INSIGHT: Explainable Weakly-Supervised Medical Image Analysis

Wenbo Zhang WZHANG66@UR.ROCHESTER.EDU
University of Rochester, Rochester, NY, USA

Junyu Chen JCHEN175@UR.ROCHESTER.EDU
University of Rochester, Rochester, NY, USA

Christopher Kanan CKANAN@CS.ROCHESTER.EDU
University of Rochester, Rochester, NY, USA

Abstract

Due to their large sizes, volumetric scans and whole-slide pathology images (WSIs) are often processed by extracting embeddings from local regions and then an aggregator makes predictions from this set. However, current methods require post-hoc visualization techniques (e.g., Grad-CAM) and often fail to localize small yet clinically crucial details. To address these limitations, we introduce INSIGHT, a novel weakly-supervised aggregator that integrates heatmap generation as an inductive bias. Starting from pre-trained feature maps, INSIGHT employs a detection module with small convolutional kernels to capture fine details and a context module with a broader receptive field to suppress local false positives. The resulting internal heatmap highlights diagnostically relevant regions. On CT and WSI benchmarks, INSIGHT achieves state-of-the-art classification results and high weakly-labeled semantic segmentation performance. Project website and code are available at: https://zhangdylan83.github.io/ewsmia/

1. Introduction

Advances in medical imaging technology have enabled clinicians to extract critical insights from increasingly large and complex datasets. However, the size of medical images presents computational and analytical challenges: whole-slide images (WSIs) of histopathology slides can contain billions of pixels (Campanella et al., 2019), and volumetric scans, such as CT or MRI, are composed of hundreds of slices. Processing such data end-to-end with deep neural networks is computationally infeasible. Instead, pipelines rely on aggregators, which synthesize local embeddings extracted from tiles (WSIs) or slices (volumes) into global predictions (Casson et al., 2024; Chang et al., 2022; Mahmood et al., 2021). While this divide-and-conquer strategy is efficient, current methods often discard spatial information during feature aggregation and depend on post-hoc visualization tools, such as Grad-CAM (Selvaraju et al., 2019), to generate interpretable heatmaps. These visualizations are prone to missing clinically significant features and introduce additional complexity.

To address these limitations, we propose **INSIGHT** (**I**ntegrated **N**etwork for **S**egmentation and **I**nterpretation with **G**eneralized **H**eatmap **T**ransmission), a novel weakly-supervised aggregator that embeds interpretability directly into its architecture. INSIGHT processes pre-trained patch- or slice-level embeddings through two complementary modules: a *Detec-*

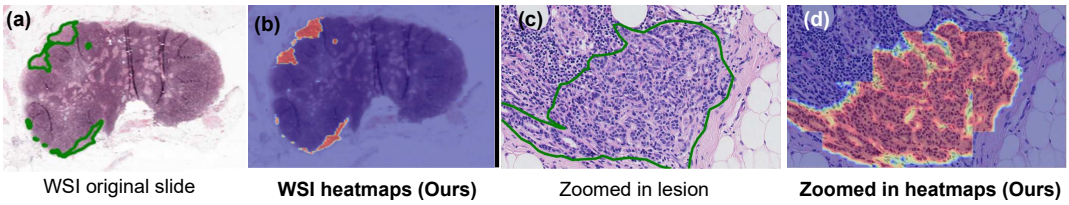

Figure 1: Qualitative visualization of interpretable heatmaps generated by INSIGHT. (a) Original whole-slide image (WSI) with annotated tumor regions outlined in green. (b) WSI-level heatmap produced by our method, highlighting predicted tumor regions. (c) Zoomed-in view of a representative lesion with expert annotation. (d) Corresponding zoomed-in heatmap showing precise localization of the lesion by INSIGHT. All heatmaps are generated using **only** image-level categorical labels.

tion Module, which captures fine-grained diagnostic features, and a *Context Module*, which suppresses irrelevant activations by incorporating broader spatial information. These outputs are fused to create interpretable *internal* heatmaps that naturally align with diagnostic regions. INSIGHT produces internal segmentation-quality heatmaps **without** requiring pixel- or voxel-level annotations as shown in Fig. 1.

Weak supervision is essential in medical imaging, where the creation of detailed annotations is expensive and labor-intensive (Campanella et al., 2019). Fully supervised methods like U-Net (Ronneberger et al., 2015) and DeepLab (Chen et al., 2017) rely on densely labeled data, which limits their scalability. In contrast, multiple instance learning (MIL) (Carbonneau et al., 2018; Campanella et al., 2019) enables models to aggregate features from local regions using only image- or slide-level labels, avoiding the need for pixel-level annotations. This approach allows models to train on significantly larger datasets, as physicians typically diagnose patients without annotating individual image regions. For instance, the FDA-cleared Paige Prostate system demonstrates the clinical viability of weak supervision by achieving high diagnostic accuracy for prostate cancer detection (Zhu et al., 2024; Campanella et al., 2019). Similarly, MIL-based methods have effectively aggregated patch- or slice-level features to make predictions on WSIs and volumetric scans (Casson et al., 2024; Chang et al., 2022).

However, current weakly supervised methods face critical limitations. They require post-hoc visualization techniques, such as Grad-CAM (Selvaraju et al., 2019) or class activation maps (CAM) (Zhou et al., 2015), to produce interpretable saliency maps, adding complexity and often failing to localize small yet clinically essential details. These saliency maps can also be biased by dataset characteristics, such as organ co-occurrence, and often struggle to balance classification accuracy with spatial localization (Cohen et al., 2020). This leaves considerable room for improvement in designing aggregators that can better capture spatial dependencies and produce clinically meaningful outputs. INSIGHT addresses these challenges by directly embedding interpretability into its architecture, eliminating the reliance on post-hoc methods and bridging the gap between computational efficiency and clinical relevance.

This paper makes the following contributions:

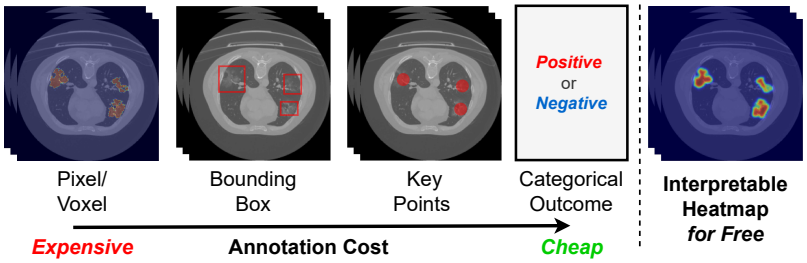

Figure 2: Semantic segmentation for medical data often requires expensive annotations that are not done routinely in clinical practice, whereas INSIGHT only needs Categorical labels (left). Unlike other weakly-labeled methods, INSIGHT has per-label heatmaps built into the architecture as a form of inductive bias, enabling it to directly produce heatmaps that closely align with ground-truth diagnostic regions *without* the use of post-hoc methods like Grad-CAM (right).

- We introduce INSIGHT, a novel weakly-supervised aggregator that directly integrates explainability into its architecture, producing high-quality heatmaps as part of its decision-making process.
- We demonstrate that INSIGHT achieves state-of-the-art performance on three challenging medical imaging datasets, excelling in both classification and segmentation tasks across CT and WSI modalities using only image-level categorical labels.
- Through extensive qualitative and quantitative analyses, we show that INSIGHT's heatmaps align closely with diagnostic regions, requiring only image-level labels for supervision.

Generalizable Insights about Machine Learning in the Context of Healthcare

INSIGHT demonstrates that fine-grained interpretability and strong predictive performance are not mutually exclusive in weakly supervised settings. By maintaining spatial resolution throughout the feature processing pipeline and embedding interpretability directly into the architecture, our method provides a general design principle for scalable and explainable medical AI systems. This architectural approach—preserving locality while incorporating contextual suppression—can be extended beyond pathology and CT to other modalities such as MRI or ultrasound. More broadly, INSIGHT challenges the common reliance on post-hoc attention visualization, offering a viable alternative that integrates decision explanation into the prediction mechanism itself.

For the clinical radiologist or pathologist, INSIGHT produces built-in heatmaps that align with diagnostic regions without requiring pixel-level annotations. These interpretable outputs help clinicians verify model predictions and identify overlooked regions, while traditional pixel-level annotation remains both labor-intensive and time-consuming. By reducing reliance on such dense annotations and producing visual explanations, INSIGHT can accelerate image review and enhance diagnostic confidence, especially for subtle or ambiguous cases.

2. Related Work

2.1. Weakly Supervised Medical Image Learning

Fully supervised methods such as U-Net (Ronneberger et al., 2015) and DeepLab (Chen et al., 2017) have achieved high segmentation accuracy in medical imaging. However,

their reliance on dense, pixel- or voxel-level annotations limits scalability in clinical workflows (Campanella et al., 2019). Weakly supervised learning mitigates this issue by enabling models to train on coarser annotations, such as image- or slide-level labels, which are more practical in large-scale clinical datasets. Methods based on MIL have proven effective for WSIs and volumetric data, where detailed annotations are unavailable. MIL aggregates features from local patches or slices to predict global outcomes, bypassing the need for pixel-level labels. For example, the FDA-cleared Paige Prostate system demonstrated the clinical viability of MIL by achieving high diagnostic accuracy for prostate cancer detection using only slide-level labels (Campanella et al., 2019; Zhu et al., 2024). MIL-based methods have also been successfully applied to breast cancer subtyping from WSIs (Casson et al., 2024) and chemotherapy response prediction from volumetric CT scans (Chang et al., 2022).

While alternative weak labeling strategies, such as scribbles (Liu et al., 2022), bounding boxes (Lu et al., 2023), or point-level markers (Laradji et al., 2020), reduce annotation demands, they still require explicit region annotations, which are rarely performed by physicians in clinical practice. INSIGHT builds on the MIL paradigm by utilizing only image-level labels for both classification and segmentation. Unlike existing methods, INSIGHT incorporates interpretability directly into its architecture, generating fine-grained heatmaps that align with diagnostic regions without requiring post-hoc visualization techniques.

2.2. Aggregation Techniques in Medical Imaging

Aggregation techniques are essential for efficiently processing large-scale medical imaging data, such as WSIs and volumetric scans, without requiring dense annotations. In WSIs, MIL-based methods treat each slide as a collection of patches and aggregate features for global predictions. For instance, MIL has been successfully applied to cancer detection using slide-level labels (Campanella et al., 2019), while advancements like STAMP (Nahhas et al., 2023) and tissue-graph approaches (Pati et al., 2023) have improved WSI analysis by refining aggregation techniques to detect patterns and segment tissues. In volumetric imaging, slice-level features are aggregated to form volume-level predictions. For example, multi-resolution systems like Saha et al. (2020) use DenseVNet-generated features to improve classification of chest CT scans, while Ye et al. (2022) developed an aggregation framework with label correction to enhance outcome predictions.

INSIGHT advances these techniques by preserving spatial resolution across slices or patches until the final pooling stage, enabling the generation of detailed, localized heatmaps. This addresses a key limitation of traditional aggregation methods, which often discard spatial information, reducing their ability to localize diagnostically relevant regions. Existing methods aim at enhancing contextual understanding in feature aggregation, but they each have inherent limitations. For instance, FPN (Lin et al., 2017) is designed for hierarchical multi-scale feature fusion but lacks an explicit mechanism to independently optimize local and global features, which is crucial for WSI. On the other hand, Dual Attention Network (Fu et al., 2019) relies on self-attention mechanisms to model long-range dependencies across spatial and channel dimensions, leading to quadratic complexity and computationally expensive for high-resolution medical images such as WSI. In contrast, unlike FPN which depends on predefined hierarchical feature aggregation, INSIGHT's Context module efficiently refines activations within the same feature scale. It also allows separate optimiza-

tions on local and global representations. Meanwhile, compared to Dual Attention Nets, INSIGHT avoids the quadratic complexity of self-attention with larger convolution kernel size, making it significantly more efficient for processing whole-slide images.

2.3. Explainable AI Methods in Medical Imaging

Explainable AI is crucial for ensuring model predictions in medical imaging are interpretable and clinically meaningful. Post-hoc methods such as class activation maps (CAM) (Zhou et al., 2015) and Grad-CAM (Selvaraju et al., 2019) are widely used to generate saliency maps that highlight regions associated with model predictions. For example, Xie et al. (2023) developed high-resolution CAMs for thoracic CT scans, and Viniavskyi et al. (2020) refined pseudo-labels for chest X-ray segmentation using CAMs. Attention mechanisms have further advanced interpretability. Self-attention, as introduced in (Vaswani et al., 2023), has been combined with Grad-CAM to localize disease-specific regions (Zhang et al., 2022), while dual-attention mechanisms, such as DA-CMIL (Chikontwe et al., 2021), produce interpretable maps for detecting conditions like COVID-19 and bacterial pneumonia.

While these methods generate useful insights, they rely on post-hoc interpretability, which can misalign with the model's decision-making process. Few existing methods directly integrate interpretability into their model architectures. One example is Shallow-ProtoPNet (Singh et al., 2025), a prototype-based network designed for fully transparent inference on small-scale datasets such as chest X-rays. In contrast, INSIGHT is built to scale to large, high-resolution inputs like CT volumes and WSIs, addressing the computational and spatial challenges that arise in real-world clinical imaging. INSIGHT eliminates this reliance by embedding interpretability directly into its architecture, producing calibrated, built-in heatmaps as part of its predictions, as shown in Fig. 3. ensures alignment between the model's outputs and diagnostic reasoning.

2.4. Pre-training in Medical Image Analysis

Pre-trained models have significantly advanced medical imaging by providing robust feature extractors. ResNet-based architectures, such as those used in (Courtiol et al., 2020; Lin et al., 2022), effectively capture local and intermediate features in weakly supervised frameworks.

On the other hand, Vision Transformers (ViTs) (Dosovitskiy et al., 2021) leverage self-attention to capture broader spatial relationships. Self-supervised foundational models are typically pre-trained on massive unlabeled image datasets, enabling the learning of meaningful representations without manual annotations. They produce high-quality features that are highly compatible with downstream tasks including those based on aggregation.

For example, UNI (Chen et al., 2024) is trained on over 100 M WSIs using a DINOv2-style (Oquab et al., 2024) self-supervised learning framework and provides pathology-specific representations optimized for fine-grained localization. Virchow 2 (Zimmermann et al., 2024), a 632-million-parameter ViT-H model, is pre-trained on 3.1 M WSIs with domain-specific augmentations tailored for histopathology tasks. INSIGHT leverages pre-trained ViTs to extract tensor embeddings from WSI patches and radiology slices. By preserving spatial resolution, INSIGHT achieves robust performance across WSI and CT datasets, supporting fine-grained analysis and interpretability.

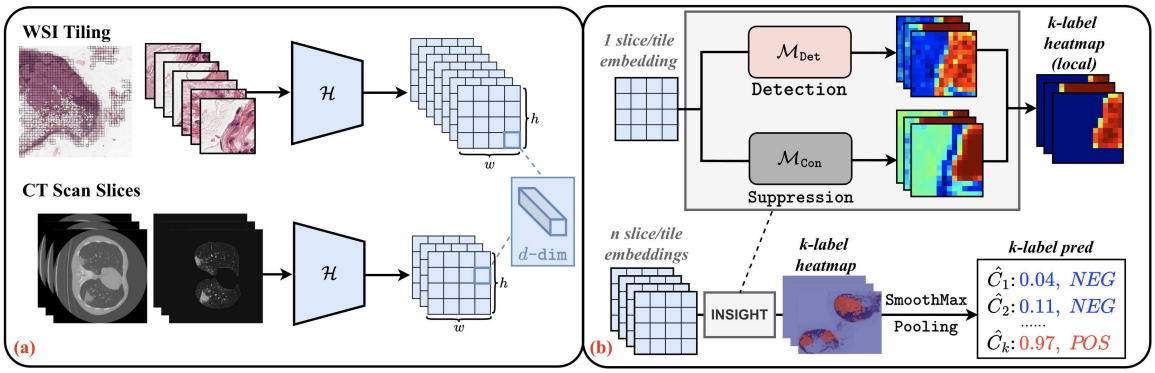

Figure 3: Overview of the INSIGHT framework. **(a)** Input images (WSIs or CT volumes) are preprocessed and transformed into spatial embeddings using a pretrained encoder. **(b)** Each slice or patch embedding is processed by INSIGHT, which consists of a detection module to capture fine-grained signals and a context suppression module to reduce false positives. This produces patch-level heatmaps for each category, which are then aggregated to generate whole-slide or volume-level heatmaps. Final predictions for each category are obtained via SmoothMax pooling over the heatmaps. Throughout this process, spatial resolution is preserved, encouraging the model to produce reliable and interpretable heatmaps.

3. Methods

Fig. 3 illustrates the high-level overview architecture of INSIGHT. INSIGHT processes pre-trained features from each slice in a volume or each patch in a WSI through dual modules—Detection and Context—to generate fine-grained heatmaps. These heatmaps are aggregated using a SmoothMax pooling strategy to produce categorical predictions, optimized using a combination of binary cross-entropy (BCE) and spectral decoupling losses. Below, we describe each component in detail.

3.1. Feature Extraction

To preserve spatial resolution and capture both local and global contextual information, INSIGHT uses pre-trained models specific to the input modality, as illustrated in Fig 3(a). To demonstrate that INSIGHT is compatible with a variety of foundation models across modalities, we evaluate it with multiple state-of-the-art encoders. For CT scan volumes, we use a ViT pre-trained with DINOv2 (Oquab et al., 2024) on ImageNet. For pathology WSIs, we evaluate INSIGHT with both UNI (Chen et al., 2024) and Virchow 2 (Zimmermann et al., 2024). UNI is pre-trained using DINOv2-style self-supervision on over 100 million WSIs, enabling it to capture rich visual features tailored to histopathology. Virchow 2 is a large ViT-H model with 632 million parameters, trained on 3.1 million WSIs using domain-specific augmentations optimized for pathology. In all cases, the extracted feature maps are denoted as $F \in \mathbb{R}^{c \times h \times w}$, where c, h, and w are the feature depth, height, and width, respectively.

3.2. INSIGHT Architecture

Heatmap Fusion. The core of INSIGHT lies in its architectural design which introduces an inductive bias that integrates both local and global spatial features. As illustrated in Fig. 3(b), each input embedding is processed by two parallel modules: the *Detection module* and the *Context modules*, which are combined to generate a refined heatmap:

$$\mathcal{H} = \sigma((1 - \sigma(\mathcal{H}_{\texttt{Con}})) \odot \mathcal{H}_{\texttt{Det}}), \tag{1}$$

where σ is the sigmoid activation, and $\odot$ denotes element-wise multiplication. The resulting heatmap $\mathcal{H}$ assigns relevance scores to regions, aligning with the target labels. *Detection module* uses small-kernel convolutions to capture fine-grained details such as textures and edges, producing a heatmap $\mathcal{H}_{\texttt{Det}}$ that highlights potential regions of interest. To mitigate false positives, the *Context module* uses larger kernels to capture broader spatial context and suppress irrelevant activations. To better align the pre-trained features with the downstream task, we apply a linear projection before passing the embeddings to both modules.

To isolate high-saliency regions within the fused heatmap, we apply dynamic thresholding with Otsu's algorithm (Otsu, 1979):

$$\mathcal{H}' = \mathcal{H} \cdot \mathbb{I}(\mathcal{H} > T), \tag{2}$$

where $T = \arg\min_t \sigma_w^2(t)$ minimizes intra-class variance and $\mathbb{I}$ is the indicator function. This step enhances localization by filtering out low-activation background regions.

Pooling and Classification. To produce a categorical outcome, heatmaps are aggregated across all slices or patches using SmoothMax pooling (Maddison et al., 2016), which is a Boltzmann-weighted operator that emphasizes regions with strong activations:

$$\hat{y} = \frac{\sum_i \mathcal{H}'_i \cdot \exp(\alpha \cdot \mathcal{H}'_i)}{\sum_i \exp(\alpha \cdot \mathcal{H}'_i)}, \tag{3}$$

where $\hat{y}$ is the predicted likelihood of a positive outcome, $\mathcal{H}'_i$ are heatmap activations, and α controls the sharpness of the weighting. For multi-label classification, the heatmap is extended across channels, with each channel representing a separate category, allowing joint optimization over multiple labels.

We empirically tested several other pooling methods, including max pooling and LP pooling to compare their effectiveness. Max pooling considers only the strongest activation, which limits spatial gradient flow during training and leads to suboptimal heatmap quality. LP pooling ($p \geq 2$)incorporates more values but often produces outputs outside the $[0, 1]$ range, requiring post-processing (e.g., output clipping or sparse regularization (Zhou et al., 2021)), which yielded no empirical gains. In contrast, SmoothMax pooling provides a principled trade-off between local emphasis and global context through exponential weighting.

3.3. Training Objective

Binary Cross-Entropy Loss. Since our model is trained using only volume- or slide-level categorical labels, we optimize the predicted probabilities using the binary cross-entropy loss $\mathcal{L}_{\text{BCE}}$. For multi-label settings, the loss is independently computed for each class.

Spectral Decoupling. To address gradient starvation and encourage robust decision boundaries, we include spectral decoupling loss (Pohjonen et al., 2022):

$$\mathcal{L}_{\text{SD}} = \frac{\lambda_{\text{SD}}}{2} \cdot \|z\|_2^2, \tag{4}$$

where z are unnormalized logits, and λ_{SD} controls regularization strength. Spectral decoupling is one of the most effective methods for overcoming dataset bias without additional sub-group labels (Shrestha et al., 2022a,b).

Total Loss. The total loss combines these objectives:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{SD}}. \tag{5}$$

3.4. Heatmap Visualization

Heatmaps are reconstructed to match the original image dimensions for intuitive interpretation. For WSIs, patch-level heatmaps are stitched together based on their spatial coordinates. For CT volumes, slice-level heatmaps are stacked in order to reconstruct the full volume. The resulting heatmaps are then upsampled using bicubic interpolation to ensure smooth and visually consistent representations. The pseudocode for the heatmap generation and classification process is provided in Appendix D.

4. Datasets and Preprocessing

We evaluate INSIGHT on three publicly available datasets representing two imaging modalities: whole slide images (WSI) and CT volumes.

4.1. Computed Tomography Dataset

MosMed (Morozov et al., 2020). The MosMed dataset comprises 1,110 chest CT studies for COVID-19 detection, split into two subsets: MosMed-A (1,060 volumes) and MosMed-B (50 volumes). MosMed-A is used for classification tasks, employing five-fold cross-validation with an 80/20 train-test split. MosMed-B, containing voxel-level annotations, is used solely to evaluate segmentation performance by comparing INSIGHT-generated heatmaps $\mathcal{H}$ against ground truth using Dice score. CT volumes are normalized to the Hounsfield Unit (HU) range $[-1000, 400]$, scaled to $[0, 1]$, and resized to $518 \times 518 \times 32$. To ensure compatibility with pre-trained ViTs, single-channel data is replicated across three channels. Lung parenchyma is isolated using Lungmask (Hofmanninger et al., 2020).

4.2. Whole Slide Image Datasets

CAMELYON16 (Litjens et al., 2018). This dataset contains 399 WSIs for detecting metastatic breast cancer in lymph nodes. Following UNI (Chen et al., 2024), we use the 20× magnification slides in our experiments, omitting one unusable slide from the original training set. We split the remaining 269 training slides into 243 for training and 26 for validation. The model checkpoint with the highest validation Dice is evaluated on the official test set of 129 slides.

BRACS (Brancati et al., 2021). This dataset contains 547 WSIs annotated with 4539 bounding boxes for breast carcinoma subtype classification: *ADH* (Atypical Ductal Hyperplasia – early premalignant change), *FEA* (Flat Epithelial Atypia – precursor lesion), *DCIS* (Ductal Carcinoma in Situ – non-invasive cancer), and *Invasive* (Invasive Carcinoma – malignant stromal invasion). As BRACS provides only the most severe label for each slide, we extracted fine-grained labels using QuPath (Bankhead et al., 2017). Following BRACS's official splits and excluding slides with incomplete labeling or no valid regions, the final dataset comprises 347, 51, and 73 slides for training, validation, and testing. We downsample all WSIs from 40× to 20× magnification.

We use the CLAM toolbox (Lu et al., 2021) for patch extraction for both CAMELYON16 and BRACS.

5. Experiments

This section presents the experimental setup and results. We provide additional implementation details, including training configurations and model architecture parameters, in Appendix A.

5.1. Compared Baselines

We compare INSIGHT to state-of-the-art methods across the MosMed, CAMELYON16, and BRACS datasets, evaluating both classification and segmentation performance under a weakly supervised setting.

CT Volumes. For the MosMed dataset, we benchmark against methods specifically designed for COVID-19 CT analysis.

- **Progressively Resized 3D-CNN** (Hasan et al., 2021), a CNN that progressively resizes input volumes to extract multi-scale information.
- **3D U-Net** (Bressem et al., 2021), a fully supervised segmentation model trained with voxel-level annotations that provides a direct comparison compared to weakly supervised methods.
- **3D GAN** (Shabani et al., 2022), a weakly supervised generative model that performs volume-level segmentation without requiring dense voxel-level annotations.

WSI Datasets. To ensure a fair comparison, all methods (including ours) use **the same foundational model** encoder for feature extraction:

- **ABMIL** (Ilse et al., 2018): An attention-based MIL approach that identifies the most informative regions within a slide, commonly used in weakly supervised WSI analysis.
- **CLAM-SB/MB** (Lu et al., 2021): A MIL method combining attention mechanisms with clustering to improve localization. CLAM-SB uses a single branch for non-overlapping classes, while -MB employs multiple branches to target specific classes.
- **TransMIL** (Shao et al., 2021): A transformer-based MIL model that captures long-range dependencies within WSIs, enhancing contextual analysis for slide-level classification.

Table 1: Performance comparison for classification and segmentation on the MosMed dataset. Methods marked with * were trained on an external dataset but tested on the same MosMed test set for consistency. Baseline details in Sec. 5.1.

Task	Method	AUC / Dice (%)
Classification	PR-3D-CNN (Five-fold CV)	0.914 ± 0.049
	INSIGHT (Five-fold CV)	**0.962 ± 0.012**
Segmentation	3D U-Net* (Voxel-level)	40.5 ± 21.3
	3D GAN (Volume-level)	41.2 ± 14.7
	INSIGHT (Volume-level)	**42.7 ± 15.3**

- **WiKG** (Li et al., 2024): A dynamic graph representation learning framework that models knowledge graphs, implementing a knowledge-aware attention mechanism to capture contextual interactions.

5.2. Quantitative Results

INSIGHT consistently outperformed baseline methods in classification and segmentation tasks across all datasets, demonstrating adaptability to diverse medical imaging modalities. Performance metrics are reported in Table 1 and Table 2. On **MosMed**, INSIGHT achieved a classification AUC of 96.2%, nearly 5% higher than the best baseline, underscoring its robustness in volumetric CT analysis for COVID-19 detection. On **CAMELYON16**, INSIGHT achieved a Dice score of 74.6%, outperforming the top competing model by 6.9%, highlighting its ability to enhance boundary precision in segmentation. On **BRACS**, INSIGHT showed strong multi-label classification performance, with AUC improvements of 3.3% for both ADH and FEA, showcasing its effectiveness in distinguishing subtle tissue variations critical for subtype classification.

INSIGHT's superior performance stems from three key architectural innovations:
Spatial Embedding. By retaining spatial resolution during feature extraction, INSIGHT produces $14 \times 14 \times 1024$ feature maps per WSI tile, preserving diagnostically meaningful structure. In contrast, baseline methods reduce feature maps to $1 \times 1 \times 1024$ via pooling, discarding fine-grained information. This enables INSIGHT to detect small but critical patterns, especially in BRACS, where distinct subtypes may co-exist within a single slide.
Adjacent Context Focus. INSIGHT's context module captures local dependencies using larger convolutional filters. This focus on adjacent spatial relationships is especially effective for multi-label datasets like BRACS, where lesion subtypes often appear in close proximity.
Broader Foreground Activation Consideration. Unlike MIL-based models that focus only on top k patches, INSIGHT's smooth pooling considers all relevant activations above a threshold. This preserves *finer* boundary details, enhancing segmentation accuracy. For example, on CAMELYON16, this broader inclusion contributed to its 6.9% Dice improvement.

We further conducted experiments using the **Virchow 2** foundational model (Zimmermann et al., 2024), a more recent state-of-the-art pretraining approach for histopathology images, to assess INSIGHT's performance and robustness against other benchmark methods. As shown in Table 3, INSIGHT again largely outperformed all baselines in segmenta-

Table 2: Performance comparison on the CAMELYON16 and BRACS official test datasets. ALL aggregators use **UNI** for embeddings. We report AUC and Dice scores (mean ± std) for CAMELYON16, and AUC scores for different categories in BRACS. Reported Dice scores for all comparison methods were fine-tuned through grid search on the validation set. Baseline details in Sec. 5.1. BRACS's subtype definitions are provided in Sec. 4.2.

Aggregator	CAMELYON16		BRACS			
	AUC	Dice	ADH	FEA	DCIS	Invasive
ABMIL	0.975	55.8±25.0	0.656	0.744	0.804	0.995
CLAM-SB	0.966	64.7±24.1	0.611	0.757	0.833	**0.999**
CLAM-MB	0.973	67.7±22.6	0.701	0.687	0.828	0.998
TransMIL	0.982	12.4±22.4	0.644	0.653	0.769	0.989
WiKG	0.967	66.1±24.2	0.454	0.653	0.771	0.990
INSIGHT	**0.990**	**74.6±19.1**	**0.734**	**0.790**	**0.837**	**0.999**

Table 3: Performance comparison on the CAMELYON16 official test dataset using **Virchow 2** for embeddings. We report AUC and Dice scores (mean ± std) for CAMELYON16. Reported Dice scores for all comparison methods were fine-tuned through grid search on the validation set. Baseline details in Sec. 5.1.

Aggregator	AUC	Dice (%)
ABMIL	0.963	59.8±25.7
CLAM-SB	0.992	66.8±25.1
CLAM-MB	0.986	66.2±21.9
TransMIL	0.980	9.3±20.1
WiKG	**0.996**	66.8±21.4
INSIGHT	0.992	**78.3±15.4**

tion task on the CAMELYON16 dataset. INSIGHT's performance improvements are further supported by qualitative results (Fig. 4) and ablation studies (Sec. 5.5), demonstrating the critical role of its architectural innovations in achieving state-of-the-art results.

5.2.1. Stratified Analysis

To assess performance across lesion scales, we extracted connected lesion components and corresponding predicted heatmaps from the entire test set. Lesions were grouped into three size-based bins based on pixel area: **Small** ($< 10^6$ pixels), **Moderate** (10^6–10^7 pixels), and **Large** ($> 10^7$ pixels). These thresholds were chosen based on the empirical area distribution in the dataset to reflect clinically meaningful differences. The bin counts are: 1415 (Small), 135 (Moderate), and 62 (Large), showing that small lesions dominate the dataset and represent a particularly challenging and clinically important subgroup. Table 4 reports the Dice scores of INSIGHT and comparison models scores across different lesion sizes using both UNI and Virchow 2 backbones, along with one-tailed permutation test results (10,000

Table 4: Performance comparison (Dice, mean ± std) on different lesion sizes. Paired one-tailed permutation hypothesis tests (10,000 iterations) compare each baseline to INSIGHT. Bolded indicates $0.01 \leq p < 0.05$, and ** indicates $p < 0.01$.

		Small		Moderate		Large	
Model	**Aggregator**	**Dice (%)**	***p*-value**	**Dice (%)**	***p*-value**	**Dice (%)**	***p*-value**
UNI	**INSIGHT**	$\mathbf{42.8 \pm 38.6}$	-	$\mathbf{74.6 \pm 27.3}$	-	$\mathbf{82.1 \pm 19.8}$	-
	CLAM-MB	20.7 ± 32.2	**	62.2 ± 27.8	**	$\mathbf{81.4 \pm 19.6}$	0.40
	CLAM-SB	20.8 ± 32.2	**	62.6 ± 27.4	**	$\mathbf{81.2 \pm 19.2}$	0.36
	ABMIL	11.0 ± 25.2	**	51.1 ± 29.8	**	72.4 ± 23.0	**
	TransMIL	30.6 ± 37.3	**	23.4 ± 32.0	**	34.1 ± 32.3	**
	WiKG	19.5 ± 31.9	**	58.7 ± 29.1	**	$\mathbf{80.3 \pm 21.0}$	0.29
Virchow2	**INSIGHT**	$\mathbf{55.5 \pm 37.0}$	-	$\mathbf{78.6 \pm 22.1}$	-	$\mathbf{81.2 \pm 19.7}$	-
	CLAM-MB	24.5 ± 33.3	**	60.9 ± 24.7	**	$\mathbf{81.4 \pm 17.7}$	0.52
	CLAM-SB	27.7 ± 34.3	**	66.6 ± 22.8	**	$\mathbf{84.1 \pm 16.5}$	0.81
	ABMIL	10.8 ± 24.4	**	50.9 ± 30.1	**	74.6 ± 23.6	0.04
	TransMIL	42.4 ± 37.0	**	35.4 ± 32.8	**	50.9 ± 29.9	**
	WiKG	24.1 ± 33.9	**	62.1 ± 26.4	**	$\mathbf{82.8 \pm 17.8}$	0.68

iterations) comparing INSIGHT to each baseline. INSIGHT consistently outperforms all baselines on **Small** and **Moderate** lesions ($p < 0.01$), which are especially critical in clinical diagnosis due to their subtle presentation. For **Large** lesions, INSIGHT performs comparably or better than baselines. These findings highlight INSIGHT's robustness across lesion scales and its particular advantage in detecting small, subtle pathological regions often missed by conventional MIL-based methods.

5.3. Qualitative Analysis

We visualize and compare heatmaps generated by INSIGHT and baseline methods in Fig. 4. Unlike baseline methods, which often rely on post-hoc adjustments, INSIGHT directly integrates interpretability into its architecture, mapping each patch's probability to the heatmap. This eliminates the need for additional calibration steps, producing well-calibrated, fine-grained heatmaps that effectively highlight diagnostically significant regions. Additional visualizations are provided in Appendix B, which further illustrate the superior interpretability and precision of INSIGHT's heatmaps across both WSIs and CT datasets.

WSI Heatmap Comparisons:

- **INSIGHT:** The heatmaps generated by INSIGHT accurately capture subtle structures and ensure comprehensive coverage of ground-truth areas. Zoomed-in views reveal precise delineation of lesion boundaries, showcasing INSIGHT's ability to localize regions critical for diagnosis. Ground-truth areas exhibit high activation values, reflecting the model's robustness in spatially aligning predictions with annotations.
- **ABMIL:** Heatmaps from ABMIL frequently produce false negatives, detecting only a limited number of patches and failing to capture a holistic view of lesion regions. This lack of coverage diminishes its diagnostic utility, particularly in cases with subtle abnormalities.

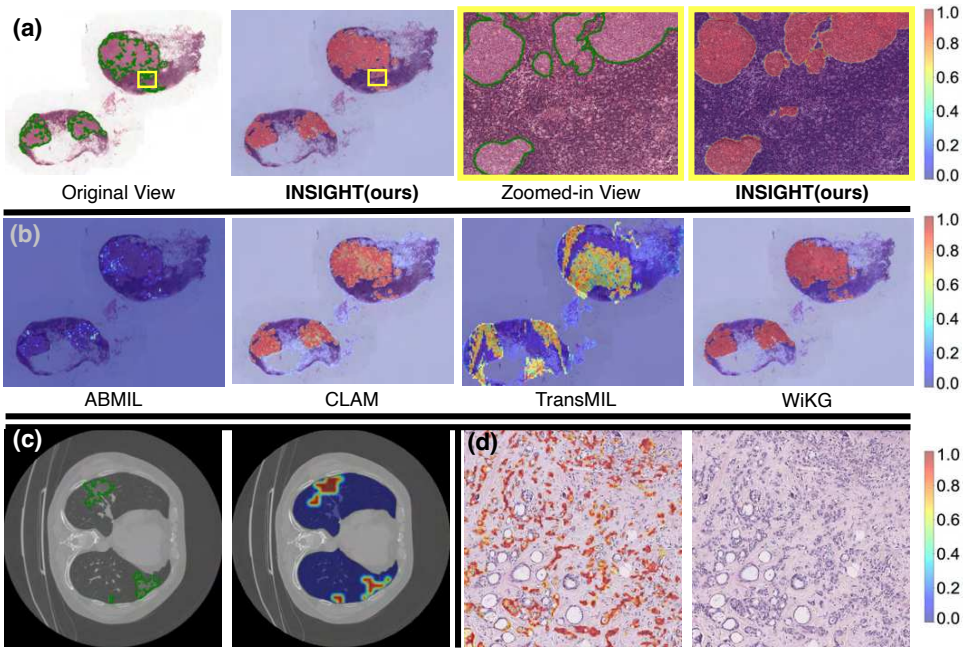

Figure 4: Qualitative comparison of heatmaps generated by INSIGHT (ours) and baseline methods using **UNI**. For CAMELYON16, we compare lesion detection heatmaps generated by (a) our INSIGHT model and (b) the baseline methods. Additionally, we present (c) heatmaps for MosMed and (d) heatmaps for BRACS. Ground-truth areas are outlined in green. Note that BRACS does not provide detailed annotations, so ground truth is not displayed in panel (d). Additional visualizations can be found in the Appendix (Sec. B) for further reference.

- **CLAM:** While CLAM's heatmaps overlap with ground-truth regions, they often exhibit poor calibration, with significant portions of the lesion areas showing low activation values. This leads to suboptimal localization, especially for boundaries.
- **TransMIL:** Heatmaps generated by TransMIL suffer from numerous false positives, marking non-lesion areas as relevant. This overactivation reduces diagnostic precision and makes the results less interpretable for clinical use.
- **WiKG:** Although WiKG captures interactions between patches through dynamic graph representation, its heatmaps occasionally exhibit overactivation in non-lesion areas. This impact boundary delineation and hurt interpretability in clinical applications.

Impact of Foundation Model. To further evaluate the generalizability of INSIGHT across different foundation models, we compare heatmaps generated using **Virchow 2** with those from the default **UNI** backbone. As shown in Fig. 5, both foundational models enable accurate localization of lesion regions; however, Virchow 2 produces slightly more refined boundaries in certain zoomed-in regions, suggesting that foundation model choice

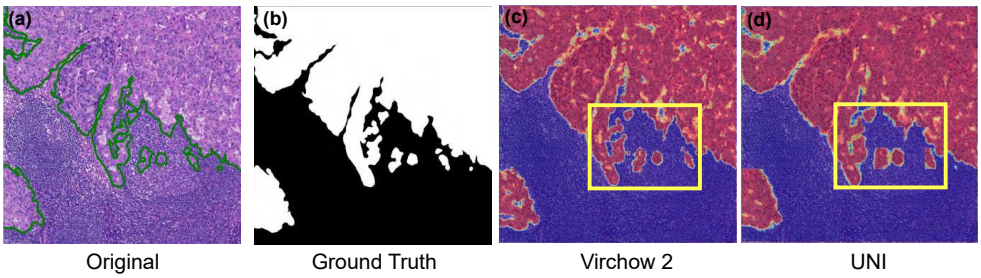

Figure 5: Qualitative comparison of heatmaps generated by INSIGHT using different foundation models. We compare lesion detection heatmaps generated from (a) the original slide, (b) the ground truth in binary (c) INSIGHT with UNI, and (d) INSIGHT with Virchow 2. Yellow box region highlights differences in boundary sharpness and focus.

can affect visual precision. Despite this, INSIGHT's overall interpretability and region alignment remain robust across backbones.

5.4. Comparison with Grad-CAM

To demonstrate the superiority of INSIGHT's built-in heatmaps over post-hoc methods, we compare them with Grad-CAM on the MosMed and CAMELYON16 datasets (Table 5). Unlike Grad-CAM, which requires additional backward passes and is sensitive to noisy gradients, INSIGHT generates interpretable heatmaps directly during forward inference without extra computation. The results show that INSIGHT achieves significantly higher Dice scores across all datasets and foundation models, indicating more accurate and fine-grained localization of diagnostically relevant regions. This intrinsic explainability ensures better alignment with clinical expectations while avoiding the instability and coarse resolution often observed in Grad-CAM.

5.5. Ablation Studies

To validate the effectiveness of INSIGHT's architectural components, we perform an ablation study analyzing three key innovations: context suppression, SmoothMax pooling, and regularization. Results for the CAMELYON16 and BRACS datasets are shown in Table 6. The study evaluates both binary (CAMELYON16) and multi-label (BRACS) tasks, revealing how each component contributes to classification and segmentation performance. Our key findings are the following.

Context Suppression. Adding context suppression (Row 2) improves Dice scores across all datasets, with a notable gain of 16.6 on CAMELYON16. This highlights the module's ability to improve spatial alignment between predictions and ground truth by focusing on relevant regions while suppressing background noise. However, context suppression alone does not maximize AUC, indicating spatial coherence is insufficient for capturing the discriminative features needed for optimal classification.

SmoothMax Pooling. Incorporating SmoothMax pooling alongside context suppression (Row 3) yields a substantial Dice increase of 34.6 on CAMELYON16, emphasizing its role

Table 5: Comparison of Dice scores (mean ± std) between INSIGHT's built-in heatmaps and Grad-CAM. Results are reported on the MosMed and CAMELYON16 datasets.

Dataset	Foundation Model	Built-in Heatmap	Grad-CAM
MosMed	DINOv2	**42.7 ± 15.3**	12.8 ± 13.9
Camelyon16	UNI	**74.6 ± 19.1**	26.4 ± 20.4
Camelyon16	Virchow2	**78.3 ± 15.4**	20.6 ± 27.4

Table 6: Ablation study results on the CAMELYON16 and BRACS datasets. INSIGHT performance is shown with different configurations of CS (Context Suppression), SM (SmoothMax), and Rg (Regularizer).

			CAMELYON16		BRACS			
CS	SM	Rg	AUC	Dice(%)	ADH	FEA	DCIS	Invasive
✗	✗	✗	**0.992**	25.5±16.3	**0.741**	0.509	0.764	0.993
✓	✗	✗	0.982	42.1±21.4	0.733	0.507	0.799	0.998
✓	✓	✗	0.969	**76.7±15.7**	0.728	0.767	0.834	0.999
✓	✓	✓	0.990	74.6±19.1	0.734	**0.790**	**0.837**	**0.999**

in refining lesion boundaries by leveraging all pixel-level activations. This pooling strategy enhances the model's ability to capture nuanced spatial details and boundary precision. However, it slightly reduces AUC for binary tasks, likely due to softened decision boundaries introduced by spatial averaging, which may blur class distinctions. Its ablation ("SmoothMax Removal") reverts to standard max pooling.

Regularization. Adding regularization (Row 4) improves both AUC and Dice across datasets, stabilizing AUC in CAMELYON16 and enhancing robustness in BRACS. Spectral decoupling mitigates overfitting in single-label tasks like CAMELYON16, ensuring broader feature utilization and improved generalization. For multi-label tasks like BRACS, label smoothing particularly benefits smaller classes such as ADH and FEA, improving classification consistency across imbalanced distributions.

Ablation with Virchow2 Backbone. To further evaluate the generalizability of INSIGHT's architectural components, we repeat our ablation study using the Virchow2 (Zimmermann et al., 2024) foundational model. This experiment is conducted on CAMELYON16 to assess how our modules perform under a different pretraining regime. As shown in Table 7, we observe consistent performance gains with the addition of each module. The full INSIGHT configuration, incorporating context suppression, SmoothMax pooling, and regularization, achieves the highest overall performance with an AUC of 0.992 and a Dice score of 78.3%. These results reinforce the robustness of our architecture and demonstrate that INSIGHT's design principles are effective across diverse pretrained backbones.

Table 7: Ablation study of INSIGHT using the Virchow2 foundational model on the Camelyon16 dataset.

			Virchow2	
CS	**SM**	**Rg**	**AUC**	**Dice (%)**
✗	✗	✗	0.968	3.8±3.0
✓	✗	✗	0.988	34.8±21.2
✓	✓	✗	0.990	78.0±17.3
✓	✓	✓	**0.992**	**78.3±15.4**

6. Discussion and Conclusion

Accurately detecting and interpreting disease-related regions in medical imaging remains a critical challenge, particularly when only image-level annotations are available. INSIGHT addresses this gap by achieving state-of-the-art performance in classification and segmentation across CT volumes and WSIs using *only* image-level labels. This capability is transformative, aligning with real-world clinical workflows where physicians rarely annotate at the pixel or voxel level. By enabling use of larger datasets, INSIGHT unlocks potential for scalable, efficient medical imaging solutions. Its architectural innovations—detection and context modules—combine local feature analysis with contextual understanding to produce interpretable, well-calibrated heatmaps. These heatmaps enhance diagnostic transparency and reduce reliance on pixel-level annotations. Such capabilities could significantly accelerate the annotation process (Raciti et al., 2020), allowing clinicians to refine rather than annotate from scratch, streamlining workflows requiring detailed labels.

Limitations. While INSIGHT has been validated on CT and WSI data, its applicability to other modalities, such as MRI and ultrasound, remains unexplored. In principle, the architecture should generalize due to its ability to process spatially embedded features and generate interpretable outputs. However, a key barrier lies in the scarcity of public datasets. Future work will focus on identifying or curating suitable datasets to assess generalizability across broader modalities.

Conclusion. INSIGHT represents a significant advancement in weakly supervised medical imaging, demonstrating strong performance across diverse tasks and modalities. By leveraging image-level labels, embedding interpretability directly into its architecture, and providing fine-grained diagnostic insights, INSIGHT lays the foundation for scalable, clinically relevant AI solutions. Future work will enhance its capabilities with advanced foundation models and extend its reach to additional modalities, broadening its impact on medical imaging research and practice.

Acknowledgments

This work was supported in part by NSF award #2326491. The views and conclusions contained herein are those of the authors and should not be interpreted as the official policies or endorsements of any sponsor. We thank Jhair Gallardo and Shikhar Srivastava for their comments on early drafts.

References

Peter Bankhead, Maurice B Loughrey, José A Fernández, Yvonne Dombrowski, Darragh G McArt, Philip D Dunne, Stephen McQuaid, Ronan T Gray, Liam J Murray, Helen G Coleman, et al. Qupath: Open source software for digital pathology image analysis. *Scientific reports*, 7(1):1–7, 2017.

Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, Maria Gabrani, Florinda Feroce, and Maria Frucci. Bracs: A dataset for breast carcinoma subtyping in h&e histology images, 2021. URL https://arxiv.org/abs/2111.04740.

Keno K. Bressem, Stefan M. Niehues, Bernd Hamm, Marcus R. Makowski, Janis L. Vahldiek, and Lisa C. Adams. 3d u-net for segmentation of covid-19 associated pulmonary infiltrates using transfer learning: State-of-the-art results on affordable hardware, 2021. URL https://arxiv.org/abs/2101.09976.

Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 2019.

Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 2018.

Adam Casson, Siqi Liu, Ran A Godrich, Hamed Aghdam, Brandon Rothrock, Kasper Malfroid, Christopher Kanan, and Thomas Fuchs. Joint breast neoplasm detection and subtyping using multi-resolution network trained on large-scale h&e whole slide images with weak labels. In *Medical Imaging with Deep Learning*, 2024.

Runsheng Chang, Shouliang Qi, Yanan Wu, Qiyuan Song, Yong Yue, Xiaoye Zhang, Yubao Guan, and Wei Qian. Deep multiple instance learning for predicting chemotherapy response in non-small cell lung cancer using pretreatment ct images. *Scientific Reports*, 2022.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, 2017. URL https://arxiv.org/abs/1606.00915.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

Precious Chikontwe, Marcel Luna, Min Kang, Kyung-Soo Hong, Jae-Ho Ahn, and Sang Hyun Park. Dual attention multiple instance learning with unsupervised complementary loss for covid-19 screening. *Medical Image Analysis*, 72:102105, 2021. doi: 10.1016/j.media.2021.102105.

Joseph Paul Cohen, Paul Bertin, and Vincent Frappier. Chester: A web delivered locally computed chest x-ray disease prediction system, 2020. URL https://arxiv.org/abs/1901.11210.

Pierre Courtiol, Eric W. Tramel, Marc Sanselme, and Gilles Wainrib. Classification and disease localization in histopathology using only global labels: A weakly-supervised approach, 2020. URL https://arxiv.org/abs/1802.02212.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019.

Md. Kamrul Hasan, Md. Tasnim Jawad, Kazi Nasim Imtiaz Hasan, Sajal Basak Partha, Md. Masum Al Masba, and Shumit Saha. Covid-19 identification from volumetric chest ct scans using a progressively resized 3d-cnn incorporating segmentation, augmentation, and class-rebalancing, 2021. URL https://arxiv.org/abs/2102.06169.

Julia Hofmanninger, Florian Prayer, Jianning Pan, et al. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European Radiology Experimental*, 4(1):50, 2020. doi: 10.1186/s41747-020-00173-2.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning, 2018. URL https://arxiv.org/abs/1802.04712.

Issam Laradji, Pau Rodriguez, Oscar Manas, Keegan Lensink, Marco Law, Lironne Kurzman, William Parker, David Vazquez, and Derek Nowrouzezahrai. A weakly supervised consistency-based learning method for covid-19 segmentation in ct images. *arXiv preprint arXiv:2007.02180*, 2020.

Jiawen Li, Yuxuan Chen, Hongbo Chu, Qiehe Sun, Tian Guan, Anjia Han, and Yonghong He. Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11323–11332, 2024.

Tiancheng Lin, Yuanfan Guo, Yi Xu, and Chang-Wen Chen. Identifying key factors in multi-instance learning for whole-slide pathological images: Designing assumption-guided feature extractor. *SSRN Electronic Journal*, 2022. doi: 10.2139/ssrn.4230902. Available at SSRN: https://ssrn.com/abstract=4230902 or http://dx.doi.org/10.2139/ssrn.4230902.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Olaf Geessink, Maschenka Balkenhol, Peter Bult, Alma Halilovic, Meyke Hermsen, Ronald van de Loo, Rob Vogels, Quirine F Manson, Nikolas Stathonikos, Alex Baidoshvili, Paul van Diest, Casper Wauters, Michel van Dijk, and Jeroen van der Laak. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. *Gigascience*, 7(6):giy065, Jun 2018. doi: 10.1093/gigascience/giy065.

Xiaoming Liu, Quan Yuan, Yaozong Gao, Kelei He, Shuo Wang, Xiao Tang, Jinshan Tang, and Dinggang Shen. Weakly supervised segmentation of covid19 infection with scribble annotation on ct images. *Pattern Recognition*, 122:108341, 2022. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2021.108341. URL https://www.sciencedirect.com/science/article/pii/S0031320321005215.

Fangfang Lu, Zhihao Zhang, Tianxiang Liu, Chi Tang, Hualin Bai, Guangtao Zhai, Jingjing Chen, and Xiaoxin Wu. A weakly supervised inpainting-based learning method for lung ct image segmentation. *Pattern Recognition*, 144:109861, 2023. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2023.109861. URL https://www.sciencedirect.com/science/article/pii/S0031320323005599.

Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Usman Mahmood, Robik Shrestha, David DB Bates, Lorenzo Mannelli, Giuseppe Corrias, Yusuf Erdi, and Christopher Kanan. Detecting spurious correlations with sanity tests for artificial intelligence guided radiology systems. *Frontiers in Digital Health*, 2021.

S. P. Morozov, A. E. Andreychenko, N. A. Pavlov, A. V. Vladzymyrskyy, N. V. Ledikhova, V. A. Gombolevskiy, I. A. Blokhin, P. B. Gelezhe, A. V. Gonchar, and V. Yu. Chernina. Mosmeddata: Chest ct scans with covid-19 related findings dataset, 2020. URL https://arxiv.org/abs/2005.06465.

Omar S. M. El Nahhas, Marko van Treeck, Georg Wölflein, Michaela Unger, Marta Ligero, Tim Lenz, Sophia J. Wagner, Katherine J. Hewitt, Firas Khader, Sebastian Foersch, Daniel Truhn, and Jakob Nikolas Kather. From whole-slide image to biomarker prediction: A protocol for end-to-end deep learning in computational pathology, 2023. URL https://arxiv.org/abs/2312.10944.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

Pushpak Pati, Guillaume Jaume, Zeineb Ayadi, Kevin Thandiackal, Behzad Bozorgtabar, Maria Gabrani, and Orcun Goksel. Weakly supervised joint whole-slide segmentation and classification in prostate cancer, 2023. URL https://arxiv.org/abs/2301.02933.

Joona Pohjonen, Carolin Stürenberg, Antti Rannikko, Tuomas Mirtti, and Esa Pitkänen. Spectral decoupling for training transferable neural networks in medical imaging. *Iscience*, 2022.

Patricia Raciti, Jillian Sue, Rodrigo Ceballos, Ran Godrich, Jeremy D Kunz, Supriya Kapur, Victor Reuter, Leo Grady, Christopher Kanan, David S Klimstra, et al. Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies. *Modern Pathology*, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

Anindo Saha, Fakrul I. Tushar, Khrystyna Faryna, Vincent M. D'Anniballe, Rui Hou, Maciej A. Mazurowski, Geoffrey D. Rubin, and Joseph Y. Lo. Weakly supervised 3d classification of chest ct using aggregated multi-resolution deep segmentation features. In Horst K. Hahn and Maciej A. Mazurowski, editors, *Medical Imaging 2020: Computer-Aided Diagnosis*. SPIE, March 2020. doi: 10.1117/12.2550857. URL http://dx.doi.org/10.1117/12.2550857.

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. *International Journal of Computer Vision*, 2019.

Sepideh Shabani, Mahdi Homayounfar, Vishal Vardhanabhuti, Mohammad Amin Nikouei Mahani, and Mohammadhadi Koohi-Moghadam. Self-supervised region-aware segmentation of covid-19 ct images using 3d gan and contrastive learning. *Computers in Biology and Medicine*, 149:106033, Oct 2022. doi: 10.1016/j.compbiomed.2022.106033. URL https://doi.org/10.1016/j.compbiomed.2022.106033.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. Transmil: Transformer based correlated multiple instance learning for whole slide image classification, 2021. URL https://arxiv.org/abs/2106.00908.

Robik Shrestha, Kushal Kafle, and Christopher Kanan. An investigation of critical issues in bias mitigation techniques. In *WACV*, 2022a.

Robik Shrestha, Kushal Kafle, and Christopher Kanan. Occamnets: Mitigating dataset bias by favoring simpler hypotheses. In *ECCV*, 2022b.

G. Singh, S.F. Stefenon, and K.C. Yow. The shallowest transparent and interpretable deep neural network for image recognition. *Scientific Reports*, 15:13940, 2025. doi: 10.1038/s41598-025-92945-2.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Ostap Viniavskyi, Mariia Dobko, and Oles Dobosevych. Weakly-supervised segmentation for disease localization in chest x-ray images. In *Artificial Intelligence in Medicine*, pages 249–259. Springer International Publishing, 2020. doi: 10.1007/978-3-030-59137-3_23. URL https://doi.org/10.1007/978-3-030-59137-3_23.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine*, pages 1–12, 2024.

Weiyi Xie, Colin Jacobs, Jean-Paul Charbonnier, and Bram van Ginneken. Dense regression activation maps for lesion segmentation in ct scans of covid-19 patients. *Medical Image Analysis*, 86:102771, 2023. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2023.102771. URL https://www.sciencedirect.com/science/article/pii/S1361841523000324.

Qinghao Ye, Yuan Gao, Weiping Ding, Zhangming Niu, Chengjia Wang, Yinghui Jiang, Minhao Wang, Evandro Fei Fang, Wade Menpes-Smith, Jun Xia, and Guang Yang. Robust weakly supervised learning for covid-19 recognition using multi-center ct images. *Applied Soft Computing*, 116:108291, 2022. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2021.108291. URL https://www.sciencedirect.com/science/article/pii/S1568494621010966.

Xin Zhang, Liangxiu Han, Wenyong Zhu, Liang Sun, and Daoqiang Zhang. An explainable 3d residual self-attention deep neural network for joint atrophy localization and alzheimer's disease diagnosis using structural mri. *IEEE Journal of Biomedical and Health Informatics*, 26(11):5289–5297, November 2022. ISSN 2168-2208. doi: 10.1109/jbhi.2021.3066832. URL http://dx.doi.org/10.1109/JBHI.2021.3066832.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization, 2015. URL https://arxiv.org/abs/1512.04150.

Xiong Zhou, Xianming Liu, Chenyang Wang, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Learning with noisy labels via sparse regularization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 72–81, 2021.

Lingxuan Zhu, Jiahua Pan, Weiming Mou, Longxin Deng, Yinjie Zhu, Yanqing Wang, Gyan Pareek, Elias Hyams, Benedito A Carneiro, Matthew J Hadfield, et al. Harnessing artificial intelligence for prostate cancer management. *Cell Reports Medicine*, 2024.

Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, David Klimstra, Razik Yousfi, Thomas Fuchs, Nicolo Fusi, Siqi Liu, and Kristen Severson. Virchow2: Scaling self-supervised

mixed magnification models in pathology, 2024. URL https://arxiv.org/abs/2408.00738.

Appendix

Appendix A. Implementation Details

Architecture Parameters. INSIGHT uses UNI (Chen et al., 2024) and Virchow (Vorontsov et al., 2024) for WSI tiles, and ViT-L/DINOv2 (Oquab et al., 2024) for CT slices. Preprocessed inputs are resized to 224×224, and embeddings from the last layer of the encoder are concatenated to retain spatial information. Dimensionality is reduced from 1024 to 128 using a linear transformation. Detection and Context modules consist of three convolutional layers with GELU activations, layer normalization, and a final layer for label-specific outputs. The Detection module uses 1×1 convolutions, while the Context module uses 3×3 convolutions.

Training Configuration. INSIGHT is trained using AdamW with a learning rate of 10^{-4} and weight decay of 10^{-4}. Training is capped at 50 epochs, with early stopping triggered after 8 epochs of no validation improvement. We performed grid search to tune hyperparameters on the Camelyon16 validation set, including SmoothMax sharpness α (4, 6, **8**, 10), learning rates ($5e^{-5}$, $\mathbf{1e}^{-4}$, $3e^{-4}$), and spectral decoupling strengths λ (0, **0.01**, 0.05). The selected configuration was then applied to BRACS and MosMed without further tuning, demonstrating INSIGHT's robustness across tasks. To ensure fair comparison, all baseline models were implemented using the original configurations specified in their respective papers.

Appendix B. Qualitative Analysis

We provide additional visualization results (Fig. 6, 7, 8, 9) to further demonstrate INSIGHT's ability to generate well-calibrated heatmaps that align with diagnostically relevant regions in the CAMELYON16 dataset. These examples include results using both UNI and Virchow 2 feature encoders, highlighting the consistency and generalizability of our method across different backbone models. Please zoom in to view finer details.

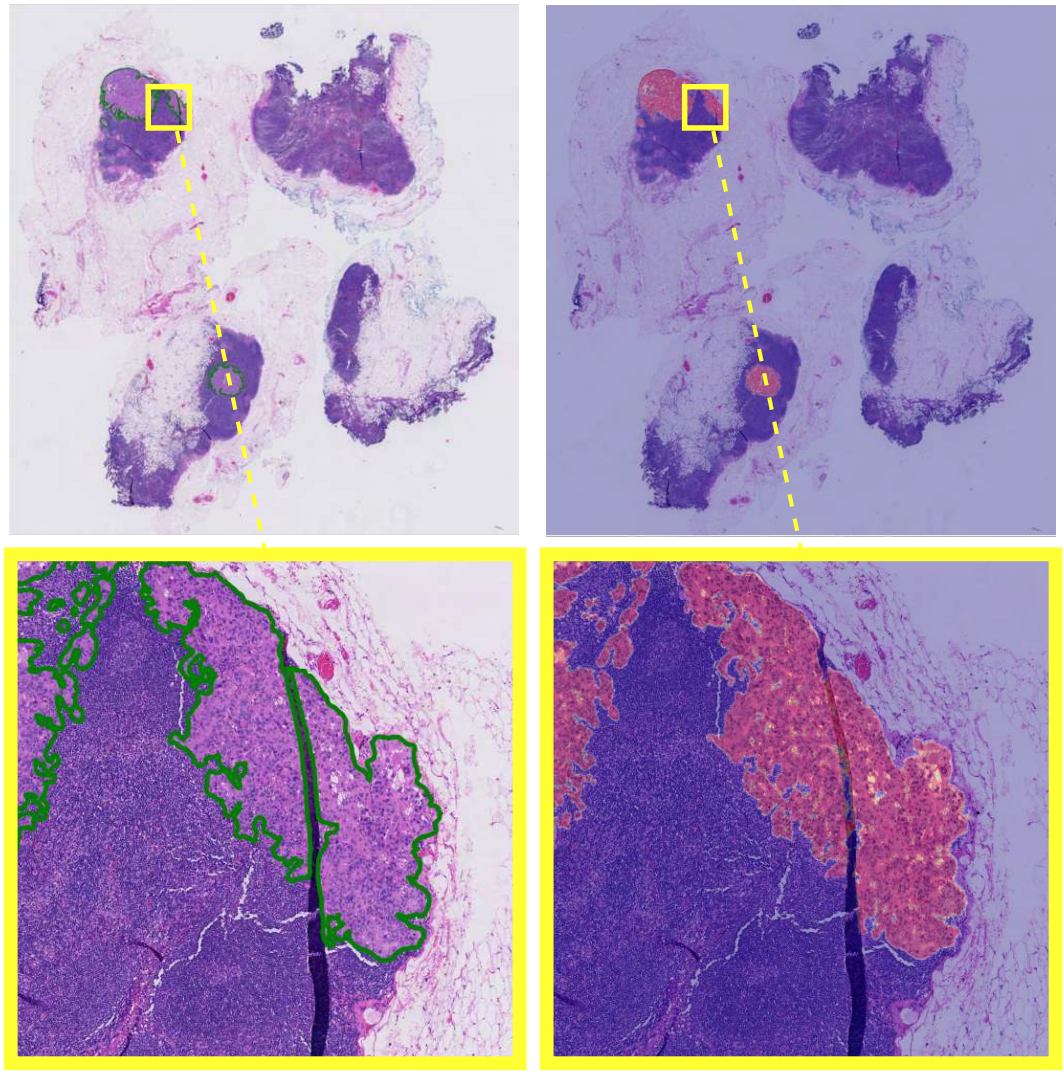

Figure 6: Original histopathology image (left) and the corresponding heatmap generated by INSIGHT using the **Virchow 2** encoder (right). The green outlines on the original image represent ground-truth regions. The bottom row provides zoomed-in views of the regions highlighted by yellow boxes.

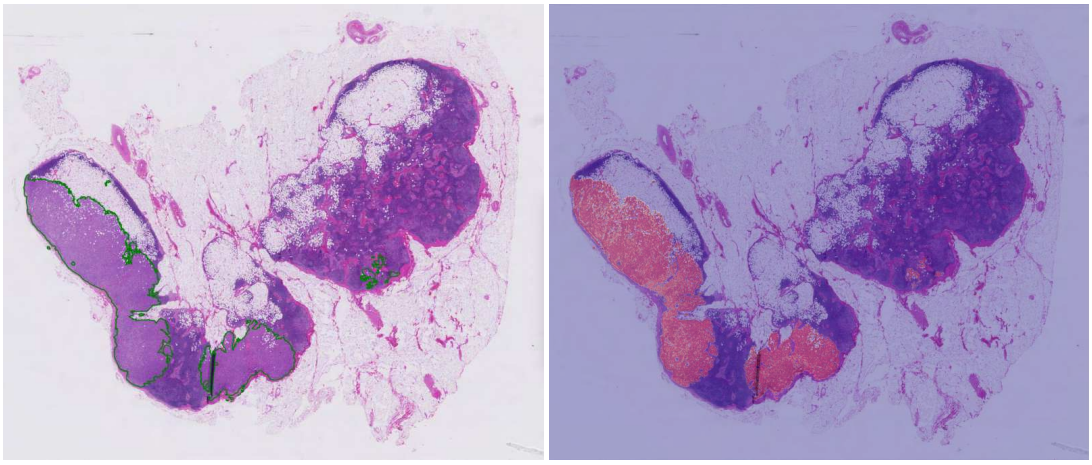

Figure 7: WSI heatmap visualization (right) generated by INSIGHT using **Virchow 2** as the feature extractor on the CAMELYON16 official dataset. The green outlines on the original image (left) represent ground-truth regions.

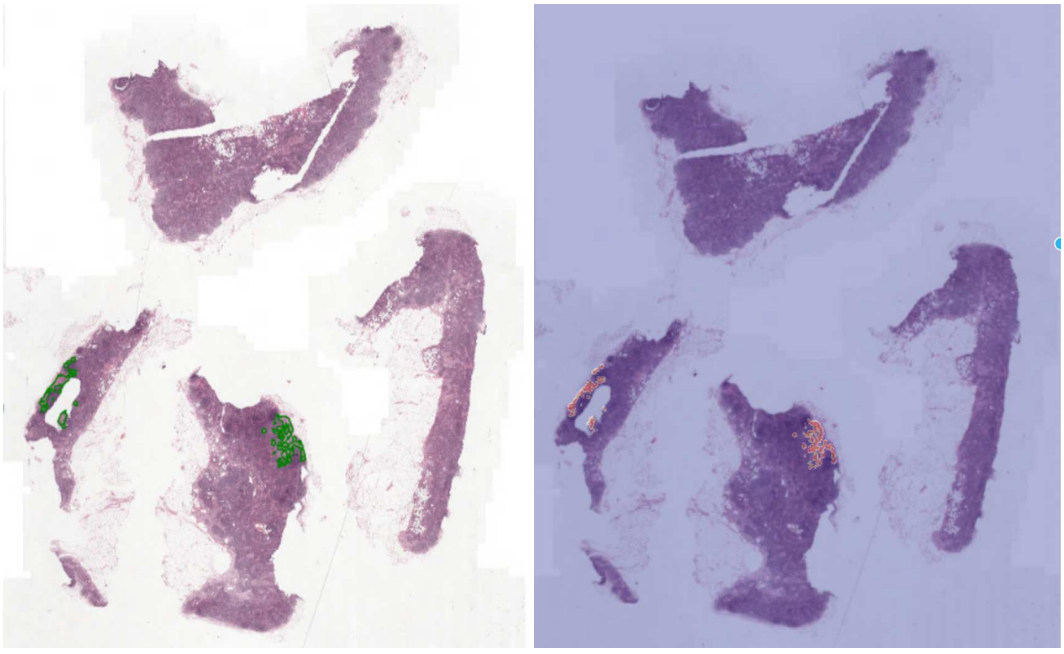

Figure 8: WSI heatmap visualization (right) generated by INSIGHT using **UNI** as the feature extractor on the CAMELYON16 official dataset. The green outlines on the original image (left) represent ground-truth regions.

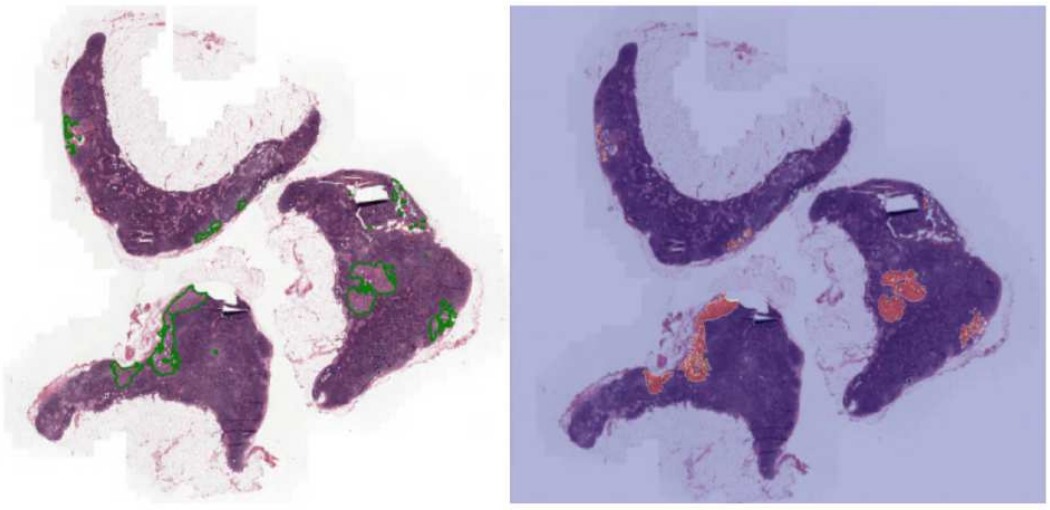

Figure 9: Additional WSI heatmap visualization (right) generated by INSIGHT using **UNI** as the feature extractor on the CAMELYON16 official dataset. The green outlines on the original image (left) represent ground-truth regions.

Appendix C. Runtime Analysis

A key advantage of INSIGHT is native support for multi-label classification. In contrast, baselines like ABMIL and CLAM variants are inherently multi-class only and require running separate models per label for multi-label tasks such as BRACS. This makes INSIGHT significantly more efficient in real-world settings. We report runtime (min/epoch) for training and inference below.

Table 8: Runtime (min/epoch) comparison across datasets and models.

Dataset (task)	**Model**	**Train**	**Inference**
Camelyon16 (binary)	INSIGHT (Ours)	81.8	4.2
	ABMIL	21.2	2.2
	CLAM-SB	19.3	2.2
	CLAM-MB	21.5	2.3
	TransMIL	28.3	2.6
	WiKG	24.7	2.4
BRACS (multilabel)	INSIGHT (Ours)	114.7	6.8
	ABMIL	111.5	12.8
	CLAM-SB	112.8	15.3
	CLAM-MB	121.4	16.6
	TransMIL	132.7	19.2
	WiKG	125.3	16.9

Appendix D. Heatmap Generation Algorithm

Algorithm 1: INSIGHT Heatmap Generation and Classification

Input: WSI tiles or CT slices $\{x_i\}_{i=1}^{N}$ with embeddings $\{e_i\}_{i=1}^{N}$, Detection module f_{Det}, Context module f_{Con}, and pooling sharpness α.

Output: Prediction $\hat{y}$ and reconstructed heatmap $\mathcal{H}_{\text{full}}$.

Step 1: Feature Projection.

for $i = 1$ **to** N **do**

  $e_i \leftarrow \text{Conv}_{1\times1}(e_i)$ `// 1×1 projection to reduce dimension`

end

Step 2: Module Outputs.

for $i = 1$ **to** N **do**

  $\mathcal{H}_{\text{Det},i} \leftarrow f_{\text{Det}}(e_i)$ `// Local detail map`

  $\mathcal{H}_{\text{Con},i} \leftarrow f_{\text{Con}}(e_i)$ `// Context map`

  $\mathcal{H}_i \leftarrow \sigma\big((1 - \sigma(\mathcal{H}_{\text{Con},i})) \odot \mathcal{H}_{\text{Det},i}\big)$

end

Step 3: Patch-to-Image Fusion.

Arrange $\{\mathcal{H}_i\}$ into a full heatmap $\mathcal{H}_{\text{full}}$ according to spatial coordinates.

Apply Otsu threshold $T \leftarrow \arg\min_t \sigma_w^2(t)$.

$\mathcal{H}_{\text{full}} \leftarrow \mathcal{H}_{\text{full}} \cdot \mathbb{I}(\mathcal{H}_{\text{full}} > T)$

Step 4: SmoothMax Pooling.

$\hat{y} \leftarrow \frac{\sum_j \mathcal{H}_{\text{full},j} \cdot \exp(\alpha \cdot \mathcal{H}_{\text{full},j})}{\sum_j \exp(\alpha \cdot \mathcal{H}_{\text{full},j})}$

Step 5: (Multi-label).

If multi-label, repeat Steps 2–4 for each class $c \in \{1, \ldots, C\}$, where C is the number of labels. This produces channel-wise outputs $\hat{y}_c$, resulting in a prediction vector $\hat{\mathbf{y}} \in \mathbb{R}^C$.