MDPI

Article

Evaluating Semantic Segmentation Performance Using DeepLabv3+ with Pretrained ResNet Backbones and Multi-Class Annotations

Matej Spajić [1], Marija Habijan [1], Danijel Marinčić [2] and Irena Galić [1]

[1] J. J. Strossmayer University of Osijek, Faculty of Electrical Engineering, Computer Science and Information Technology Osijek, Croatia
[2] General Hospital Dr. Josip Benčević Slavonski Brod and International Medical Center Priora
* Correspondence: mspajic1@etfos.hr, marija.habijan@ferit.hr

Abstract

Semantic segmentation is a critical task in computer vision, enabling dense classification of image regions. This work investigates the effectiveness of the DeepLabv3+ architecture for binary semantic segmentation using annotated image data. A pretrained ResNet-101 backbone is employed to extract deep features, while Atrous Spatial Pyramid Pooling (ASPP) and a decoder module refine the segmentation outputs. The dataset provides per-image annotations indicating class presence, which are leveraged to approximate segmentation masks for training purposes. Various data augmentation techniques and training strategies were applied to support effective learning and reduce overfitting. Experimental results on the MHIST dataset show that the proposed pipeline achieves strong performance despite the lack of pixel-level annotations, with a mean Intersection-over-Union (mIoU) of 0.76 and a mean Dice coefficient of 0.84. These confirm the potential of weakly supervised segmentation using class-aware CAMs and deep pretrained encoders for structured pixel-level prediction tasks in medical imaging.

Keywords: Artificial Intelligence; DeepLabv3+; Histopatology; Medical Image Analysis; Semantic Segmentation

1. Introduction

Semantic segmentation of medical images is a critical yet challenging task in computer vision, where each pixel of an image must be classified into a meaningful category. In digital pathology, precise segmentation can highlight regions of interest, such as lesions or tissue types, which benefits the diagnosis process. Histopathology involves examining tissue sections under a microscope to identify disease patterns, including tumor margins, inflammatory regions, or morphological subtypes. Accurate delineation of these structures is essential for grading malignancies, assessing surgical margins, and guiding treatment decisions [1]. Obtaining the ground truth per pixel in histopathology is extremely labor intensive since pathologists must carefully outline the regions, which often requires specialized expertise to distinguish subtle morphological features. This annotation process is not only time-consuming but also subject to inter-observer variability and fatigue-related errors. This creates a need for methods that can learn from weaker forms of annotation, such as image-level labels or points, to enable scalable and reproducible segmentation of histological structures [2,3].

Received:
Revised:
Accepted:
Published:

Citation: Spajić, M.; Habijan, M.; Marinčić, D.; Galić, I. . Title. *Journal Not Specified* **2025**, *1*, 0. https://doi.org/

1.1. Related Research

Semantic segmentation has long been studied in computer vision, with Fully Convolutional Networks [4] and encoder-decoder architectures [5] achieving state-of-the-art results when pixel-wise labels are available. In biomedical imaging, the U-Net architecture [6] became a baseline, enabling accurate segmentation of cells and tissues from relatively small datasets. Recent segmentation models such as DeepLabv3+ [7] extend these ideas by incorporating multi-scale context and improved decoding of object boundaries. This architecture's success has carried over to medical image segmentation tasks, especially when combined with transfer learning. For feature extraction, deep residual networks (ResNets) are commonly used backbones [8] and they have been widely adopted in segmentation frameworks as encoders, since they provide rich feature hierarchies learned from large-scale data.

Due to the high cost of obtaining pixel-level annotations in domains like pathology, there is growing interest in weakly supervised semantic segmentation (WSSS) that learns from coarse labels [9]. In WSSS, the model is given labels such as image-level tags or bounding boxes instead of detailed masks. A common approach is to use Multiple Instance Learning (MIL) or global pooling schemes to infer pixel labels from image labels [10]. MIL-based methods treat each image as a bag of instances (pixels or patches) and attempt to identify which regions correspond to the positive class, often using attention or self-training mechanisms. Alternatively, pseudo-labeling approaches generate artificial segmentation masks from the weak labels and then train a segmentation network as if it were fully supervised. The method proposed in this work falls into this latter category.

Class Activation Maps (CAMs) are a widely used technique to obtain coarse localization from image classifiers by highlighting image regions that most strongly influence the class prediction [11]. Prior works have used CAMs as initial seeds for segmentation, but raw CAM outputs are typically diffuse and fail to capture precise object boundaries. This has been observed in natural images and in histology, where CAM-based masks tend to be blobby or incomplete. To mitigate this, researchers have proposed various refinement techniques. For example, smooth pooling and thresholding can tighten CAM maps, and Conditional Random Fields (CRFs) have been used as a post-processing step to sharpen boundaries. More recent CAM improvements include Grad-CAM++ [12], Score-CAM [13] and Smooth Grad-CAM++ [14].

In the medical imaging domain, weakly supervised semantic segmentation has become an active area of research, particularly for tasks such as tumor and lesion detection in histopathology slides. Recent studies have explored multiple instance learning (MIL) frameworks to infer pixel-level labels from slide-level annotations, demonstrating their potential in cancer detection. For example, Campanella et al. [1,2] presented a large-scale MIL approach achieving clinical-grade performance on whole-slide images. Complementary to MIL, pseudo-labeling strategies have also gained traction, where coarse localization maps such as class activation maps (CAMs) are refined into training masks [15]. However, raw CAM outputs often lack boundary precision and can produce incomplete segmentations, motivating the development of advanced refinement techniques like score-based weighting, guided filtering, or DenseCRF post-processing [16]. More recent improvements in explainability, such as Smooth Grad-CAM++ and related gradient-based methods, further enhance localization quality in weakly supervised settings [14]. Unlike more complex pipelines that integrate knowledge distillation or synthetic data augmentation, our method adopts an effective yet straightforward approach: generating refined CAM-based pseudo-masks through morphological operations and DenseCRF, followed by training a DeepLabv3+ segmentation network.

1.2. Research Contributions

In this work, we tackle semantic segmentation with only image-level class labels (no manual masks), using the Minimalist Histopathology Image Analysis Dataset (MHIST) of colorectal polyp images. Each image in MHIST is labeled as either a hyperplastic polyp (HP) or a sessile serrated adenoma (SSA) (HPs are benign, while SSAs are precancerous). Our goal is to segment the regions corresponding to these classes despite not having explicit pixel labels during training. To address this challenge, we adopt a weakly supervised segmentation strategy. We train a ResNet classifier using only image-level labels and employ Class Activation Mapping (CAM) to localize class-specific regions in each image. The coarse CAM heatmaps are refined via morphological post-processing and a dense Conditional Random Field (DenseCRF) to obtain binary pseudo-segmentation masks. These masks are then used to supervise a DeepLabv3+ semantic segmentation model with a pretrained ResNet encoder. DeepLabv3+ leverages atrous spatial pyramid pooling (ASPP) to capture rich multi-scale context and includes a decoder module to recover sharp boundaries, making it well-suited for dense prediction tasks. Our approach produces accurate and interpretable segmentation maps that delineate HP and SSA regions despite never seeing ground-truth masks.

The main contributions of this work are as follows. Firstly, we demonstrate that effective semantic segmentation of histopathology images can be achieved using only slide-level class labels without any pixel-wise annotations, specifically evaluated on the MHIST dataset of colorectal polyp images. Secondly, we leverage Class Activation Mapping (CAM) in combination with morphological filtering and DenseCRF refinement to generate high-quality pseudo-masks that serve as supervision during training. Thirdly, we train a DeepLabv3+ segmentation network equipped with a pretrained ResNet encoder, enabling accurate multi-class segmentation predictions from these pseudo-labels. Finally, our approach achieves strong performance, which demonstrates that structured pixel-level predictions can be reliably learned from weak supervision alone.

2. Method

Our method consists of a three-stage pipeline: (1) train a classification model to obtain class activation maps for each image, (2) post-process and refine these CAMs into pseudo-ground-truth masks, and (3) train a DeepLabv3+ segmentation model on the pseudo-masks to predict pixel-level class labels as shown in Figure 1.

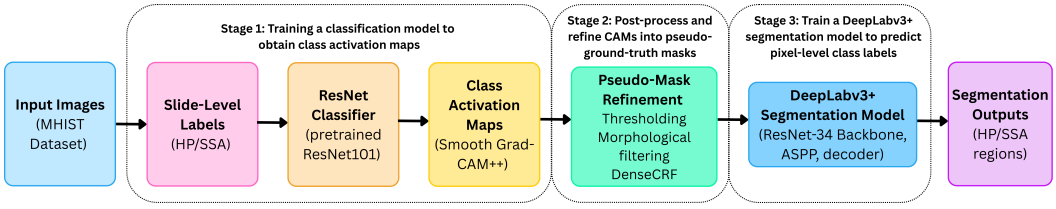

Figure 1. Overview of the proposed weakly supervised segmentation pipeline.

2.1. Data Preparation and Preprocessing

We use the MHIST dataset, which contains 3,152 fixed-size H&E-stained images of colorectal polyps (224×224 pixels). Each image has an image-level label indicating the predominant histologic class: HP or SSA. We split the dataset into training and validation sets (i.e., 80/20) while preserving the class distribution. Before feeding images to the network, we apply standard augmentation and normalization techniques. Color jittering and rotation/flips (via the Albumentations library) help the model generalize to variations in staining and orientation. We also performed stain normalization using the Macenko

method and tissue masking to remove white background, ensuring that color and tissue presence are consistent across samples. This preprocessing is important in histology images to reduce variability that is not related to tissue class.

2.2. Classification Model and CAM Generation

In the first stage, we train a convolutional classifier to recognize the presence of each class in an image. We choose ResNet-101 as the classifier architecture, initialized with ImageNet pretrained weights for transfer learning. ResNet-101 offers a balanced trade-off between depth and complexity, and its residual connections help in training stability. The classifier is trained on a binary task to distinguish between SSA and HP images, where HP images serve as the negative class. We use a combination of binary cross-entropy (BCE) and Focal Tversky loss (with $\alpha = 0.7$, $\beta = 0.3$, $\gamma = 1.33$). The BCE loss handles the main classification objective, while the Focal Tversky component is included to address class imbalance and focus learning on hard-to-classify regions. We train the classifier with Stochastic Gradient Descent (SGD) using a batch size of 32, learning rate of 0.01, and weight decay of 0.01, following a one-cycle learning rate schedule. We apply a progressive layer unfreezing strategy: pretrained layers are frozen for the first 5 epochs, then sequentially unfrozen in stages over a total of 25 epochs to prevent early overfitting. All training parameters, including learning rate, batch size, weight decay, and unfreezing schedule, are empirically optimized. After convergence, we use the trained model to produce Class Activation Maps using Smooth Grad-CAM++, extracting the CAM for the ground-truth class from the final convolutional layer to obtain a heatmap indicating how strongly each image region contributes to the predicted class score.

2.3. Pseudo-Mask Refinement

The raw CAMs often highlight the central polyp regions but may miss boundaries or include scattered false positives. To create usable segmentation masks, we apply post-processing steps. First, we threshold the continuous CAM heatmap at the 0.5 quantile to obtain a binary mask that retains confident activations while discarding background. Next, morphological opening and closing with kernel sizes of 3 and 5 pixels, respectively, remove small isolated blobs and fill internal holes, producing smoother and more contiguous regions. Small-object removal further excludes connected components smaller than 64 pixels, assuming true polyp regions are larger. To refine boundaries and better align them with tissue structures, we apply a Dense Conditional Random Field (DenseCRF) for 5 iterations. The CRF includes a Gaussian pairwise term ($s_{xy} = 3$, compatibility $= 3$) and a bilateral term ($s_{xy} = 80$, $s_{rgb} = 13$, compatibility $= 10$) to encourage spatial and color consistency in the mask. All hyperparameters were empirically optimized to balance segmentation sharpness and noise suppression. The resulting refined pseudo-masks serve as surrogate ground truth for the segmentation network. For each training image, we store the corresponding masks: images labeled as HP use uniform HP masks, while images labeled as SSA include masks that identify SSA regions.

2.4. Segmentation Model Training

In the third stage, we train a DeepLabv3+ segmentation model to predict pixel-wise class labels using the pseudo-masks from Stage 2 as targets. DeepLabv3+ is configured with a ResNet-101 backbone pretrained on ImageNet to leverage strong feature extraction. The model is trained for a binary segmentation task and outputs a single probability map corresponding to the SSA class, treating all other regions (including HP tissue) as background. Since each image contains at most one polyp class, we enforce consistency by attaching a slide-level classification head (global pooling and sigmoid) predicting the SSA presence. We apply a multi-task loss: a binary cross-entropy (BCE) classification

loss, weighted by 0.2, to ensure agreement with the known label, and a segmentation loss combining binary Dice loss and Lovász-Softmax loss, weighted as Dice $+ 0.5 \times$ Lovász. Dice loss optimizes mask overlap and handles class imbalance, while Lovász-Softmax encourages higher Intersection-over-Union, improving segmentation quality over cross-entropy alone. Optimization uses Stochastic Gradient Descent (SGD) with a learning rate of 1×10^{-3}, momentum of 0.9, and weight decay of 1×10^{-3}, along with gradient clipping. As a training strategy, we progressively introduce negative samples: for the first 10 epochs, the model trains only on SSA images, after which HP images are introduced with a sampling probability that increases linearly to a maximum of 0.3 throughout 35 epochs. This curriculum prevents spurious segmentations and improves specificity. All the values for the post-processing steps, such as the minimum object size and CRF parameters, were determined through empirical optimization. We experimented with a range of plausible values and selected those that yielded the highest performance on a held-out validation set. Training continues until the validation metric, Intersection-over-Union, calculated at a fixed prediction threshold of 0.5, stops improving, with checkpointing to save the best model. The final network predicts pixel-wise maps for SSA in new histology images.

3. Results

Before evaluating segmentation, a clinical expert visually reviewed the Smooth Grad-CAM++ maps from the ResNet classifier to confirm they highlighted relevant tissue regions. The generated pseudo-masks were then compared to the initial CAMs and inspected against slide images, with post-processing (morphological filtering and DenseCRF) noticeably improving alignment with glandular structures. Although still approximate, the refined masks offered a much stronger supervisory signal than raw CAMs. Figure 2 shows an example segmentation result for an SSA-labeled image, including the original slide, refined pseudo-mask, model prediction, and overlays, illustrating strong visual agreement between predicted and pseudo-ground truth masks.

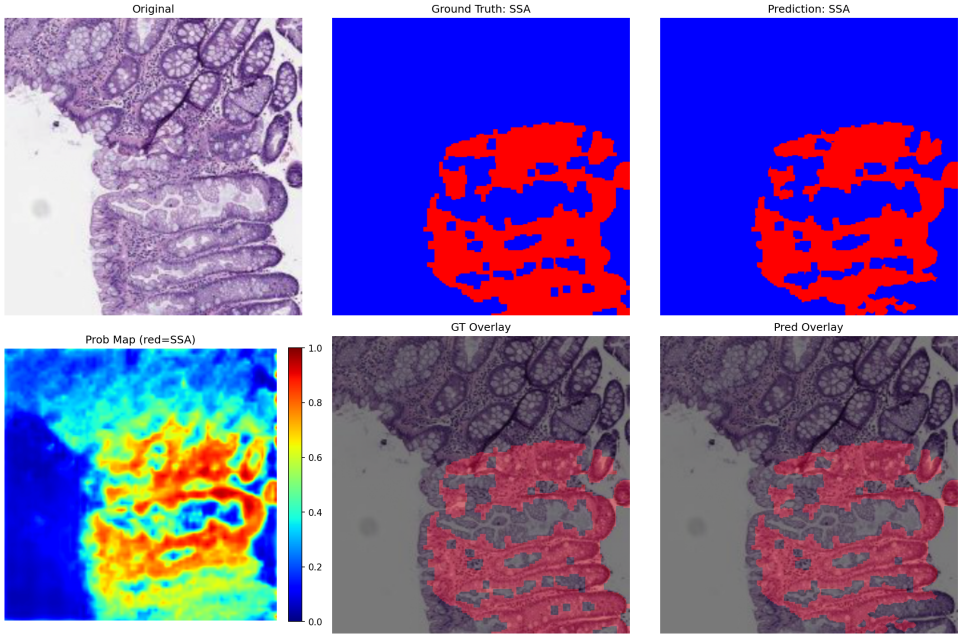

Figure 2. Example segmentation results on an SSA-labeled histopathology image. Top row shows (from left to right): the original H&E-stained image, the refined pseudo-mask used as ground truth, and the model's predicted mask. Bottom row displays the predicted SSA probability map, the pseudo-mask overlaid on the original image, and the predicted segmentation overlaid for visual comparison.

Figure 3. Example error visualization on an SSA-labeled histopathology image, created by overlaying model predictions onto ground truth masks. White indicates true positives, red false positives, blue false negatives, and black true negatives.

To provide a comprehensive overview of model performance, we include the classifier's validation accuracy curve, loss progression, and segmentation IoU metrics. Additionally, we present a series of visualizations: representative segmentation outputs, an error map that overlays true positive, false positive, true negative, and false negative regions to facilitate detailed interpretation of prediction accuracy, and an uncertainty graph that highlights regions of low model confidence. Figure 3 provides an error visualization where correctly and incorrectly classified regions are color-coded, offering insight into model behavior and areas of misclassification.

Table 1 summarizes the key performance metrics of our classification and segmentation pipeline on the MHIST test set. The reported values include the classifier's validation accuracy, final loss, and the segmentation model's Intersection-over-Union (IoU) score. All metrics were automatically tracked and aggregated using the Weights & Biases (WandB) platform, ensuring reliable and reproducible monitoring throughout the training and evaluation process. The training process and all reported results were obtained under fully deterministic conditions by setting fixed random seeds for all sources of stochasticity and by employing deterministic algorithms wherever possible.

Table 1. Performance summary of the proposed method on the MHIST test set.

Metric	Value
Classifier Validation Accuracy	**0.8485**
Classifier Loss	**0.5149**
Segmentation IoU	**0.5774**

Finally, Figure 4 displays uncertainty maps to visualize model confidence, highlighting regions where the segmentation model is indecisive, such as complex tissue boundaries or ambiguous structures. To explicitly address how these were derived, the uncertainty U for each pixel was computed directly from its corresponding output probability p using the formula $U = 1 - 2 \times |p - 0.5|$. This calculation yields maximum uncertainty ($U = 1$) when the model's prediction is most ambiguous ($p \approx 0.5$) and minimum uncertainty ($U = 0$) when the model is most confident ($p \approx 0$ or $p \approx 1$). These maps are valuable for identifying challenging cases and can inform targeted improvements, revealing where additional annotation or supervision might be most beneficial. Together, these visualizations offer insight into the model's predictive behavior and confidence, supporting the viability of the proposed method in a weak supervision setting.

4. Conclusions

In this work, we presented a weakly supervised framework for multi-class semantic segmentation of histopathology images, using only image-level class annotations to train

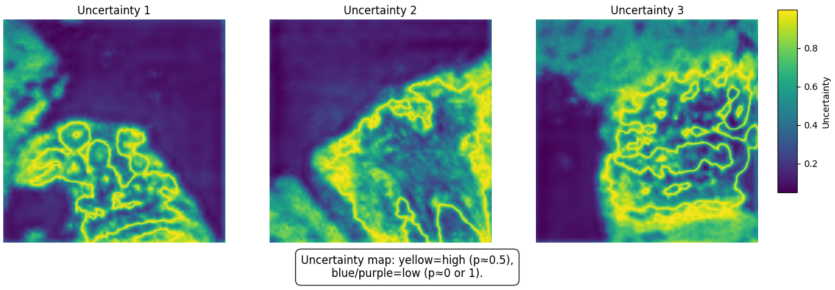

Figure 4. Uncertainty maps that highlight regions where the segmentation model exhibits low confidence, helping to identify challenging or ambiguous cases in the test set.

a segmentation model. By leveraging a ResNet-based classifier to generate CAMs and refining these into pseudo-masks, we trained a DeepLabv3+ segmentation network that accurately delineates colorectal polyp subtypes in H&E images.

The integration of advanced techniques like Smooth Grad-CAM++ for localization, morphological filtering and DenseCRF for mask refinement, and a powerful DeepLabv3+ architecture with specialized loss functions (Dice and Lovász) was key to the success of our approach. Despite the absence of manual pixel labels, the model achieved consistently strong performance on the evaluated dataset. These findings confirm that if suitable priors and training strategies are employed, deep models can perform structured pixel-level predictions in complex image domains like pathology even under weak supervision.

In conclusion, our work demonstrates the feasibility of high-quality segmentation using only multi-class image annotations. This has practical significance: obtaining slide-level diagnoses is far easier than detailed annotations. Therefore, such techniques can unlock segmentation capabilities for many existing labeled datasets. For the MHIST dataset, our results show that one can not only classify but also localize the polyps, potentially aiding pathologists in identifying regions of interest.

In the future, we plan to extend this approach by incorporating additional weak cues (such as point annotations) to further improve mask accuracy. Another avenue is applying our pipeline to other histology tasks (e.g., tumor vs. normal segmentation) and exploring MIL-based pseudo-labels in conjunction with CAMs for even better initial masks. Additionally, we plan to provide more comprehensive quantitative evaluation of the pseudo-masks themselves using metrics such as region overlap and boundary accuracy. Overall, this work shows that weak supervision can nearly match fully supervised performance in medical image segmentation, greatly reducing annotation effort with little loss in accuracy.

Author Contributions: Conceptualization, M.S. and M.H.; methodology, M.S.; software, M.S.; validation, M.S., M.H., D.M; formal analysis, M.S.; writing—original draft preparation, M.S., M.H. and I.G; writing— review and editing, M.S., M.H. and I.G.; visualization, M.S.; supervision, M.H. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: https://github.com/Matej293/mhist-cam-segmentation

Acknowledgments: This research has been partially supported by the Croatian Science Foundation under project number IP-2024-05-9492. The authors thank International Medical Center Priora for the specific types of support provided by the hospital.

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

References

1. Campanella, G.; Hanna, M.G.; Geneslaw, L.; Miraflor, A.; Werneck Krauss Silva, V.; Busam, K.J.; Brogi, E.; Reuter, V.E.; Klimstra, D.S.; Fuchs, T.J. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine* **2019**, *25*, 1301–1309.
2. Lu, M.Y.; Chen, R.J.; Williamson, D.F.; Zhao, M.; Shady, M.; Lipkova, J.; Mahmood, F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering* **2021**, *5*, 555–570.
3. He, Y.; Li, X.; Zemp, R.J. Enhancing Weakly-Supervised Histopathology Image Segmentation with Knowledge Distillation on MIL-Based Pseudo-Labels. *ArXiv* **2024**, *abs/2407.10274*.
4. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440.
5. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2017**, *39*, 2481–2495.
6. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, 2015, pp. 234–241.
7. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2018, pp. 801–818.
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
9. Zhou, Z.H. A brief survey on weakly supervised semantic segmentation. *AI Open* **2020**, *1*, 44–52.
10. Pathak, D.; Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Multi-Class Multiple Instance Learning. In Proceedings of the International Conference on Learning Representations (ICLR) Workshops, 2015.
11. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921–2929.
12. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In Proceedings of the Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 839–847.
13. Wang, H.; Du, Z.; Cui, Z.; Zhu, H.; Chen, H.; Ma, Z. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 110–119.
14. Omeiza, D.; Speakman, S.; Cintas, C.; Weldermariam, K. Smooth Grad-CAM++: An Enhanced Inference Level Visualization Technique for Deep Convolutional Neural Network Models. In Proceedings of the 2019 IEEE International Conference on Computer Vision Workshops (ICCVW). IEEE, 2019, pp. 614–623.
15. Zhao, Y.; Yang, H.; Yan, K.; Liu, Y.; Ma, Z. Multi-scale and multi-level weakly supervised learning for histopathology image segmentation. *IEEE Transactions on Medical Imaging* **2022**, *41*, 573–584.
16. Ahn, J.; Kwak, S. Weakly Supervised Semantic Segmentation Using Pixelwise Localization Networks and Affinity Learning. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5242–5251.