



Multi-class semantic segmentation of oral cancer tissues on whole-slide images: A SuperPixel enhanced deep learning approach

Fabian León^{a,b,c}, Anne Champagnac^d, Mathieu Dupoy^a, Jean-Michel Fayette^d, Tamás Dániel Csúry^{b,c}, Laurent Duraffourg^{a,*,1}, Pierre Saintigny^{b,d,e,*,1}

^a Active Digital Multispectral InfraRed (ADMIR), Moirans, France

^b University of Lyon, Claude Bernard Lyon 1 University, INSERM 1052, CNRS 5286, Centre Léon Bérard, Cancer Research Center of Lyon, Lyon, France

^c CASTING, Cancer Dynamics, Adaptation and Modeling, Inria, Inserm, Ecole Normale Supérieure de Lyon, Centre Léon Bérard, Cnrs, Université Claude Bernard Lyon 1, Lyon, France

^d Department of Biopathology, Centre Léon Bérard, Lyon, France

^e Department of Medical Oncology, Centre Léon Bérard, Lyon, France

ARTICLE INFO

Keywords:

Oral cancer
Histopathological images
Deep learning
Semantic segmentation
Superpixels

ABSTRACT

Oral cancer remains a significant global health challenge, where early and accurate histopathological diagnosis is essential for improving patient outcomes. Whereas deep learning has shown promise in assisting pathologists, current methods face significant challenges by the need for exhaustive manual annotations, class imbalance, and limited contextual insight from patch-based analyses. This work introduces a deep learning method for the automated semantic segmentation of 10 different tissue types within oral cancer tumor regions of interest (ROIs) on whole-slide images (WSIs). To reduce annotation burden and enhance precision, a superpixel-based approach is proposed for data annotation and incorporated a superpixel refinement step during inference to improve boundary delineation while reducing noisy pixels. The model was rigorously evaluated across nine combinations of architectures and backbones, using 444,038 patches from 10 different patients for training and validation (via a 5-fold Leave-One-Fold-Out Group K-Fold Cross-Validation) and 78,103 patches from two independent patients for testing. The present approach achieved a robust average accuracy of 0.921 and micro Intersection over Union (IoU) of 0.859, demonstrating robust results even in under-represented and morphologically diverse tissue classes. Despite challenges related to class imbalance and local context limitations, the proposed methodology consistently provided high-quality predictions. Furthermore, seamless integration into the open-source QuPath platform enables WSI ROI reconstruction, highlighting its potential for real-world clinical deployment in oral cancer diagnostics.

Introduction

Oral cancer remains a significant global health concern around the world, with 389,846 new cases and 188,438 deaths in 2022, representing the sixteenth most common cancer worldwide in the same year.¹ A critical challenge with this disease is the very poor survival rate, mainly associated with late diagnosis in advanced stages (stages III and IV),² which implies aggressive therapies and mutilations that can constitute a remarkable decrease in the patient's quality of life.³ For instance, the 5-year survival rate drops from 80% in early stages of the disease to 20% in stages III and IV.⁴ Within oral cancers, oral squamous cell carcinoma (OSCC) is the most common histopathological subtype, accounting for about 90% of cases.^{5,6} Oral potentially malignant disorders (OPMDs) may precede OSCC. These lesions are described as various mucosal disorders and include

leukoplakia, erythroplakia, oral submucous fibrosis, and oral lichen planus, among others disorders.^{7,8} At the histological level, OPMD may harbor dysplasia at various grades and risk of oral cancer transformation.

In the literature, several methods have been proposed for the early diagnoses of OSCC and OPMDs; however, visual examination and oral palpation remain the most common screening standards.⁹ Whereas other techniques such as autofluorescence imaging, Raman spectroscopy, saliva analysis, and buccal brushing have been explored,¹⁰ their adoption in clinical routines is limited by a lack of robust validation and standardization across larger populations.⁸ Histopathology analysis remains the gold-standard for the diagnosis of OSCC and oral dysplasia. This examination involves the analysis of hematoxylin and eosin (H&E)-stained tissue samples under microscopy by pathologist. Nevertheless, this analysis could be influenced by human biases and totally depend on expert's experience and

* Corresponding authors at: Active Digital Multispectral InfraRed (ADMIR), Moirans, France.

E-mail addresses: laurent.duraffourg@admir-analysis.com (L. Duraffourg), pierre.saintigny@lyon.unicancer.fr (P. Saintigny).

¹ Authors contributed equally and are co-corresponding authors.

expertise to identify and interpreted subtle architectural and cellular alterations.¹¹ For instance, Ranganathan et al.¹² reported only slight agreement (average kappa = 0.139) between six pathologists in the grading of 69 tissue samples of oral epithelial dysplasia following the World Health Organization system. Similarly, Rahul et al.¹³ demonstrated slight agreement (kappa = 0.130) between two general pathologists in assessing resection margins based on the presence/absence and grade of dysplasia. This evidence underscores that support tools may help pathologists in assessing oral cancer specimens.

Artificial intelligence, particularly deep learning algorithms, has emerged in recent years as a promising tool to support clinicians in the diagnosis of oral cancer across a wide range of clinical applications. These methods take advantage of the versatility and capability of deep learning to learn intricate and complex patterns and relationships autonomously, without the need for human intervention. For instance, several studies have explored the use of different technologies, such as photographic images,¹⁴ tomography,¹⁵ and Raman spectroscopy,¹⁶ for the early detection of OSCC and OPMDs. Other works have studied the application of deep learning architectures for classification or segmentation tasks on histological images, including the classification of tissue artifacts,¹⁷ tumor/non-tumor classification,¹⁸ tumor segmentation¹⁹ among other applications.

Despite significant advances in deep learning algorithms and ongoing efforts in the context of oral cancer and OPMD, substantial limitations persist regarding their practical implementation and clinical usability. The time-consuming nature of data annotation remains a major barrier, as deep learning model performance heavily relies on both the quantity and quality of annotations, as well as the number of classes. Furthermore, many existing studies often omit crucial details about test data acquisition or rely on conventional 80%/20% random data splits. Such splits can inadvertently share data distribution, introducing a bias that compromises generalizability to external datasets. Additionally, a prevalent focus on image patch-level analysis in numerous studies limits their practical utility in real-world clinical settings.

To overcome these limitations, this work introduces a novel deep learning architecture, rigorously evaluated using a 5-fold Leave-One-Fold-Out (Group K-Fold) Cross-Validation scheme, that leverages superpixel-based annotation and refinement post-processing, for the automated semantic segmentation of regions of interest (ROIs) within WSIs. The proposed method includes a precisely annotation of 10 different tissue types (i.e., classes), including: epithelial tumor cells, stroma, muscle, necrosis, keratinization, salivary gland, vessel, nerve, blood, and white artifact. The approach was evaluated across multiple combinations of architectures and backbones using two independent WSIs (comprising over 70,000 patches) derived from different patients, achieving an remarkable accuracy of 0.921 and an Intersection over Union (IoU) of 0.859. This methodology advances beyond existing oral cancer segmentation approaches by expanding the classification framework to a more comprehensive set of 10 different tissue types, while also demonstrating robust performance at the ROI reconstruction level. Additionally, the generated predictions were integrated into the open-source QuPath software,²⁰ facilitating their analysis and modification by pathologists, thereby enhancing their usability and potential for clinical application.

Related works

Over the past few years, deep learning has changed how new technologies and methods are applied and integrated within clinical workflows. This revolution allows researchers to focus on the development and refinement of innovative technologies and methodologies, whereas deep learning acts as a powerful tool to unlock their potential applications. For instance, several authors have explored deep learning applications in clinical photography, smartphone-acquired imagery, contrast-enhanced computed tomographies, Raman spectroscopy, optical coherence tomography, and hyperspectral data cubes.^{14,16,21} However, the application of these technologies remains limited, and their validation, often conducted on restricted datasets under specific conditions, obstructs their widespread adoption. In

contrast, the application of deep learning directly to digital slides and whole-slide images (WSIs) has emerged as a major focus to address tissue-level complexities.

The application of deep learning algorithms in the analysis of OSCC histopathological images has been also widely explored for tasks such as classification (the categorization of an input into a defined class), detection (the identification of the spatial coordinates of an object within an input), and segmentation (the individual pixel classification within an input). For instance, Panigrahi et al.¹⁸ proposed an evaluation of different convolutional neural network architectures, including VGG16, VGG19, ResNet50, InceptionV3, and MobileNet, for the binary classification of benign and malignant H&E tissue patches of OSCC at 100× magnification in a set of 4000 patches (20% for test). The best accuracy was achieved by the ResNet50 architecture with 96.6%. Kaori et al.²² proposed similar methodology but incorporated two additional categories: a suspicious class for patches, where pathologists could not establish a definitive diagnosis, and an unknown class for patches exhibiting tissue artifacts such as defocus. In this case, the EfficientNet-B0 architecture was implemented for the classification of 90,059 tissue patches (with a 25% random split for testing) across different magnifications, they reported a top-five cross-validation accuracy of 99%. Despite the promising differentiation between tumor and no tumor H&E patches observed in both studies, there remains a potential concern regarding the mixing of training and testing data distributions, which could introduce significant bias during model training and consequently lead to decreased performance when evaluated on an independent test set from different patients.

Other works have explored the tissue semantic segmentation of OSCC samples, a task that is significantly more challenging than classification tasks, necessitating subsequent pre- and post-processing steps to extract image patches from the WSIs and subsequently reconstruct the predicted patches into the original WSI. For instance, in dos Santos et al.,¹⁹ a U-Net architecture was proposed for the automated segmentation of tumor regions and stroma classes. To do so, 1050 image patches (20% for test) at 20× magnification were randomly extracted from 15 WSIs to train and evaluate the U-Net architecture based on the LAB color representation, which can be more robust to brightness and contrast changes. The authors report an accuracy of 97.6% and an IoU of 85.2% at patch level. An additional step was employed to segment regions in the entire WSI with the respective reconstruction; however, quantitative results for this stage were not provided. Furthermore, the validation was conducted on a limited dataset, and it was not specified whether the training and testing sets were derived from the same WSIs. To address the challenge of data scarcity, Dalí et al.²³ studied the impact of data augmentation techniques during training in the task of image segmentation of tumor areas in OSCC. The authors trained and evaluated a U-Net architecture on two distinct datasets: one comprising 8962 image patches (50% for test) and the other being the dataset proposed by dos Santos et al.¹⁹ The applied transformations included horizontal and vertical flips, rotation, transposition, grid distortion, optical distortion, color transfer, and inpainting. Their findings indicated that these augmentations could yield an increase in accuracy from 82.73 to 84.60% and in IoU from 65.46 to 69.69% for the first dataset, and from 92.44 to 96.99% and from 51.98 to 81.92% for the second dataset, respectively. On the other hand, Muhammad et al.²⁴ extended the histological analysis and proposed FABnet, a deep learning algorithm for the segmentation of micro-vessels and nerves, important prognostic factors associated with cancer recurrence, metastasis, and poor patient survival.²⁵ The authors proposed an encoder-decoder architecture with Xception residual blocks, followed by separable convolutions and atrous spatial pyramid pooling. Following the principles of U-Net, the authors also implemented a form of attention skip connections between the encoder and decoder to enhance low-level image features necessary for the segmentation of small vessels and nerves. The proposed network was tested on 756 patches at 20× magnification extracted from three independent WSIs, obtaining an average IoU of 77.8% and average accuracy of 96.5%. Although high accuracies were reported in the study, the analysis was performed only at patch level and not at the WSI level, which significantly limits its potential application in clinical routines.

Materials and methods

Dataset

Head and neck tissue specimens from patients who underwent surgical resection for oral cancer were collected from the Centre Léon Bérard (Lyon, France), with due consideration for patient privacy and under the approval of the institutional ethics committee. The pathologist selected the test samples based on two criteria: the inclusion of the maximum number of tissue classes and the minimal presence of cautery or scanning artifacts, such as blurring. A total of 12 samples (one sample per patient) were included in this study and digitized using a 3DHISTECH Panoramic 150 digital scanner at 20× magnification with a resolution of 0.325 μm per pixel. The tissue sections were stained following classical French protocols using hematoxylin, phloxine, and saffron (HPS), which helps to distinguish collagen fibers from other connective tissue components, including muscle.

The digital slides were precisely annotated at the highest magnification using the QuPath software by an expert pathologist. Initially, the pathologist identified a ROI surrounding the tumor location. Subsequently, they performed the annotation of 10 different tissue classes within these ROIs, namely: epithelial tumor cells, stroma, muscle, necrosis, keratinization, salivary gland, vessel, nerve, blood, and white artifact. For the stroma class, we included the tumoral stroma compartment, encompassing the extracellular matrix, fibroblasts, and immune cells. Additionally, for the white artifact class, we included adipose tissue due to its strong morphological similarity to this class.

A subsequent patch extraction step yielding patches of size 256 × 256 pixels was performed for training, validation, and testing the models, as further detailed in the “Patch Extraction” section. As shown in Table 1 and Supplementary Table 1, each tissue sample varied in its composition of tissue classes, number of patches, and ROI areas depending on the size of the tumor. For this study, two independent samples (designated as WSI 11 and WSI 12) were strictly held out as an independent patient-level test set. The remaining 10 samples were allocated for model training and validation using a 5-fold Leave-One-Fold-Out (Group K-Fold) cross-validation scheme. To address the fact that not all images contained every tissue class, the folds were carefully selected to ensure maximum representation of all 10 tissue classes across each fold. This setup guaranteed that patches originating from the same patient/WSI were strictly kept within the same fold, preventing any patient-level data leakage between the training and validation sets.

Superpixel-based annotation approach

A critical bottleneck in current artificial intelligence (AI) development is the time-consuming and error-prone process of accurate, pixel-level annotation, especially for semantic segmentation tasks. Expert clinicians often require hours per case²⁶ to perform the meticulous freehand delineation necessary for large histopathological images (up to 100,000 × 100,000

pixels), a method that is susceptible to misclassifications. Furthermore, the inclusion of different classes significantly increases the complexity of this task. To overcome this annotation bottleneck and improve delineation quality, a superpixel-based annotation approach was implemented in this work. This strategy segments images into meaningful regions (superpixels), allowing pathologists to annotate coherent areas rather than individual pixels. This not only accelerates the annotation workflow but also inherently enhances precision by grouping spatially and chromatically similar pixels, thereby reducing the errors commonly associated with freehand pixel-level delineation.

Specifically, we implemented the Simple Linear Iterative Clustering (SLIC) algorithm²⁷ to generate these superpixels. The SLIC algorithm iteratively groups pixels based on their color similarity and spatial proximity, resulting in uniform and compact segments that follow the different histological structures and morphologies with reasonable precision. This method offers a significant advantage over pixel-wise annotation by allowing pathologists to label larger, semantically coherent regions with a single action.

Different superpixel sizes were visually evaluated by experts pathologist to determine the optimal balance between annotation speed and the granularity of the resulting segments. Larger superpixels facilitated faster annotation but could potentially over-segment or merge distinct tissue structures. Conversely, smaller superpixels provided finer boundary detail but increased the annotation time and could reintroduce the challenges associated with pixel-level precision delineation. With these considerations, SLIC superpixels with an average size of 25 μm were selected as the most suitable for this work. This size provided a robust trade-off, enabling efficient annotation while preserving the necessary detail to accurately delineate different histological structures. As illustrated in Fig. 1, pixel-level annotation resulted in several errors and inaccurate regions, whereas superpixel-based annotation yielded more precise and smoother segmentations with a substantial reduction in annotation time.

Patch extraction

Contrary to some conventional segmentation tasks involving natural images, scenes, or autonomous driving, histopathological images require a preprocessing step due to the large size of WSIs, which often exceeds the processing capabilities of current computational architectures. To make the WSIs usable, here is employed two different patch extraction techniques from the annotated ROIs within the WSIs.

For training, as illustrated in Supplementary Fig. 1A, only fully annotated patches were considered, discarding any patch containing even a single unannotated pixel. This strict criterion aimed to train the models exclusively on densely annotated regions. Additionally, is applied a 50% overlap between extracted patches for the following three reasons: to allow data augmentation of training samples, to mitigate boundary artifacts commonly observed in convolutional neural networks, and, as illustrated in Fig. 2, to ensure that complete morphological structures, potentially cut during patch extraction, were captured within at least one patch.

For testing, it is important to cover every pixel within the ROI. As shown in Fig. 1B of the supplementary material, if a patch contained at least one pixel belonging to the ROI, the entire patch was included. This strategy guaranteed that all pixels within the ROI received predictions. Similar to the training phase, a 50% overlap was applied during test patch extraction. Following this process, as detailed in Table 1, a total of 444,038 and 78,103 patches were extracted for training and testing, respectively. All patches were of size 256 × 256 pixels and with a resolution of 0.325 μm per pixel. For validation, the test patch extraction approach was adopted to ensure the prediction of all patches within the ROI.

Data augmentation

Data augmentation is a well-known technique in which different rigid and non-rigid transformations are applied to images, generating new data with slight variations. This process enhances the generalization capabilities of AI models by artificially increasing the number of training samples.

Table 1
Summary of region of interest (ROI), total surface areas, and extracted patch counts per patient sample for the training/validation and independent testing partitions.

Sample	ROI area (mm ²)	Number of patches
WSI 1	120.3	68,204
WSI 2	23.3	12,621
WSI 3	91.6	51,738
WSI 4	17.3	9132
WSI 5	11.2	6093
WSI 6	19.9	10,961
WSI 7	211.8	120,069
WSI 8	34.0	18,985
WSI 9	217.7	123,801
WSI 10	36.5	22,434
Total	783.6	444,038
WSI 11	49.7	30,283
WSI 12	78.1	47,820
Total	127.8	78,103

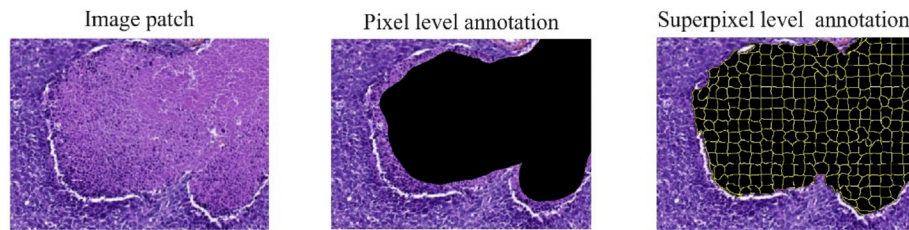


Fig. 1. Comparison between pixel level and superpixel level annotations. Superpixels allow more smooth and fast annotation by grouping meaningful pixels by their color and spatial relationships in one unique cluster.

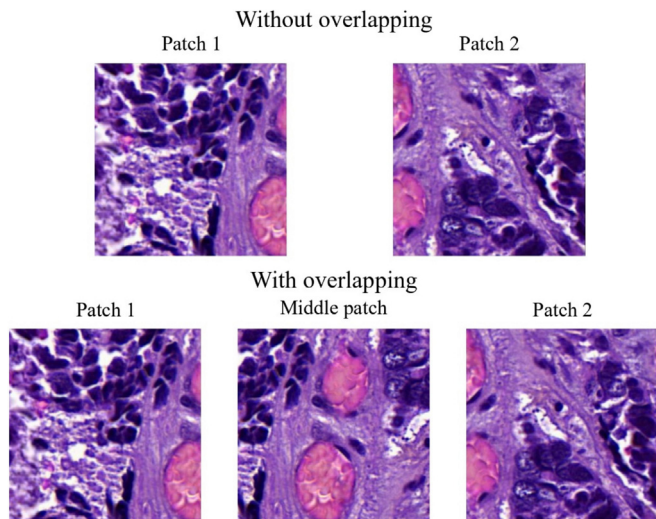


Fig. 2. Patch extraction and overlapping. During the patch extraction process, some biological structures can be cut and not completely captured. With a 50% overlap, it is possible to reduce these cutting issues while simultaneously increasing the number of training patches.

Following the recommendations of David et al.,²⁸ is implemented the following transformations, each with a 50% independent probability of application: vertical and horizontal flipping, scaling (factors between 0.8 and 1.2), rotations (between -10 and 10 degrees), brightness adjustment (ratios between 0.8 and 1.1), contrast adjustment (ratios between 0.7 and 1.3), and Gaussian blur (3×3 kernel, sigma between 0.1 and 1.0). Nearest-neighbor interpolation was used for rotations and scaling when

necessary. Brightness, contrast, and Gaussian blur were specifically applied to mitigate staining and optical variations within the dataset. This augmentation pipeline was implemented online, meaning transformations were applied on-the-fly as data batches were loaded into memory during training.

Image reconstruction

Training and inference process were performed at the patch level. Consequently, a post-processing step was required to reconstruct the predictions into a format that is more interpretable and relevant for pathologists.

As illustrated in Fig. 3, the inference process generated 10 probability maps for each patch, with each map representing the probability of a pixel belonging to a specific tissue class. As described previously, each patch included 50% of overlapping between adjacent patches, these areas were used to merged the probability maps by averaging each overlapping region. This process generated a smoother reconstruction, effectively reducing tile artifacts and boundary effects generated from the convolutional process. Following the merging step, an argmax function was applied to each pixel to assign the tissue class with the highest probability. This process is applied to every patch in the test set until reconstruct the entire ROI.

Ablation study

To evaluate the impact of different network architectures and feature extraction backbones within the proposed superpixel-based annotation and refinement process, an ablation study was performed. As illustrated in Table 2, three of the most common semantic segmentation architectures in medical imaging were tested including:

- U-Net + + : a U-Net-based approach including nested and dense skip connections to facilitate the aggregation of multi-scale features.²⁹

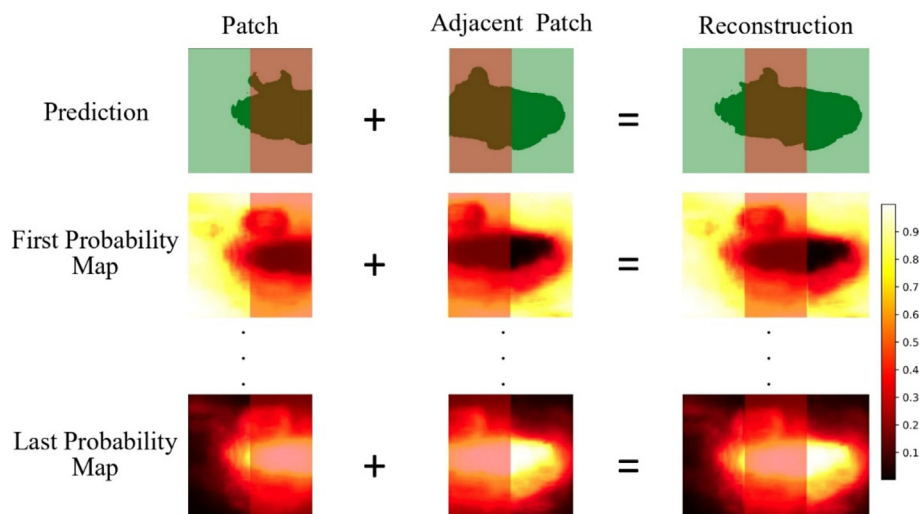


Fig. 3. Whole-slide image reconstruction methodology. Overlapping areas were used to reconstruct patch predictions into original ROIs. First, overlaped probabilities maps were merged by averaging their values and then the class with the highest probability was selected by applying argmax function.

Table 2

Overview of architectures and backbones considered in the ablation study. Three different network architectures and three feature extraction backbones were used to train and test the present approach for a total of nine segmentation models.

Architecture	Description (Key Features)
U-Net + +	U-Net based, nested & dense skip connections for multi-scale feature aggregation.
DeepLabV3 +	Atrous convolutions for multi-scale contextual information without resolution loss.
SegFormer	Transformer-based, designed for global context modeling and efficient segmentation.

Backbone	Description (Key Features)
ResNet101	Residual connections for deeper networks; addresses vanishing gradient.
Xception	Depthwise separable convolutions for efficiency and spatially distinct features.
EfficientNet-B5	Compound scaling (depth, width, resolution) for balanced accuracy and efficiency.

- DeepLabV3 + : an architecture that employs atrous convolutions to capture multi-scale contextual information without losing resolution.³⁰
- SegFormer: a transformer-based architecture designed for global context modeling and efficient segmentation.³¹

These architectures were implemented with three top state-of-the-art feature extractor backbones:

- ResNet101: a backbone that implements residual connections to enable the training of deeper networks and capture complex histological patterns while mitigating vanishing gradient problems.³²
- Xception: an architecture that utilizes depthwise separable convolutions to enhance efficiency and learn distinct spatial features.³³
- EfficientNet-B5: an approach that employs a compound scaling strategy to achieve a balanced trade-off between accuracy and efficiency by optimally adjusting network depth, width, and resolution.³⁴

All possible combinations of these architectures and backbones were systematically trained and evaluated.

Superpixel base refinement approach

Despite the contribution of superpixels during the image annotation step, it is essential to use pixel-level representations during model training

to enable models to learn the intrinsic and complex relationships and patterns from these smaller units. However, as shown in Fig. 4, this level of representation during inference can generate some isolated and misclassified pixels, resulting in a noisy prediction mask. To mitigate this noise and inspired by the superpixel-based annotation process, here is implemented a superpixel refinement. In this step, superpixel polygons were generated with an average size of 25 μm on the HPS ROI and overlaid onto the prediction mask. Pixels within a superpixel are expected to exhibit similar color and spatial relationships, indicating a high likelihood of belonging to the same tissue class. Leveraging this spatial coherence, an argmax function was applied within each superpixel to identify the most frequent class, which was then assigned to all pixels within that superpixel. As illustrated in Fig. 4, this refinement process generated cleaner predictions that more accurately followed the different tissue structures by removing isolated pixels while preserving boundary precision.

Comparison baselines post-processing methods

To evaluate the efficacy of the proposed superpixel refinement method against other common methodologies, the raw model probability predictions were also processed using three widely established smoothing and post-processing techniques. The parameters for all baseline methods were optimized on the validation folds of the most computationally efficient architecture (DeepLabV3 + with an EfficientNet-B5 backbone), where larger kernels and sigma values were found to yield better performance by closely approximating the 25 μm spatial scale of the superpixels. These methodologies include:

- Gaussian Blur: a linear smoothing filter used to reduce high-frequency noise and soften segmentation boundaries. A Gaussian kernel of size 61×61 pixels with $\sigma = 19$ was applied.
- Morphological opening and closing: standard non-linear operations used to clean up the masks. Opening (erosion followed by dilation) was used to remove isolated noise pixels, whereas closing (dilation followed by erosion) was applied to fill small holes within the segmented tissue classes. Both operations utilized a disk-shaped structuring element with a radius of 25 pixels.
- Conditional random field (CRF): a probabilistic graphical model used to model pairwise pixel interactions. The CRF refines boundaries by combining the network's pixel-level class probabilities (unary potentials) with pixel color and spatial proximity features (pairwise potentials) to maximize label consistency across visually homogeneous regions.

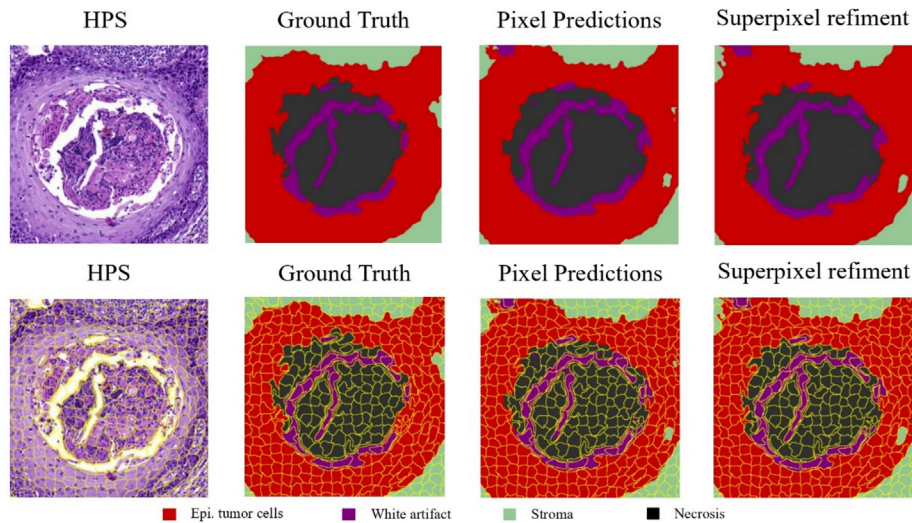


Fig. 4. Superpixel-based refinement and pixel comparison. The top row shows how pixels predictions can generate some isolated misclassified points and how superpixels can mitigate these errors by grouping pixels in meaningful irregular clusters. At the bottom row, superpixel polygons are overlaid into the different images. Polygons do follow the different tissue structures, giving more clear and precise boundary delineations when superpixel refinement was applied as compared to pixel predictions.

Experimental setup and hyperparameters

To ensure a consistent and comparable evaluation across all tested network architectures and backbone combinations, certain training and inference hyperparameters, as well as the computational resources, were fixed. For training, testing, and hyperparameter fine-tuning, the Jean Zay supercomputer was used, equipped with an Intel Xeon Platinum 8468 CPU, 128 GB of DDR5 RAM, and an NVIDIA H100 SXM5 GPU with 80 GB of dedicated memory. For specific tasks, such as measuring the computational time of the different baseline post-processing approaches under a more controlled scenario, a local workstation was used, equipped with an Intel Core i9-14900KF CPU, 64 GB of DDR5 RAM, and an NVIDIA GeForce RTX 4090 GPU with 24 GB of dedicated memory.

Different hyperparameter combinations were tested using the validation folds and the DeepLabV3+ architecture with an EfficientNet-B5 backbone. The best-performing combination was then uniformly applied to all models, which included a batch size of 128 and the Adam optimizer with an initial learning rate of 0.001. Additionally, several loss functions widely used in medical image segmentation were evaluated to address potential class imbalances, including cross-entropy, focal loss, Dice loss, and their combinations (cross-entropy + Dice, and focal + Dice). Because the combination of cross-entropy and Dice loss consistently showed the highest overall IoU and the most stable optimization across the validation folds, it was selected as the definitive objective function for training. A Softmax function was selected as the classification function for inference. Training was performed for a maximum of 15 epochs with an early stopping criterion if no validation loss improvement was observed for five consecutive epochs, as no significant performance improvement was noted beyond this point. Furthermore, all feature extraction backbones (ResNet101, Xception, and EfficientNet-B5) were initialized with pre-trained weights from the ImageNet dataset to leverage prior knowledge and accelerate convergence.

Due to the limited size of the dataset (10 WSIs available for training and validation), a final model was trained using all 10 available samples simultaneously to maximize feature representation before evaluation on the independent holdout test set. To prevent overfitting without a separate validation split, the number of training epochs for this final model was fixed at exactly 10 epochs. This number was selected empirically based on the previous cross-validation loops, where no significant performance or optimization improvements were consistently observed beyond 10 epoch.

Metrics

Accurate evaluation and appropriate metric selection are critical points for precisely measuring the real impact of the different approaches and their interpretation in clinical scenarios. To this end, two different groups of metrics were considered: pixel- and spatial-level metrics.

Pixel-level metrics in semantic segmentation treat each pixel prediction as an independent point in space, without accounting for neighborhood relationships. Accuracy, which measures the overall percentage of correct classifications, and the F1-score, defined as the harmonic mean of precision and recall (Eq. 1) are included within this group. Whereas accuracy can be useful for evaluating individual class performance, it is highly susceptible to misleading results in the presence of class imbalance, especially when evaluating overall performance. The F1-score, by contrast, offers a more balanced performance assessment by considering both false positives and false negatives, making it more robust for overall results. In the context of semantic segmentation, precision for a given class quantifies the proportion of correctly classified pixels for that class out of all pixels predicted by the model as belonging to that class. Conversely, recall for a given class measures the proportion of correctly classified pixels for that class out of all actual pixels belonging to that class in the ground truth. By combining these two, the F1-score provides a comprehensive view, balancing the model's ability to avoid misclassifications (precision) with its ability to find all relevant pixels (recall) for each class.

$$F1\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

Spatial metrics, unlike their pixel-level counterparts, explicitly account for neighborhood relationships and spatial interactions. IoU and the Dice score are included in this group. IoU quantifies the ratio of overlap between predictions and ground truth, calculated as the intersection area divided by the union area (Eq. 2). Similarly, the Dice score also quantifies overlapping areas, but it is calculated by taking twice the intersection area divided by the sum of the individual predicted and ground-truth areas (Eq. 3). Analogous to accuracy, IoU can exhibit sensitivity to classes with low occurrences, whereas the Dice score provides a more robust and balanced evaluation for overall results, particularly in the context of class imbalance.

$$IoU = \frac{|\text{Predicted} \cap \text{Ground Truth}|}{|\text{Predicted} \cup \text{Ground Truth}|} \quad (2)$$

$$\text{Dice Score} = \frac{2 \times |\text{Predicted} \cap \text{Ground Truth}|}{|\text{Predicted}| + |\text{Ground Truth}|} \quad (3)$$

Given the multiclass nature of the segmentation problem, the F1-score, IoU, and Dice score were calculated using their multi-label formulations. To comprehensively assess model performance under the significant class imbalance observed in the dataset, both macro- and micro-averaging strategies were implemented. The micro-average provides an overall measure of pixel level accuracy across all categories, whereas the macro-average evaluates the model's capability across all 10 tissue classes equally, regardless of their size.

Statistical analysis

To evaluate whether the performance differences among the tested segmentation models were statistically significant, a Friedman test was performed. The Friedman test is a non-parametric alternative to the repeated measures ANOVA, utilized here to compare and rank the overall model metrics obtained across the five cross-validation folds. A p -value of less than 0.05 was considered the threshold for statistical significance. This analysis was used to determine if any specific network architecture or backbone combination consistently outperformed the others across the different validation subsets.

Results

As illustrated in Fig. 5, the inference workflow comprised different steps ranging from patch extraction, model inference, image reconstruction, and superpixel refinement to obtaining the delineated ROI prediction mask. To test the present workflow, different architectures and backbones were evaluated on two ROIs within WSIs from different patients. These ROIs varied in tissue area, class distribution, and the number of patches. Pre- and post-processing steps were essential to extract patches from the ROIs and subsequently reconstruct the predicted patches into their original format. All metrics and analyses were conducted at the WSI level, providing a more comprehensive view of the results rather than just a localized patch analysis.

As a remark, throughout the document, the individual results correspond to the DeepLabV3+ architecture with an Xception backbone trained using all 10 available training and validation WSIs. This combination was mainly selected due to the implementation of atrous convolutions, which enhance the field of view, and its faster inference speed compared with U-Net++ and SegFormer.

Post-processing and refinement evaluation

To validate the efficacy of the proposed superpixel refinement strategy, the processed segmentation masks were evaluated against the raw pixel-level probability predictions as well as three well know post-processing

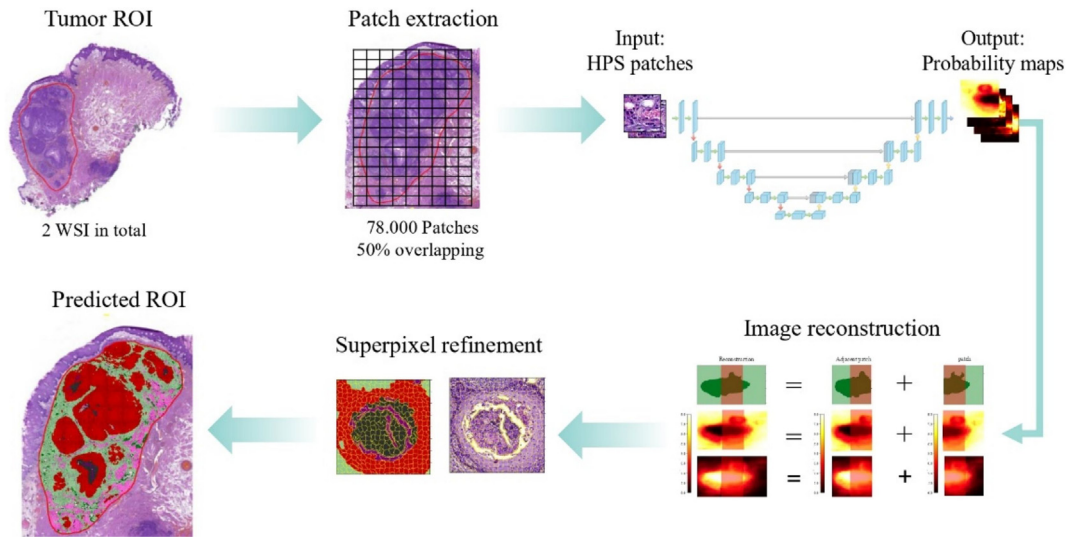


Fig. 5. Inference workflow. The testing workflow includes different steps and processes ranging from patch extraction to superpixel refinement until obtain the complete prediction of the delineated ROI.

methods: Gaussian blur, morphological operations, and conditional random fields (CRFs). As summarized in Table 3, performance was evaluated across multiple metrics using both macro- and micro-averaging schemes.

The quantitative results consistently demonstrate that the proposed superpixel-based refinement yielded superior performance across all evaluated metrics. Compared to the raw pixel-level predictions, the superpixel refinement achieved average absolute gains of 1.6% in micro F1-score, 1.5% in accuracy, 2.5% in micro IoU, and 1.5% in micro Dice score. Notably, the most prominent improvements were observed in the macro-averaged metrics, where macro IoU increased by 3.0% (from 0.661 to 0.691) and macro Dice score advanced by 2.3% (from 0.777 to 0.800). Alternative techniques like Gaussian blur and morphological operations offered marginal smoothing adjustments, whereas the CRF approach showed a modest performance increase over the pixel baseline but fell short of the superpixel method across both macro- and micro-distributions.

Ablation study results

First, to evaluate the training stability of the networks, the mean training and validation loss curves across the five cross-validation folds for the DeepLabV3+ with an EfficientNet-B5 backbone were monitored. As illustrated in Fig. 2 of the supplementary material, the training loss steadily decreased across all 20 epochs, indicating robust convergence. However, the validation loss reached its optimal, lowest plateau between epochs 6 and 9. Beyond the 10th epoch, a visible divergence in validation loss accompanied by an expansion in variance ($\pm$ STD) was observed, signaling the start of overfitting on the training set.

Following stability verification, a comprehensive ablation study was performed to identify the optimal combination of segmentation network architectures and backbones. A total of nine unique configurations derived

from three different architectures (DeepLabV3+, U-Net++ and SegFormer) and three backbone networks (ResNet101, Xception, and EfficientNet-B5) were evaluated. Each configuration was trained using identical optimization hyperparameters across a 5-fold cross-validation totaling 810 h of training time to complete the study.

The validation and test results are shown in Supplementary Table 2 and the test box plots in Fig. 6. All nine architectural combinations showed highly competitive and consistent behavior, with overall performance differences of less than 1.5% across metrics. These uniform results across the diversity of the different models demonstrate that the proposed superpixel schema is decoupled from specific deep learning design choices. Notably, the standard deviations across the validation folds remained consistently tight (ranging from ± 0.001 to ± 0.055), validating the stability and generalization capacity of the pipeline despite the inherently limited dataset size.

Across the validation folds, macro-averaged metrics (with macro F1-scores centering around 68%–70% and macro IoU around 54%–56%) clearly highlighted the impact of high-class imbalance among the 10 different tissue classes. However, when evaluating the models on the independent holdout test set, a consistent performance leap was observed across all architectures, with accuracy clustering closely between 0.914 and 0.919, and macro F1-scores consistently exceeding 0.810. To assess whether any specific architectural combination provided a statistically superior advantage on the unseen holdout data, a Friedman test was performed across the test set metrics. The test revealed no statistically significant differences among the nine models ($p > 0.05$). This uniform top-tier performance on independent data shows that the proposed superpixel annotation and refinement pipeline standardizes model outputs, delivering stable, and architecture-independent segmentation masks.

Final holdout evaluation

To rigorously assess the final capabilities of the proposed approach, a DeepLabV3+ model with an Xception backbone was deployed. The model was trained on the pooled training and validation datasets (comprising 10 WSIs) for a fixed duration of 10 epochs and subsequently evaluated on the two independent holdout images (WSI 11 and WSI 12). The raw pixel-level predictions were compared directly against the superpixel-refined boundaries across both micro- and macro-averaged metrics.

As presented in Table 4, the holdout validation results demonstrate that the superpixel-based refinement consistently yields higher segmentation metrics across both independent cases compared to the raw pixel baseline. Specifically, the superpixel framework achieved an absolute enhancement of 1.5% in global accuracy, 1.6% in micro F1-score, and 2.5% in micro IoU.

Table 3

Segmentation performance evaluation of various post-processing and smoothing methods against the proposed superpixel refinement on the unseen holdout test set.

Post-processing	Accuracy	Macro F1-score	Micro F1-score	Macro IoU	Micro IoU	Macro Dice	Micro Dice
Pixels	0.906	0.777	0.905	0.661	0.834	0.777	0.905
Gaussian blur	0.906	0.779	0.905	0.662	0.834	0.779	0.905
Morpho. operations	0.907	0.777	0.906	0.661	0.834	0.777	0.906
CRF	0.913	0.781	0.911	0.662	0.837	0.781	0.913
SuperPixels	0.921	0.800	0.921	0.691	0.859	0.800	0.921

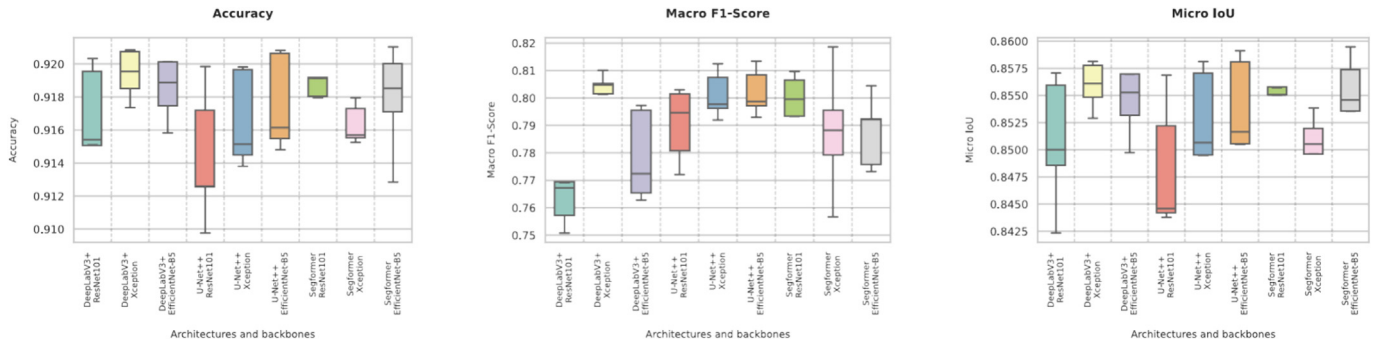


Fig. 6. Architecture models box plots. In general, models show similar results with not statically difference (Friedman test $p > 0.05$). DeepLabV3+ with Xception backbone shows more stability with a tight standard deviation across the different folds.

Table 4

Individual whole-slide image (WSI) performance metrics and overall averages for the raw pixel baseline versus the proposed superpixel refinement strategy on the unseen holdout data.

	Accuracy	Macro F1-score	Micro F1-score	Macro IoU	Micro IoU	Macro Dice	Micro Dice
<i>Pixels</i>							
WSI 11	0.911	0.761	0.911	0.633	0.843	0.761	0.911
WSI 12	0.901	0.794	0.900	0.688	0.824	0.794	0.900
Average	0.906	0.777	0.905	0.661	0.834	0.777	0.905
<i>Superpixels</i>							
WSI 11	0.923	0.785	0.922	0.663	0.862	0.922	0.785
WSI 12	0.920	0.815	0.919	0.719	0.856	0.919	0.815
Average	0.921	0.800	0.921	0.691	0.859	0.921	0.800

Interestingly, the most remarkable improvements were observed within the macro-averaged distributions. The average macro IoU increased by 3.0% (from 0.661 to 0.691), whereas the average macro Dice score improved by 2.3% (from 0.777 to 0.800). This stable baseline correction across unseen patient tissue structures demonstrates that the spatial and color properties provided by the superpixel refinement effectively eliminate local classifier noise.

Class performance results

In addition to spatial and pixel-level metrics, confusion matrices were generated from the two test WSIs to evaluate and compare individual class performance, while also including relevant class metrics such as precision, recall, Dice score, and IoU scores as detailed in Table 3 of the supplementary material. As illustrated in Fig. 7, for WSI 11, the current approach effectively recognizes most tissue structures. However, it shows a notable decrease in performance for vessels and keratinization. This is primarily attributed to the challenge of differentiating the different shapes and sizes of vessel structures (showing an IoU of 0.532), and the difficulty of the model to well differentiate the boundaries between epithelial tumor cells and keratinization. Interestingly, the model accurately predicted the absence of salivary glands in this slide, making no predictions for this class.

For WSI 12, where all classes are present, the model demonstrated significant accuracy for salivary glands (97.15% in the confusion matrix and 0.961 IoU in Table 3 of the supplementary material). A specific area of confusion observed in this WSI was between blood and vessels, where 35.51% of blood pixels were predicted as vessels, likely due to the presence of blood within smaller vessel lumens, which explains the lower recall value of 0.590 for the blood class. Importantly, as the unified confusion matrix in Supplementary Fig. 3 illustrates, the model exhibited high accuracy (up



Fig. 7. Confusion matrix from the two test WSI using the DeepLabV3 architecture and Xception backbone. For WSI 11, a notably decrease was observed for vessel and keratinization classes mainly due to the difficulty of the model to differentiate the limits between some tumor cells and necrotic areas. Similar results were observed in WSI 12 but in this case, a confusion between blood and vessels was noted mainly to be attributed to red blood cells inside small vessels.

to 93%) in differentiating epithelial tumor cells from stroma, underscoring its robustness in distinguishing tumor cells from benign regions a critical point for diagnostic applications.

As illustrated in Fig. 8, specific predicted regions from different classes and locations were extracted to provide a visual assessment. The current approach demonstrates a strong ability to accurately differentiate diverse tissue structures, regardless of their size. This performance, particularly in achieving precise boundary delineations, can be largely attributed to the superpixel refinement process.

Large tissue structures, such as muscle, epithelial tumor cells, salivary glands, and stroma, consistently showed notable segmentation results, with accuracy and Dice scores exceeding 0.900 due to the relatively high prevalence of these classes within the training dataset. Furthermore, despite the limited presence of the nerve class in the training set, the model effectively recognized nerve structures (recall of 0.860 and IoU of 0.647), even when they exhibited cutting artifacts as observed in Fig. 8. Conversely, keratinization presented a major challenge, showing the lowest IoU of 0.320 because it was frequently confused with tumor cells (33.4% in WSI 11) or necrosis (46.93% in WSI 12). Another notable observation is the model's capacity to accurately identify white artifacts (Dice score of 0.800), including small regions that were occasionally overlooked during the manual annotation process. An example of the full ROI prediction is provided in Fig. 4 of the supplementary material.

Loss function evaluation

To determine the optimal objective function for the proposed approach under the high class imbalance, an evaluation was performed across five different loss formulations: cross-entropy (CE), CE + Dice, Dice, Focal, and Focal + Dice. As illustrated in Supplementary Fig. 5, pure Dice loss demonstrated the lowest overall performance, particularly for the macro F1-score which was around 68%. Whereas Focal loss and Focal + Dice showed competitive metrics, they exhibited a wider performance variance across the folds, showing less stability.

Conversely, the combined CE + Dice loss consistently demonstrated the highest results and the tightest variance distributions across all metrics, achieving an accuracy of 0.921, a macro F1-score of 0.810, and a micro IoU of 0.857. This behavior suggested that combining the pixel-level stability of CE with the overlap optimization of Dice loss provides a robust training strategy for these data, justifying its selection as the objective function for the current approach.

Computational execution time analysis

To evaluate the computational efficiency of the proposed approach compared to the alternative post-processing methods, a time measurement experiment was performed. The average execution time per WSI

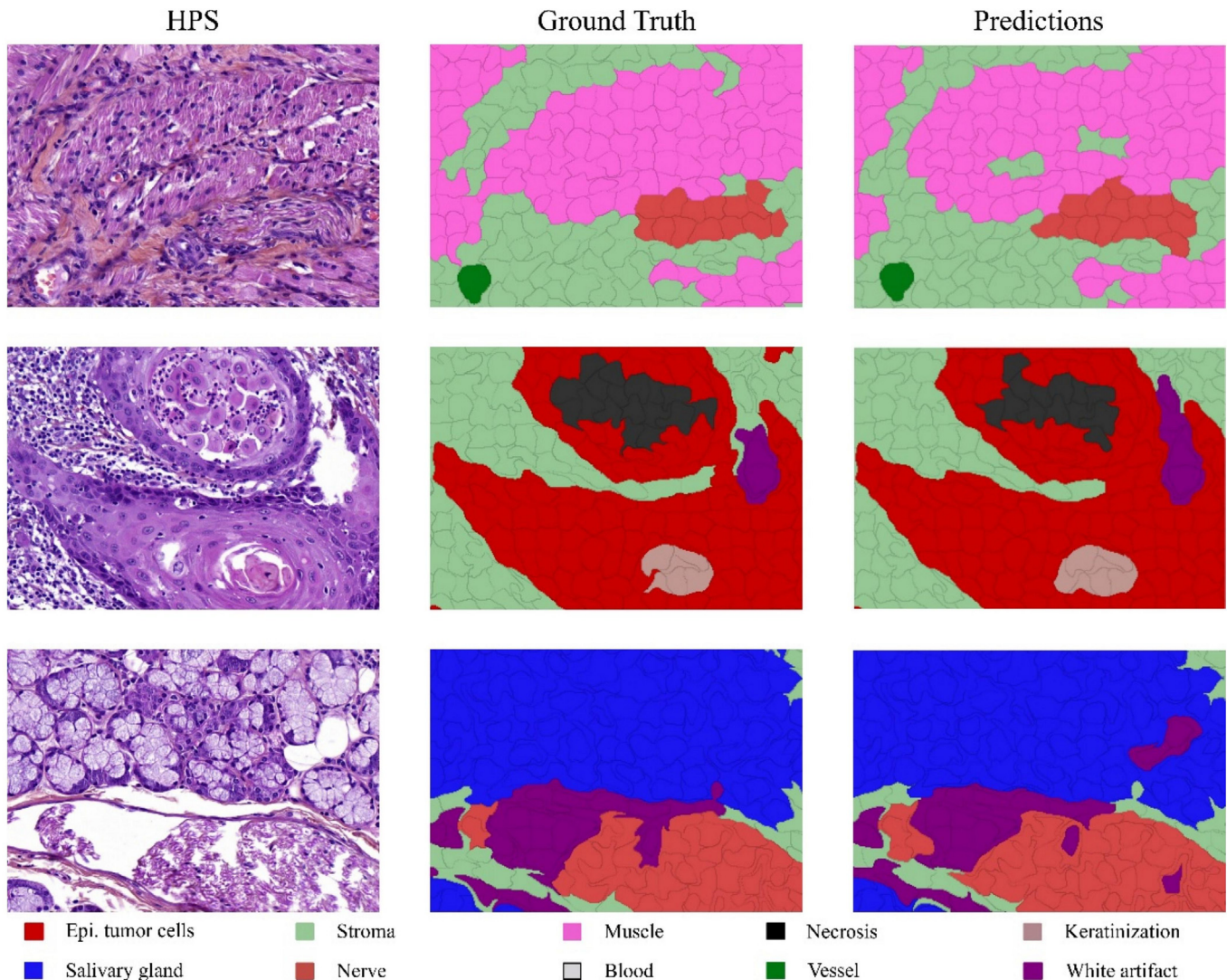


Fig. 8. DeepLabV3 backbone and Xception architecture segmentations of different areas. In general, model's predictions show considerable ability to segmented the different tissue classes independent of their size and location.

Table 5

Computational time execution for the different baseline post-processing methods. In general, reconstruction + patch extraction takes around 5 min, whereas the proposed superpixel refinement takes 3.5 min just 1 min slower compare to opening and closing morphological operations.

Task	Average execution time (minutes)
Patch extraction	0.2
Prediction + reconstruction	5
Morphological operation	2.5
Gaussian blur	7.5
CRF	60
Superpixels	3.5

was measured for each pipeline step. As summarized in Table 5, preprocessing tasks, including patch extraction, required only 0.2 min, whereas the model prediction and spatial reconstruction step took 5 min per slide.

When comparing the post-processing refinement methods, significant differences in processing time were observed. Morphological operations were the fastest, requiring 2.5 min, followed closely by the proposed superpixels refinement, which took 3.5 min per slide. In contrast, Gaussian blur required 7.5 min, and CRF method demonstrated an extreme computational bottleneck, taking 60 min to complete a single WSI. This behavior suggested that while CRFs are computationally heavy and impractical for large-scale digital pathology workflows, the proposed superpixel refinement provides a highly efficient and fast alternative, delivering excellent boundary precision with minimal computational overhead.

QuPath integration

Finally, to significantly enhance visualization and model usability for clinical application, the model's predictions were integrated into the QuPath software.²⁰ This integration involved several steps: first, the superpixel predictions were converted into polygons, including their absolute coordinates relative to the HPS slide. Next, this geometric information, along with other parameters such as class, color, and type, was formatted into a GeoJSON file to ensure compatibility with QuPath's specifications.

As Fig. 9 illustrates, this integration provides more than just a visual interface to easily superimpose and switch between the original HPS sample and the model's predictions. It also empowers pathologists to directly modify and refine the model's segmentations using QuPath's extensive suite of built-in tools. Furthermore, QuPath's capabilities extend to incorporating different stains, performing various nuclei segmentations, and offering numerous other functionalities, maximizing the overall utility of the predictions in a clinical research environment.

Discussion

This work introduces an automated semantic segmentation algorithm for oral cancer of 10 different tissue classes within tumor ROIs in HPS slides. The evaluation was carried out in a set of 9 different architectures and backbones using in a set of 444,038 training/validation patches and 78,103 test patches with size of 256 × 256 pixels, derived from 12 different patients. The study included the use of superpixels to enhance and speed up the annotation process allowing pathologist to efficiently make fine-grained annotation at a higher resolution (0.325 μm pixel size). This level of precision was crucial for achieving accurate boundary segmentation,

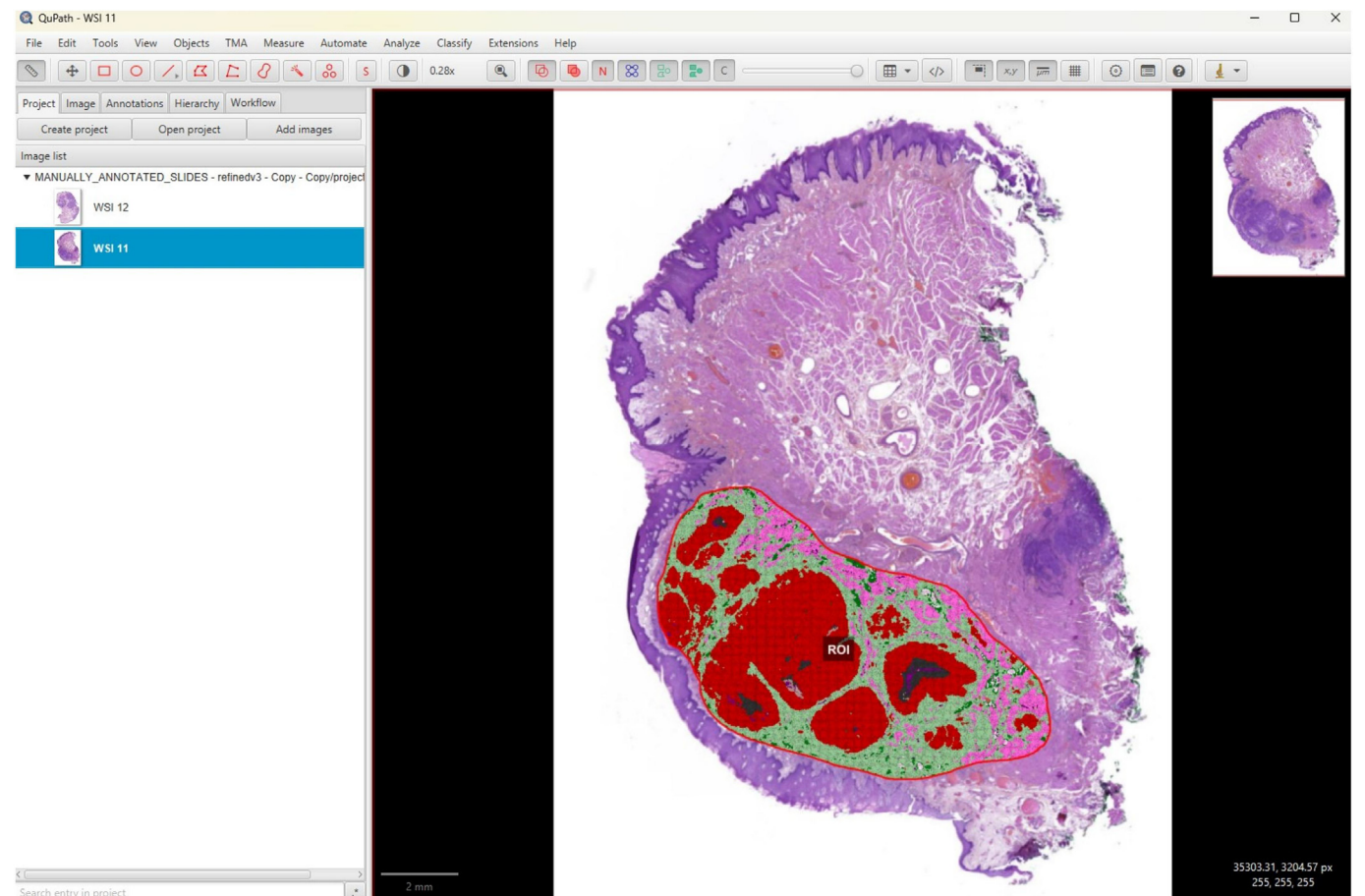


Fig. 9. QuPath software interface. Model predictions are imported into QuPath, allowing clinicians to use all the software's built-in tools and functions over the predictions.

evidenced by a micro IoU of up to 0.859 and an accuracy of up to 0.921. Additionally, no statistically significant differences (Friedman test, $p > 0.05$) were found across the different architecture/backbone combinations, with remaining models showing less than 2% performance difference, underscoring the robustness and consistency of the entire workflow. Furthermore, the present superpixel-based refinement strategy significantly improved performance, yielding an increase of 3.0% in macro IoU and 2.3% in macro Dice score compared to pixel-level predictions, by effectively removing small, isolated pixels and providing superior boundary delineations through grouping pixel predictions by their spatial and color relationships.

The clear advantage of the proposed superpixel refinement over traditional post-processing baselines is attributed to its capacity to leverage the inherent spatial and color coherence of tissue morphologies. Whereas alternative techniques like Gaussian blur and morphological operations smooth predictions, they lack mechanisms to handle complex boundaries, often erasing smaller tissue structures. Whereas scaling the baseline filters (such as Gaussian blur and morphological operations) to a 25 μm kernel size successfully removed isolated noise, it inherently caused a loss of fine structural details and blurred class boundaries. In contrast, the proposed superpixel refinement achieved the same level of noise reduction at that scale while strictly preserving sharp, biologically accurate boundaries.

When considering the balance between accuracy and computational speed, the proposed superpixels refinement demonstrated a highly favorable tradeoff. Although it required an additional 1 min of processing time compared to simple morphological operations (3.5 min vs. 2.5 min), this minor increase was heavily justified by the substantial improvements in boundary precision and macro-averaged metrics. Furthermore, it completely bypassed the extreme 60-min computational bottleneck seen with CRFs, showing that the current approach is both highly accurate and fast enough for practical, large-scale digital pathology workflows taking in average 9 min to process one WSI.

Despite the considerable results obtained, some issues still persist that warrant discussion. A major challenge in the current approach arises from the significant class imbalance observed in both the training and test sets. For example, as detailed in Supplementary Table 1, classes with relatively low occurrences in the training set such as nerve (0.84%), blood (0.57%), and vessels (2.36%) show a degradation in performance in both test WSIs compared to classes with higher occurrences. This can be primarily attributed to two factors: the high morphological variation in shape and size within these tissue classes, and the insufficient training data to learn such diverse representations. It is important to note, however, that even with this high class imbalance, the model is still able to identify these challenging structures, achieving a robust recall of 0.860 for nerve, a precision of 0.910 for blood, and a balanced performance for vessels with an IoU of 0.532.

A notable observation in the semantic segmentation results is the overlap between the epithelial tumor cells and keratinization classes. Whereas this might initially seem like a misclassification, it reflects a biologically accurate representation of OSCC pathology. Keratinization is the definitive sign of squamous differentiation and is inherently part of the malignant cells in OSCC. Therefore, the model's ability to identify and classify keratinized regions as part of the epithelial tumor cells class is not an error; instead, it demonstrates the model's capacity to recognize a fundamental histological characteristic of well-differentiated tumors.

Another significant challenge arises from current computational limitations that necessitate working with local patches, thereby limiting the field of view compared to the vast size of WSI ROIs. This loss of contextual information makes accurate segmentation of large structures difficult, forcing models to rely on local color and texture patterns for such patches. Consequently, this dependence on local visual features highlights a potential limitation regarding stain generalization, as the model was trained on HPS-stained tissues and may face domain shifts if applied to H&E stains due to alterations in color and texture patterns. For instance, as Figs. 6A and 6B of the supplementary material demonstrate, some structures are misclassified mainly due to their local appearance similarity with other

tissue types. A similar issue is observed in Supplementary Fig. 6C, where large white areas within vessel lumens are misclassified as white artifacts. This particular misclassification highlights the critical loss of spatial context from distant but related structures, such as endothelial cells, which would otherwise help in correctly identifying these lumens as part of the vessel.

In the literature, several works have explored the application of deep learning models to oral cancer analysis in histopathology images. Some studies focus on the classification of tumor versus benign patches,¹⁸ or include additional classes for areas that are diagnostically ambiguous or contain tissue artifacts.²² However, these methods typically offer only local analyses due to their patch-based nature, which significantly limits their clinical applicability and usability by providing restricted patch interpretability.

Other studies applied deep learning for semantic segmentation for different tasks, such as the segmentation of tumor areas,¹⁹ reporting a patch accuracy of 0.976 and an IoU of 0.852, but only offering qualitative assessments for the WSI segmentation. Similarly, works exploring the segmentation of micro-vessels and nerves report high accuracy (0.965) and IoU (0.780), but again, these analyses are restricted to a limited number of classes and remain patch-based.²⁴

Approaching these critical limitations in the tissue segmentation of oral cancer specimens, the present work not only significantly surpasses current approaches by considering a comprehensive set of 10 tissue classes for semantic segmentation but also demonstrates robust inference across different patients through quantitative analysis of two independent WSIs. Furthermore, the WSI ROI reconstruction and deployment in QuPath significantly enhances its potential for future clinical impact and usability in clinical scenarios.

Conclusions and future work

This work introduces an automated semantic segmentation algorithm for oral cancer, specifically designed to segment 10 different tissue classes within ROIs on WSIs. The present approach demonstrated the methodology's robustness through an extensive evaluation across nine unique architecture and backbone combinations, using a considerable dataset from 12 independent patients. A key innovation of the approach is the integration of superpixels for both efficient annotation and effective post-prediction refinement, which provided accurate boundary segmentations while mitigating possible noise predictions. This strategy yielded considerable results, including a micro IoU of up to 0.859 and an accuracy of up to 0.921, consistently across different model configurations.

Despite the challenges posed by significant class imbalance and the inherent limitations of patch-based analysis in capturing global WSI context, the present approach still demonstrated a significant ability to identify even morphologically diverse and under-represented classes. Crucially, the full WSI ROI reconstruction and integration of the predictions into the open-source QuPath software significantly enhance the clinical usability and potential for direct translation into diagnostic workflows. Future work will focus on addressing current limitations by exploring techniques to handle extreme class imbalance and the inclusion of complementary information, such as the biochemical composition of the tissue from infrared imaging, to enrich feature representation beyond color and texture features. This new infrared information will help to mitigate the limited field of view by adding the biochemical representation of each tissue class instead of only relying on their morphological representation limited by the patch size. Furthermore, prospective validation studies in diverse clinical settings and tasks are expected to assess generalizability and potential clinical applications.

Funding

The study was funded by LYRICAN + (INCA-DGOS-INSERM-ITMO cancer_18003) to PS, by Agence National pour la Recherche (TUMADMIR N°ANR-25-LCV1-0005) to PS and LD, and by Plan France 2030 AuvergneRhône-Alpes region through the "Collaborative Projects I-

Demo" (MIRFLECT N°DOS0267087/00) to PS and LD. The authors would like to thank the French National Association for Research and Technology (ANRT) for funding the PhD thesis of Mr. Fabian León (grant No. 2023/1224).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 20XX-AD010316592 made by GENCI.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jpi.2026.100699>.

References

- Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2024;74(3):229–263.
- Warnakulasuriya S. Global epidemiology of oral and oropharyngeal cancer. *Oral Oncol* 2009;45(4–5):309–316.
- González-Ruiz I, Ramos-García P, Ruiz-Ávila I, González-Moles MÁ. Early diagnosis of oral cancer: a complex polyhedral problem with a difficult solution. *Cancers* 2023;15(13):3270.
- Sinevici N, O'sullivan J. Oral cancer: deregulated molecular events and their use as biomarkers. *Oral Oncol* 2016;61:12–18.
- Lewin F, Norell SE, Johansson H, et al. Smoking tobacco, oral snuff, and alcohol in the etiology of squamous cell carcinoma of the head and neck: a population-based case-referent study in Sweden. *Cancer Interdisc Int J Am Cancer Soci* 1998;82(7):1367–1375.
- González-Moles MÁ, Aguilar-Ruiz M, Ramos-García P. Challenges in the early diagnosis of oral cancer, evidence gaps and strategies for improvement: a scoping review of systematic reviews. *Cancers* 2022;14(19):4967.
- Sarode SC, Sarode GS, Tupkari JV. Oral potentially malignant disorders: a proposal for terminology and definition with review of literature. *J Oral Maxillofac Pathol* 2014;18(suppl 1):S77–S80.
- Bouvard V, Nethan ST, Singh D, et al. IARC perspective on oral cancer prevention. *N Engl J Med* 2022;387(21):1999–2005.
- Sankaranarayanan R, Ramadas K, Amarasinghe H, Subramanian S, Johnson N. Oral cancer: prevention, early detection, and treatment. *Cancer: Disease Control Priorities*. 3rd ed. Washington, DC: The International Bank for Reconstruction and Development/The World Bank; 2015. p. 85–99.
- Warnakulasuriya S. Oral potentially malignant disorders: a comprehensive review on clinical aspects and management. *Oral Oncol* 2020;102: 104550.
- Zohdy M, Cazzaniga S, Nievergelt H, et al. Inter-observer and intra-observer variations in the assessment of epithelial dysplasia in oral lichenoid diseases. *Dermatopathology* 2021;8(2):84–88.
- Ranganathan K, Kavitha L, Sharada P, et al. Intra-observer and inter-observer variability in two grading systems for oral epithelial dysplasia: a multi-centre study in India. *J Oral Pathol Med* 2020;49(9):948–955.
- Mohandas R, Ramani P, Mohapatra S. Interobserver audit of surgical margins in patients with recurrence of oral squamous cell carcinoma– a retrospective study. *Oral Maxillofac Pathol J* 2023;14(2).
- Warin K, Limprasert W, Suebnukarn S, Jinaporntham S, Santana P. Performance of deep convolutional neural network for classification and detection of oral potentially malignant disorders in photographic images. *Int J Oral Maxillofac Surg* 2022;51(5):699–704.
- Ariji Y, Fukuda M, Kise Y, et al. Contrast-enhanced computed tomography image assessment of cervical lymph node metastasis in patients with oral cancer by using a deep learning system of artificial intelligence. *Oral Surg Oral Med Oral Pathol Oral Radiol* 2019;127(5):458–463.
- Chang X, Yu M, Liu R, et al. Deep learning methods for oral cancer detection using raman spectroscopy. *Vib Spectrosc* 2023;126, 103522.
- Hossain MS, Shahriar GM, Syeed MM, et al. Tissue artifact segmentation and severity assessment for automatic analysis using WSI. *IEEE Access* 2023;11:21977–21991.
- Panigrahi S, Nanda BS, Bhuyan R, Kumar K, Ghosh S, Swarnkar T. Classifying histopathological images of oral squamous cell carcinoma using deep transfer learning. *Heliyon* 2023;9(3).
- dos Santos DF, de Faria PR, Travençolo BA, do Nascimento MZ. Automated detection of tumor regions from oral histological whole slide images using fully convolutional neural networks. *Biomed Signal Process Control* 2021;69, 102921.
- Bankhead P, Loughrey MB, Fernández JA, et al. Qupath: open source software for digital pathology image analysis. *Sci Rep* 2017;7(1):1–7.
- Jeyaraj PR, Nadar ERS, Panigrahi BK. Resnet convolution neural network based hyperspectral imagery classification for accurate cancerous region detection. 2019 IEEE Conference on Information and Communication Technology. IEEE; 2019. p. 1–6.
- Oya K, Kokomoto K, Nozaki K, Toyosawa S. Oral squamous cell carcinoma diagnosis in digitized histological images using convolutional neural network. *J Dental Sci* 2023;18(1):322–329.
- Dos Santos DF, de Faria PR, Travençolo BA, do Nascimento MZ. Influence of data augmentation strategies on the segmentation of oral histological images using fully convolutional neural networks. *J Digit Imaging* 2023;36(4):1608–1623.
- Fraz MM, Khurram SA, Graham S, et al. Fabnet: feature attention-based network for simultaneous segmentation of microvessels and nerves in routine histology images of oral cancer. *Neural Comput Appl* 2020;32:9915–9928.
- Al-Sukhni E, Attwood K, Gabriel EM, LeVe CM, Kanehira K, Nurkin SJ. Lymphovascular and perineural invasion are associated with poor prognostic features and outcomes in colorectal cancer: a retrospective cohort study. *Int J Surg* 2017;37:42–49.
- Nagpal K, Foote D, Liu Y, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit Med* 2019;2(1):48.
- Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans Pattern Anal Mach Intell* 2012;34(11):2274–2282.
- Tellez D, Litjens G, Bárdi P, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med Image Anal* 2019;58, 101544.
- Zhou Z, Rahman Siddique MM, Tajbakhsh N, Liang J. Unet++: a nested u-net architecture for medical image segmentation. *International Workshop on Deep Learning in Medical Image Analysis*. Springer; 2018. p. 3–11.
- Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)*; 2018. p. 801–818.
- Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. Segformer: simple and efficient design for semantic segmentation with transformers. *Adv Neural Inf Process Syst* 2021;34:12077–12090.
- He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2016. p. 770–778.
- Chollet F. Xception: deep learning with depthwise separable convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2017. p. 1251–1258.
- Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning. PMLR*; 2019. p. 6105–6114.