$$

where τ_{attr} is the temperature parameter for attribute alignment, N_a is the number of valid samples for attribute a, and m indexes all possible attribute texts for attribute a. The final attribute-level loss $\mathcal{L}_{\text{attr}}$ is computed as the average of $\mathcal{L}^a_{\text{attr}}$ across attributes. Consequently, VL loss is defined as

$$\mathcal{L}_{\text{CL}} = \lambda_{fix}\mathcal{L}_{fix} + \lambda_{attr}\mathcal{L}_{attr}. \tag{4}$$

Class Token and Semantic Interaction Module. As shown in Fig. 5, there is intra-class variability within a same class. To capture diverse intra-class variations and maintain consistent class semantics across organs, we further learn multiple class tokens per class. We introduce a set of $(C+1)\times k$ learnable class tokens, where C denotes the number of foreground nucleus classes and the additional one corresponds to the background category. For each class, k queries are assigned to represent intra-class variations, functioning as multiple local prototypes rather than a single centroid. This design enables the model to capture diverse morphological appearances within a class, such as variations in nuclear size, shape, or staining intensity.

We devise a Semantic Interaction Module to aggregate query evidence and inject global context back into instance representations. During decoding, the class tokens (CTs) interact bidirectionally with object queries (OQs) through a dual-attention mechanism. (Fig. 3) In the first direction (OQ→CT), the base class tokens serve as the query, while the object queries act as key–value inputs in cross-attention, generating dynamically updated representations denoted as *dynamic CTs*. Each dynamic CT aggregates class-relevant features from all object queries within the batch, effectively summarizing the visual evidence of that class. As training progresses, these aggregated representations evolve into dataset-level class prototypes that encode the global feature distribution across all nuclei. In the opposite direction (CT→OQ), the dynamic class tokens are used as key–value inputs, while the object queries serve as the query. Through this second cross-attention, followed by an FFN layer, object queries are refined into *enhanced OQs* that inherit class-level semantic information learned from the dynamic CTs. This process allows object queries to not only benefit from image-level evidence but also from the accumulated class knowledge across the dataset, reinforcing class consistency and improving semantic discriminability.

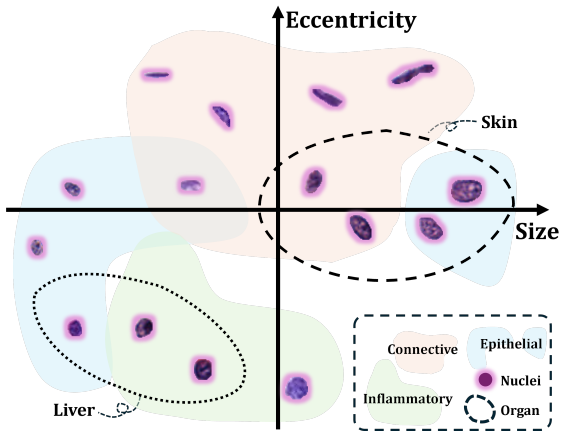


Figure 5. Simplified illustration of **intra-class variability** in nucleus classification. Nuclei from the same organ cluster more tightly than same-class nuclei across organs, complicating class-level representation learning.

To provide semantic anchoring, we align the mean representation of each group of class tokens with the corresponding class text embedding. Formally, the averaged feature of k queries belonging to class c is encouraged to be close to the text embedding T_c in the shared vision-language space:

$$\mathcal{L}_{\text{cent}} = -\frac{1}{C}\sum_{c=1}^{C} \log \frac{\exp(\bar{q}_c^\top T_c/\tau_{\text{CT}})}{\sum_{j=1}^{C}\exp(\bar{q}_c^\top T_j/\tau_{\text{CT}})}, \tag{5}$$

where $\bar{q}_c = \frac{1}{k}\sum_{i=1}^{k} q_{c,i}$ denotes the mean representation of class c, T_c is the text embedding of the corresponding class description, and τ_{CT} is the temperature parameter. This loss encourages each group of class tokens to form a compact local cluster around its text embedding. It also allows the text feature to act as a semantic anchor while individual class tokens capture intra-class diversity across decoding layers.

Method	Detection			Classification																	
	Average			Neoplastic			Epithelial			Inflammatory			Connective			Dead			Average		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Mask-RCNN [11]	0.76	0.68	0.72	0.55	0.63	0.59	0.52	0.52	0.52	0.46	0.54	0.50	0.42	0.43	0.42	0.17	0.30	0.22	0.42	0.47	0.45
DIST [27]	0.74	0.71	0.73	0.49	0.55	0.50	0.38	0.33	0.35	0.42	0.45	0.42	0.42	0.37	0.39	0.00	0.00	0.00	0.34	0.34	0.33
StarDist [32]	0.85	0.80	0.82	0.69	0.69	0.69	0.73	0.68	0.70	0.62	0.53	0.57	0.54	0.49	0.51	0.39	0.09	0.10	0.59	0.50	0.52
Micro-Net [30]	0.78	0.82	0.80	0.59	0.66	0.62	0.63	0.54	0.58	0.59	0.46	0.52	0.50	0.45	0.47	0.23	0.17	0.19	0.51	0.46	0.48
Hover-Net [10]	0.82	0.79	0.80	0.58	0.67	0.62	0.54	0.60	0.56	0.56	0.51	0.54	0.52	0.47	0.49	0.28	0.35	0.31	0.50	0.52	0.50
CPP-Net [5]	0.87	0.78	0.82	0.74	0.67	0.70	0.74	0.70	0.72	0.60	0.57	0.58	0.57	0.49	0.53	0.41	0.36	0.38	0.61	0.56	0.58
PointNu-Net [40]	0.81	0.81	0.81	0.74	0.72	0.73	0.75	0.71	0.73	0.63	0.57	0.60	0.61	0.55	0.58	0.45	0.24	0.31	0.64	0.56	0.59
CellViT-H [12]	0.84	0.81	0.83	0.72	0.69	0.71	0.72	0.73	0.73	0.59	0.57	0.58	0.55	0.52	0.53	0.43	0.32	0.36	0.60	0.57	0.58
PromptNucSeg-H [34]	0.82	0.85	0.84	0.70	0.72	0.71	0.73	0.78	0.76	0.58	0.61	0.59	0.55	0.55	0.55	0.44	0.49	0.46	0.60	0.63	0.61
Ours-H	0.86	0.87	**0.87**	0.77	0.80	**0.79**	0.79	0.81	**0.80**	0.70	0.71	**0.70**	0.67	0.65	**0.66**	0.58	0.43	**0.49**	0.70	0.68	**0.69**

Table 1. Precision (P), Recall (R), and F1-score (F1) for detection and classification across three folds by nucleus type. The best and second-best F1-scores are shown in **bold** and underline, respectively. Following [10], a classification is treated as a true positive (TP) when a predicted nucleus is within $\approx 3\mu$m of a ground-truth nucleus with the same class label; otherwise, it is considered a false positive (FP).

Feature Fusion. Following prior vision-language models [18, 19], we integrate vision features with text embeddings to narrow the modality gap and provide early semantic guidance to the visual representation. This fusion step allows the visual features to absorb textual context such as organ and class semantics, which facilitates more grounded instance decoding. Specifically, the text embeddings are projected into the same dimensional space as the visual tokens and interact with multi-scale visual features through cross-attention, allowing semantic cues to modulate spatial representations before decoding. Unlike instance-level textual alignment, we employ an image-level text representation that summarizes all nucleus categories present in the image. For example, if an image contains both neoplastic and connective nuclei, the text description becomes "a photo of a neoplastic nuclei. a photo of a connective nuclei." This aggregated text serves as a global semantic context, ensuring that the fused features encode class-level semantics. During inference, we use a fixed sentence that lists all class descriptions, without any per-image labels.

4. Experiment

4.1. Dataset

PanNuke contains 7,899 images (256×256) from 19 tissue types with 189,744 nuclei in total. Experiments follow the three-fold cross-validation protocol of [9].
MoNuSeg includes 44 pathological images (1,000×1,000) from 9 organs, comprising 28,846 manually annotated nuclei. Each image was divided into 250 patches for training.
CPM17 consists of 32 training and 32 testing tiles from 4 cancer types further split into 250–300 patches for training.

4.2. Implementation Details

We train with AdamW [20] (lr = 1e-4, batch = 8) at 3,000 steps using a lr schedule on NVIDIA A100 GPUs. The model uses 200 object queries. The total loss is $\mathcal{L} = \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{CL}}\mathcal{L}_{\text{CL}} + \lambda_{\text{cent}}\mathcal{L}_{\text{cent}}$, where $\lambda_{seg} = 5$, $\lambda_{cls} = 2$, $\lambda_{CL} = 1$, $\lambda_{cent} = 2$, and within λ_{CL} we set $\lambda_{fix} = 1, \lambda_{attr} = 0.3$. The decoder and the definitions of λ_{seg}, λ_{cls} follow [6]. We used SAM [15] ViT encoder and the interface between SAM and Mask2Former follows [41]. We set $\tau = 0.07, \tau_{CT} = 0.07, \tau_{attr} = 0.15$ and $k = 5$.

4.3. Main Results

Table 1, Table 2, and Table 3 summarize the quantitative results on the PanNuke dataset. Our model achieves the highest detection and classification accuracy, with F1-scores of 0.87 and 0.69, respectively (Table 1). For instance-level segmentation, our approach attains an average binary PQ (bPQ) of 0.6976 and a multi-class PQ (mPQ) of 0.5459 (Table 2), outperforming PromptNucSeg by +0.005 and +0.034 points, respectively. These consistent improvements across both detection and segmentation tasks demonstrate that our vision–language framework effectively aligns instance-level and semantic representations.

At the class level (Table 3), the largest PQ gains appear in *inflammatory*, *connective*, and *dead*: the first two are frequently confused due to similar morphology; our attribute-guided texts inject fine-grained shape and intensity cues that disambiguate *inflammatory* vs. *connective*. This indicates that the integration of textual priors enhances class discrimination under confusing morphology. The proposed model also improves the recognition of minor class such as *dead* cells, showing the most improvement among the classes. This demonstrates that our model mitigates class imbalance problems. These findings highlight the strength of our vision–language alignment and semantic feedback design in capturing both class semantics and local morphology.

Beyond PanNuke, we further evaluate the proposed method on two additional benchmarks, MoNuSeg and CPM17, as shown in Table 5. Our model achieves higher PQ and Aggregated Jaccard Index (AJI) scores than SOTA on MoNuSeg and CPM17, even with vit-L backbone.

The overall performance gain primarily arises from the synergy between the fixed and attribute-level vi-

Tissue	Mask R-CNN [11]		StarDist [32]		Hover-Net [10]		CPP-Net [5]		PointNu-Net [40]		CellViT-H [12]		PromptNucSeg [34]		Ours	
	bPQ	mPQ	bPQ	mPQ	bPQ	mPQ	bPQ	mPQ	bPQ	mPQ	bPQ	mPQ	bPQ	mPQ	bPQ	mPQ
Adrenal	0.5546	0.3470	0.6972	0.4868	0.6962	0.4812	0.7066	0.4944	0.7134	0.5115	0.7086	0.5134	0.7227	0.5128	**0.7357**	**0.5871**
Bile Duct	0.5567	0.3536	0.6690	0.4651	0.6696	0.4714	0.6768	0.4670	0.6814	0.4868	0.6784	0.4887	0.6976	0.5012	**0.7148**	**0.5322**
Bladder	0.6049	0.5065	0.6986	0.5793	0.7031	0.5792	0.7053	0.5936	0.7226	0.6065	0.7068	0.5844	0.7212	0.6043	**0.7349**	**0.6145**
Breast	0.5574	0.3882	0.6666	0.5064	0.6470	0.4902	0.6747	0.5090	0.6709	0.5147	0.6748	0.5180	**0.6842**	0.5322	0.6841	**0.5503**
Cervix	0.5483	0.3402	0.6690	0.4628	0.6652	0.4438	0.6912	0.4792	0.6899	0.5014	0.6872	0.4984	0.6983	0.5118	**0.7143**	**0.5910**
Colon	0.4603	0.3122	0.5779	0.4205	0.5575	0.4095	0.5911	0.4315	0.5945	0.4509	0.5921	0.4485	0.6096	0.4690	**0.6112**	**0.4806**
Esophagus	0.5691	0.4311	0.6655	0.5331	0.6427	0.5085	0.6797	0.5449	0.6766	0.5504	0.6682	0.5454	0.6920	0.5711	**0.6947**	**0.5850**
Head & Neck	0.5457	0.3946	0.6433	0.4768	0.6331	0.4530	0.6523	0.4706	0.6546	0.4838	0.6544	0.4913	0.6695	0.5104	**0.6813**	**0.5109**
Kidney	0.5092	0.3553	0.6998	0.4880	0.6836	0.4424	0.7067	0.5194	0.6912	0.5066	0.7092	0.5366	**0.7115**	**0.5786**	0.6801	0.5540
Liver	0.6085	0.4103	0.7231	0.5145	0.7248	0.4974	0.7312	0.5143	0.7314	0.5174	0.7322	0.5224	**0.7372**	0.5333	0.7370	**0.5655**
Lung	0.5134	0.3182	0.6362	0.4128	0.6302	0.4004	0.6386	0.4256	0.6352	0.4048	0.6426	0.4314	**0.6580**	0.4398	0.6470	**0.4826**
Ovarian	0.5784	0.4337	0.6668	0.5205	0.6309	0.4863	0.6830	0.5313	0.6863	0.5484	0.6722	0.5390	0.6856	0.5442	**0.6928**	**0.5736**
Pancreatic	0.5460	0.3624	0.6601	0.4585	0.6491	0.4600	0.6789	0.4706	0.6791	0.4804	0.6658	0.4719	**0.6863**	0.4974	0.6781	**0.5395**
Prostate	0.5789	0.3959	0.6748	0.5067	0.6615	0.5101	0.6927	0.5305	0.6854	0.5127	0.6821	0.5321	0.6983	0.5456	**0.7067**	**0.5677**
Skin	0.5021	0.2665	0.6289	0.3610	0.6234	0.3429	0.6209	0.3574	0.6494	0.4011	0.6565	0.4339	0.6613	0.4113	**0.6952**	**0.4496**
Stomach	0.5976	0.3684	0.6944	0.4477	0.6886	0.4726	0.7067	0.4582	0.7010	0.4517	0.7022	0.4705	**0.7115**	0.4559	0.7045	**0.5475**
Testis	0.5420	0.3512	0.6869	0.4942	0.6890	0.4754	0.7026	0.4931	0.7058	0.5334	0.6955	0.5127	0.7151	0.5474	**0.7202**	**0.5793**
Thyroid	0.5712	0.3037	0.6962	0.4300	0.6983	0.4315	0.7155	0.4392	0.7076	0.4508	0.7151	0.4519	0.7218	0.4721	**0.7397**	**0.5360**
Uterus	0.5589	0.3683	0.6599	0.4480	0.6393	0.4393	0.6615	0.4794	0.6634	0.4846	0.6625	0.4737	0.6743	0.4955	**0.6818**	**0.5259**
Average	0.5528	0.3688	0.6692	0.4744	0.6596	0.4629	0.6798	0.4847	0.6808	0.4957	0.6793	0.4980	0.6924	0.5123	**0.6976**	**0.5459**
Std	0.0076	0.0047	0.0014	0.0037	0.0036	0.0076	0.0015	0.0059	0.0050	0.0082	0.0318	0.0413	0.0093	0.0147	0.0009	0.0061

Table 2. Performance comparison on the PanNuke dataset. Following [12, 34], both binary panoptic quality (bPQ) and multi-class panoptic quality (mPQ) across 19 organs are reported. The best and second-best scores in each row are shown in **bold** and underlined, respectively. Standard deviations are computed over three-fold experiments following [9].

MethodClass	Neoplastic	Epithelial	Inflammatory	Connective	Dead
Mask-RCNN [11]	0.472	0.403	0.290	0.300	0.069
DIST [27]	0.439	0.290	0.343	0.275	0.000
StarDist [32]	0.547	0.532	0.424	0.380	0.123
Micro-Net [30]	0.504	0.442	0.333	0.334	0.051
HoVer-Net [10]	0.551	0.491	0.417	0.388	0.139
CPP-Net [5]	0.571	0.565	0.405	0.395	0.131
PointNu-Net [40]	0.578	0.577	0.433	0.409	0.154
CellViT-H [12]	0.581	0.583	0.417	0.423	0.149
PromptNucSeg-H [34]	0.598	0.582	0.441	0.433	0.161
Ours-H	**0.605**	**0.591**	**0.481**	**0.452**	**0.204**

Table 3. Average panoptic quality (PQ) over three folds for each nucleus category in the PanNuke dataset.

	VL	Attr	Entr	SIM	FF	det-F1	cls-F1	PQ	AJI
(1)						78.5	63.1	57.3	61.6
(2)	✓					83.8	67.1	62.3	65.3
(3)	✓	✓				83.9	67.4	63.8	65.5
(4)	✓	✓	✓			84.5	67.7	65.4	66.1
(5)	✓	✓	✓	✓		86.7	69.3	66.4	67.5
(6)	✓	✓	✓	✓	✓	**87.0**	**69.5**	**67.3**	**68.3**

Table 4. Ablation study of key components in the proposed framework. Attr: attribute prompts; Entr: entropy-based binning; FF: feature fusion. Row 3 denotes attribute prompts from quantile (equal-count) bins rather than entropy-optimized bins.

Method	MoNuSeg [16]		CPM17 [39]	
	AJI	PQ	AJI	PQ
U-Net [31]	0.556	0.478	0.666	0.625
DCAN [3]	0.525	0.492	0.561	0.545
Mask-RCNN [11]	0.546	0.509	0.684	0.674
DIST [27]	0.559	0.443	0.616	0.504
Micro-Net [30]	0.560	0.519	0.668	0.661
CIA-Net [44]	0.620	0.577	–	–
Full-Net [28]	0.601	0.620	0.702	0.686
Hover-Net [10]	0.618	0.597	0.705	0.697
Triple U-Net [42]	0.621	0.601	0.711	0.685
FEEDNet [8]	0.616	0.613	0.701	0.705
HARU-Net [4]	0.613	0.572	0.721	0.701
PointNu-Net [40]	0.606	0.603	0.712	0.706
PromptNucSeg-B [34]	0.614	0.620	0.731	0.726
PromptNucSeg-L [34]	0.621	0.626	0.734	0.730
PromptNucSeg-H [34]	0.622	0.627	0.740	0.733
Ours-B	0.664	0.647	0.729	0.727
Ours-L	0.666	0.665	0.741	0.734
Ours-H	**0.689**	**0.696**	**0.743**	**0.748**

Table 5. Performance comparison on the MoNuSeg and CPM17 datasets. The best results are highlighted in **bold**, and the previous state-of-the-art scores are underlined.

sion–language objectives. The fixed prompts act as global semantic anchors that encode nucleus type and organ context, while the attribute prompts provide finer morphological cues such as size, shape, and stain intensity. Combined with the class tokens and the Semantic Interaction Module (SIM), this dual alignment mechanism enables the model to learn class-consistent yet visually adaptive features.

4.4. Ablation Studies

We conduct ablation experiments to evaluate the contribution of each proposed component in our framework (Table 4). The baseline (Row 1) represents the model trained without any textual or semantic modules. When instance-level vision–language alignment (VL) is applied using only fixed prompts such as "a {class} nucleus in {organ}"

(Row 2), both detection and classification F1-scores, showing that even fixed prompts helps regularize the feature space. Adding attribute-guided prompts (Attr) (Row 4) and Entropy-based binning (Entr) can further improve the performance as fine-grained morphological cues enhance inter-class separation beyond fixed textual anchors. Incorporating the class-token and Semantic Interaction Module (SIM) (Row 5) yields an additional performance gain by enabling bidirectional communication between instance queries and class-level prototypes. Finally, Feature Fusion (FF) (Row 6) achieves the best overall performance. These results demonstrate that each component contributes cumulatively, and their combination yields robust instance-level understanding and improved segmentation quality.

4.5. Further Analysis

GT Attributes Analysis. To assess whether morphological attributes carry discriminative information beyond pixel intensities, we examine their distributions across the five PanNuke classes. Fig. 4 plots violin distributions for representative attributes. Each class exhibits a distinct profile: neoplastic nuclei tend to have larger equivalent diameters and connective nuclei have high perimeter ratios. Notably, *dead* cells occupy the extreme low end of the size spectrum; *equivalent diameters* are consistently below 20 for this class, which means that any nucleus larger than 20 pixels is highly unlikely to belong to the *dead* class.

Feature Space Visualization. To show how the learned features separate classes, we embed the instance-level representations from our model and a baseline using UMAP [23] and t-SNE [38]. Fig. 6 highlights two regions of interest. In the green dotted box, our vision–language features separate the inflammatory and connective nuclei more clearly than the baseline, where these morphologically similar classes heavily overlap. In the red dotted box, our model forms distinct subclusters for neoplastic and dead nuclei, and the text anchor (large square) for the dead class lies at the far end of its distribution, away from the neoplastic cluster; in contrast, the baseline embedding shows complete overlap in this region. Overall, text anchors lie within their respective class regions and away from decision boundaries, showing that attribute-guided prompts and SIM organize the feature space around semantic categories and morphological cues, yielding more discriminative and generalizable representations.

5. Limitation

For the dead class, two tokens drift into areas where dead instances are scarcely observed (Fig. 6). Because the dead class is a minority class, a token budget of $k = 5$ is excessive, causing SIM to concentrate on a few tokens while under-attended tokens receive minimal supervision. Since $\mathcal{L}_{cent}$ constrains only the mean of each token group, these under-attended tokens can be nudged by noisy gradients toward off-manifold regions. This suggests that class-adaptive token budgets or monitoring mechanisms that pruning or reinitialization could be beneficial. Fig. 6 also indicates that Connective and Inflammatory features heavily overlap in feature space, yet all five Connective tokens remain concentrated near the Connective text anchor. This configuration underrepresents boundary-region variability and may limit discrimination in the mixed region. To better separate such overlapping classes, a potential direction is to introduce a mild diversification constraint that increases within-class tokens dispersion and encourages more explicit boundary-oriented submode specialization.

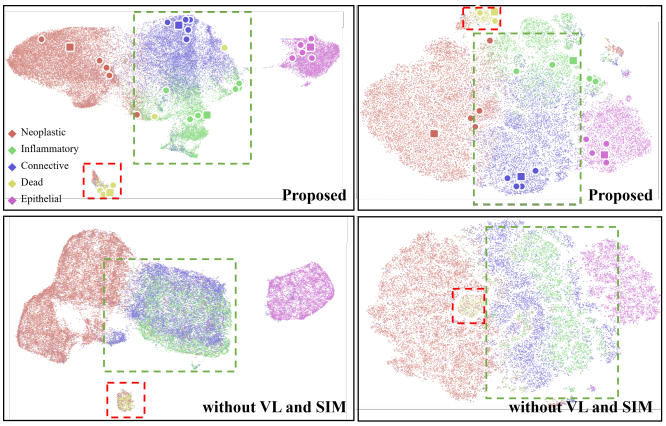


Figure 6. UMAP (left) and t-SNE (right) comparison between w/ and w/o VL alignment and SIM. The green and red boxes highlight whether inflammatory and connective categories are distinguishable and whether dead cells are clearly separated. The colored squares and circles with white edge marks the text anchors and the cluster tokens, respectively.

6. Conclusion

Nuclei segmentation and classification are challenging due to large intra-class variability and class imbalance across organs and stains. We addressed these issues with an instance-level vision–language alignment that uses attribute-guided prompts derived from ground-truth masks to steer learning toward intrinsic nuclear morphology rather than spurious context, and with class-specific prototype tokens coupled with a Semantic Interaction Module to inject dataset-level class semantics while preserving legitimate submodes. This combination reduces intra-class variance and improves cross-organ consistency without manual text labels. Experiments on PanNuke, MoNuSeg, and CPM17 show consistent gains in segmentation and classification, indicating that instance-level language guidance plus prototype-based semantic feedback yields more discriminative and generalizable nuclei representations.

References

[1] Faruk Ahmed, Andrew Sellergren, Lin Yang, Shawn Xu, Boris Babenko, Abbi Ward, Niels Olson, Arash Mohtashamian, Yossi Matias, Greg S Corrado, et al. Pathalign: A vision-language model for whole slide images in histopathology. *arXiv preprint arXiv:2406.19578*, 2024. 3

[2] Daphne W. Bell, Daniel C. Crum, and David A. Haber. Origins and molecular pathology of ovarian cancer. *Nature Reviews Cancer*, 5(11):819–831, 2005. 2, 4

[3] Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng. Dcan: deep contour-aware networks for accurate gland segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2487–2496, 2016. 7

[4] Junzhou Chen, Qian Huang, Yulin Chen, Linyi Qian, and Chengyuan Yu. Enhancing nucleus segmentation with harunet: a hybrid attention based residual u-blocks network. *arXiv preprint arXiv:2308.03382*, 2023. 7

[5] Shengcong Chen, Changxing Ding, Minfeng Liu, Jun Cheng, and Dacheng Tao. Cpp-net: Context-aware polygon proposal network for nucleus segmentation. *IEEE Transactions on Image Processing*, 32:980–994, 2023. 6, 7

[6] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 3, 4, 6

[7] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 2013. 4

[8] Gayatri Deshmukh, Onkar Susladkar, Dhruv Makwana, Sparsh Mittal, et al. Feednet: A feature enhanced encoder-decoder lstm network for nuclei instance segmentation for histopathological diagnosis. *Physics in Medicine & Biology*, 67(19):195011, 2022. 7

[9] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benes, Simon Graham, Mostafa Jahanifar, Syed Ali Khurram, Ayesha Azam, Katherine Hewitt, and Nasir Rajpoot. Pannuke dataset extension, insights and baselines. *arXiv preprint arXiv:2003.10778*, 2020. 6, 7

[10] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis*, 58:101563, 2019. 2, 6, 7

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6, 7

[12] Fabian Hörst, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel, Jens Siveke, Barbara Grünwald, Jan Egger, et al. Cellvit: Vision transformers for precise cell segmentation and classification. *Medical Image Analysis*, 94:103143, 2024. 1, 2, 6, 7

[13] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023. 2, 3

[14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 6

[16] Neeraj Kumar, Ruchika Verma, Deepak Anand, Yanning Zhou, Omer Fahri Onder, Efstratios Tsougenis, Hao Chen, Pheng-Ann Heng, Jiahui Li, Zhiqiang Hu, et al. A multi-organ nucleus segmentation challenge. *IEEE transactions on medical imaging*, 39(5):1380–1391, 2019. 7

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2

[18] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022. 2, 3, 6

[19] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 2, 3, 6

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[21] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024. 2, 3

[22] Marc Macenko, Marc Niethammer, James S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE international symposium on biomedical imaging: from nano to macro*, pages 1107–1110. IEEE, 2009. 4

[23] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv e-prints*, abs/1802.03426, 2020. Originally submitted 2018; this version revised Sept. 2020. 8

[24] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European conference on computer vision*, pages 728–755. Springer, 2022. 3

[25] Siwoo Nam, Jaehoon Jeong, Miguel Luna, Philip Chikontwe, and Sang Hyun Park. Pronet: Point refinement using

shape-guided offset map for nuclei instance segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 528–538. Springer, 2023. 2

[26] Siwoo Nam, Hyun Namgung, Jaehoon Jeong, Miguel Luna, Soopil Kim, Philip Chikontwe, and Sang Hyun Park. Instasam: Instance-aware segment any nuclei model with point annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 232–242. Springer, 2024.

[27] Peter Naylor, Marick Laé, Fabien Reyal, and Thomas Walter. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE transactions on medical imaging*, 38(2):448–459, 2018. 2, 6, 7

[28] Hui Qu, Zhennan Yan, Gregory M Riedlinger, Subhajyoti De, and Dimitris N Metaxas. Improving nuclei/gland instance segmentation in histopathology images by full resolution neural network and spatial constrained loss. In *International conference on medical image computing and computer-assisted intervention*, pages 378–386. Springer, 2019. 7

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[30] Shan E Ahmed Raza, Linda Cheung, Muhammad Shaban, Simon Graham, David Epstein, Stella Pelengaris, Michael Khan, and Nasir M Rajpoot. Micro-net: A unified model for segmentation of various objects in microscopy images. *Medical image analysis*, 52:160–173, 2019. 6, 7

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 7

[32] Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. Cell detection with star-convex polygons. In *International conference on medical image computing and computer-assisted intervention*, pages 265–273. Springer, 2018. 6, 7

[33] Jiangbo Shi, Chen Li, Tieliang Gong, Yefeng Zheng, and Huazhu Fu. Vila-mil: Dual-scale vision-language multiple instance learning for whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024*, 2024. 3

[34] Zhongyi Shui, Yunlong Zhang, Kai Yao, Chenglu Zhu, Sunyi Zheng, Jingxiong Li, Honglin Li, Yuxuan Sun, Ruizhe Guo, and Lin Yang. Unleashing the power of prompt-driven nucleus instance segmentation. In *European conference on computer vision*, pages 288–304. Springer, 2024. 1, 2, 6, 7

[35] Indra Singh, Subhasis Dasgupta, and Ramesh K. Aggarwal. Nuclear morphological abnormalities in cancer. In *Results and Problems in Cell Differentiation*, pages 45–61. Springer Nature, 2022. 2, 4

[36] Charles Spearman. The proof and measurement of association between two things. 1961. 4

[37] John W. Tukey. *Exploratory Data Analysis*. Addison–Wesley, Reading, MA, 1977. 4

[38] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (86):2579–2605, 2008. 8

[39] Quoc Dang Vu, Simon Graham, Tahsin Kurc, Minh Nguyen Nhat To, Muhammad Shaban, Talha Qaiser, Navid Alemi Koohbanani, Syed Ali Khurram, Jayashree Kalpathy-Cramer, Tianhao Zhao, et al. Methods for segmentation and classification of digital microscopy tissue images. *Frontiers in bioengineering and biotechnology*, 7:53, 2019. 7

[40] Kai Yao, Kaizhu Huang, Jie Sun, and Amir Hussain. Pointnu-net: Keypoint-assisted convolutional neural network for simultaneous multi-tissue histology nuclei segmentation and classification. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(1):802–813, 2023. 1, 2, 6, 7

[41] Tao Zhang, Xingye Tian, Yikang Zhou, Shunping Ji, Xuebo Wang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, and Yu Wu. Dvis++: Improved decoupled framework for universal video segmentation. *arXiv preprint arXiv:2312.13305*, 2023. 6

[42] Bingchao Zhao, Xin Chen, Zhi Li, Zhiwen Yu, Su Yao, Lixu Yan, Yuqian Wang, Zaiyi Liu, Changhong Liang, and Chu Han. Triple u-net: Hematoxylin-aware nuclei segmentation with progressive dense feature aggregation. *Medical Image Analysis*, 65:101786, 2020. 7

[43] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16793–16803, 2022. 2

[44] Yanning Zhou, Omer Fahri Onder, Qi Dou, Efstratios Tsougenis, Hao Chen, and Pheng-Ann Heng. Cia-net: Robust nuclei instance segmentation with contour-aware information aggregation. In *International conference on information processing in medical imaging*, pages 682–693. Springer, 2019. 7