

Uncertainty-aware automatic segmentation with interactive refinement of head and neck squamous cell carcinoma and pathological lymph nodes

Ziping Chu^{1,2,3} 

ZIPING.CHU@UNSW.EDU.AU


¹ *School of Computer Science and Engineering, University of New South Wales, Sydney, Australia*

² *Ingham Institute for Applied Medical Research, Medical Physics, Liverpool, Australia*

³ *South Western Sydney Clinical Campus, University of New South Wales, Liverpool, Australia*

Sonit Singh¹ 

SONIT.SINGH@UNSW.EDU.AU

Lois Holloway^{2,3,4,5} 

LOIS.HOLLOWAY@UNSW.EDU.AU

⁴ *Liverpool and Macarthur Cancer Therapy Centre, Liverpool, NSW, Australia*

⁵ *Institute of Medical Physics, University of Sydney, Sydney, New South Wales, Australia*

Arcot Sowmya¹ 

A.SOWMYA@UNSW.EDU.AU

Editors: Under Review for MIDL 2026

Abstract

Precise voxel-level annotation of head and neck tumours is crucial for radiotherapy planning but remains challenging due to the small size and irregular shape of tumours and low tumour-tissue contrast in the head and neck region. This study proposes HN-UISeg, an uncertainty-aware automatic segmentation and interactive refinement framework for head and neck squamous cell carcinoma and pathological lymph nodes. The framework was developed using the public HECKTOR and H&N1 datasets and evaluated on a local radiotherapy cohort with paired pre-treatment whole-body PET/CT and planning CT scans. HN-UISeg supports both PET/CT and CT-only input volumes, enabling tumour segmentation even when PET is unavailable. A clinically-oriented pre-processing strategy was adopted, which achieved comparable segmentation performance while maintaining explainability and plug-and-play capability. Uncertainty estimates derived from a combination of Monte Carlo dropout and test-time augmentation were used to generate voxel-wise uncertainty and probability maps, which guide an interactive refinement stage driven by simulated user clicks. Across primary tumours and nodal metastases, HN-UISeg delivers competitive automatic performance and further gains under interactive refinement, supporting more reliable and efficient contouring in head and neck radiotherapy planning. Code will be released on completion of this project.

Keywords: Head and Neck tumour segmentation, Human-in-the-loop medical image segmentation, head and neck squamous cell carcinoma, uncertainty-aware segmentation

1. Introduction

Accurate segmentation of head and neck squamous cell carcinoma and pathological lymph nodes is critical for radiotherapy planning but remains challenging. Difficulties arise from small and irregular primary lesions, numerous pathological lymph nodes, potentially occult primary sites, and the low contrast between tumour voxels and surrounding tissues on CT and PET imaging. These factors complicate both manual annotation and fully automatic contouring, particularly in cases of unknown primary site.

Despite these challenges, recent advances in deep learning (DL)-based medical image segmentation provide robust automated baseline solutions for tumour delineation. The encoder–decoder architecture, introduced by U-Net (Ronneberger et al., 2015), has become the backbone of modern medical image segmentation methods. U-Net and its 3D variants (Çiçek et al., 2016; Milletari et al., 2016) employ a symmetric encoder–decoder architecture with skip connections that combine high-resolution local information with deep contextual features, and have become standard baselines for organ and tumour segmentation. Transformer-based (Dosovitskiy, 2020; Vaswani et al., 2017; Liu et al., 2021) variants such as TransBTS (Wang et al., 2021) and SwinUNETR (Hatamizadeh et al., 2021) further incorporate the self-attention mechanism to capture long-range dependencies, while nnU-Net (Isensee et al., 2021, 2023, 2024) provides a self-adapting framework that is able to adjust the training strategy to any given dataset. These approaches have established robust automatic segmentation baselines. However, due to the lack of uncertainty quantification and limited support for interactive optimisation within clinical workflows, additional dataset-specific optimisation is typically required.

Recent work has shifted towards promptable foundation models and interactive segmentation frameworks. Foundation models refer to large-scale networks pre-trained on diverse data and adaptable to unseen tasks, while promptable indicates that the outputs can be guided during inference through user inputs such as points and boxes. Early Convolutional Neural Network (CNN)-based interactive methods such as DeepGrow (Sakinis et al., 2019) and DeepEdit (Diaz-Pinto et al., 2022; Cardoso et al., 2022) encode user clicks and scribbles as additional input channels, enabling iterative correction of 3D segmentation masks. The Segment Anything Model (SAM) (Kirillov et al., 2023) and its medical variants such as MedSAM (Ma et al., 2024a) and SAM-Med3D (Wang et al., 2025), enable flexible point and box based user interactions and produce probability maps (Wei et al., 2024) across a wide range of image modalities. Language-guided segmentation models (Zhao et al., 2023) have further explored text-driven segmentation, although current language-conditioned approaches remain sub-optimal for small targets compared to task-specific networks. Among these, nnInteractive (Isensee et al., 2025) and LesionLocator (Rokuss et al., 2025) have demonstrated powerful 3D promptable segmentation and longitudinal lesion tracking, highlighting the potential of prompt-based, human-in-the-loop workflows for clinically focused segmentation frameworks.

Head and neck tumour segmentation remains an active area of research due to its complexity and critical importance in radiotherapy planning. Pathological lymph nodes in this region are often small and numerous, may arise from an occult or unknown primary, are distributed across multiple cervical sub-sites and appear as low-contrast lesions on CT and PET imaging, all of which make this segmentation task extremely challenging.

Interactive and uncertainty-aware approaches have recently been proposed to better align automation with clinical practice. The 2S-ICR framework (Saukkoriipi et al., 2025) proposes a two-stage click-based refinement strategy for primary gross tumour volume segmentation in oropharyngeal cancer, achieving improvements in segmentation performance with minimal user interaction. De Biase *et al.* (De Biase et al., 2024) introduced probability maps generated by deep learning for head and neck tumour segmentation on PET/CT scans, with a graphical interface that allows radiation oncologists to threshold and edit these maps. This approach provides a more explainable representation of model prediction than a single hard segmentation mask. However, prior studies focused solely on primary tumours and have not achieved joint segmentation and refinement of lymph nodes within a unified framework incorporating uncertainty quantification.

This study introduces *HN-UISeg*, an uncertainty-aware automatic segmentation and interactive refinement framework for head and neck squamous cell carcinoma and pathological lymph nodes. The framework was developed using two public datasets and validated on a local radiotherapy cohort with paired pre-treatment whole-body PET/CT and chest-up planning CT scans for each patient. The proposed HN-UISeg supports both PET/CT and CT-only inputs, ensuring automatic tumour segmentation and interactive refinement even when PET is unavailable. HN-UISeg utilises a clinically oriented pre-processing strategy and generates voxel-level probability and uncertainty maps to guide the interactive refinement stage for both primary tumours and nodal metastases. On the local South Western Sydney Local Health District (SWSLHD) cohort, HN-UISeg achieved competitive automatic segmentation performance compared to strong baselines such as nnU-Net and SwinUNETR, and further improved segmentation accuracy under interactive refinement.

2. Methods

2.1. Network Architecture

An overview of the proposed HN-UISeg framework is shown in Figure 1. Although Transformer-based models have achieved strong performance on general computer vision tasks, U-Net variants (Roy et al., 2023; Chu et al., 2023) still dominate 3D medical image segmentation because of limited data and the high GPU memory cost of 3D inputs in this field. Consequently, a 3D Residual Encoder U-Net backbone (Isensee et al., 2024) was used for both automatic segmentation and interactive refinement of head and neck primary tumour (GTVp) and pathological lymph nodes (GTVn). The network takes either CT-only volumes or paired PET/CT volumes as input, and produces three output channels corresponding to background, GTVp and GTVn.

The backbone follows the standard Encoder-Decoder design with residual connections (He et al., 2016). The encoder comprises a series of 3D residual blocks that use convolutional operations with a stride of 2 for downsampling, progressively increasing the number of feature channels from 32 to 256. The decoder upsamples features at each resolution and fuses high-level semantic information with spatial detail via skip connections from the corresponding encoder stages. The model accepts four input channels: CT, PET and two prompt channels aligned with the image grid, corresponding to primary tumour and nodal metastasis prompts respectively. All channels are concatenated along the channel dimension at the input stage, enabling early fusion of CT and PET imaging with prompt information. For

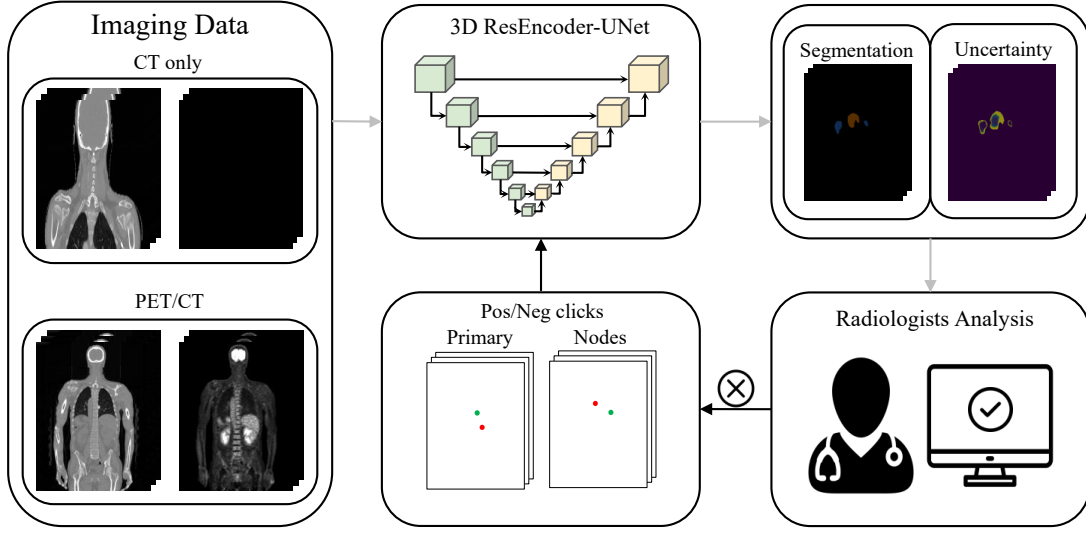


Figure 1: Overview of the proposed HN-UISeg workflow for automatic and interactive segmentation of head and neck squamous cell carcinoma and pathological lymph nodes from PET/CT or CT-only imaging.

cases without PET data, the PET channel is zero-filled. The output layer maps decoder features into voxel-wise logits using a $1 \times 1 \times 1$ convolution, followed by a Softmax to obtain the final class probability maps. Hard segmentation maps are obtained by voxel-wise Argmax, while uncertainty estimates are derived using Monte Carlo dropout and test-time augmentation.

2.2. Training Pipeline

The proposed framework was trained in three stages to alleviate the complications between the automatic and interactive branches within a single network. Circles with a radius of three voxels were used to simulate clicks, as individual voxel clicks did not provide sufficient spatial context. Unless otherwise stated, training used the AdamW (Loshchilov and Hutter, 2017) optimiser with a cosine learning rate schedule including 10 warm-up epochs, and executed for 300 epochs without early stopping. Input volumes and corresponding ground-truth annotations were used to optimise the segmentation backbone with a combined cross-entropy and Dice loss.

Stage 1: Pre-training on public datasets. The first stage utilised the HECKTOR 2022 and H&N1 datasets for pre-training. To familiarise the network with prompt information, the prompt channels had a 50% probability of being activated to simulate user clicks on primary tumours and nodal metastases. When the prompt channels were activated, one simulated positive click was generated for each connected component of the GTVp and GTVn masks by placing a circular mask with a radius of three voxels within the lesion in the corresponding prompt channel. For each tumour type, an equal number of simulated

negative clicks were generated by placing circles at random locations outside its ground-truth region in the corresponding prompt channel. This stage produced an initial set of weights that was used to initialise the network for local fine-tuning.

Stage 2: Automatic fine-tuning on private dataset. In the second stage, the automatic segmentation branch was fine-tuned on 80% of the private dataset using standard supervised learning without prompts, ensuring that images from different stages for the same patient only appeared in either the training or the validation set, but not both. Only the imaging volumes and ground-truth labels were provided as input to and supervision of the automatic segmentation branch, while both prompt channels were set to zero. This stage aimed to adapt the model to the imaging characteristics and contouring protocol of the private dataset, ensuring that the model was aligned with the cohort prior to interactive fine-tuning.

Stage 3: Interactive fine-tuning on private dataset. In the final stage, the model was fine-tuned using the same subset as in the second stage under an interactive setting. For each training volume, an initial segmentation mask was first obtained from the automatic segmentation branch. Subsequently, the differences between the automatic prediction and the ground-truth annotation were analysed at the connected component level, with negative clicks placed in false-positive regions and positive clicks in false-negative regions. These clicks were encoded as circular masks in the corresponding prompt channels and concatenated with the imaging inputs to form the input to the interactive fine-tuning branch. This stage aimed to train the model to integrate user prompts to correct errors from automated segmentation, thereby enabling precise annotation of primary tumour and pathological lymph node contours.

2.3. Loss Function

The proposed network was trained using a combination of cross-entropy loss and soft Dice loss to mitigate the strong class imbalance between tumour and background voxels. For a training patch, let $G \in \{0, 1\}^{I \times C}$ denote the one-hot encoded ground-truth labels and $P \in [0, 1]^{I \times C}$ the corresponding class probability predictions, where I denotes the number of voxels and C the number of classes. $G_{i,c}$ and $P_{i,c}$ are the ground-truth label and predicted probability for voxel i and class c respectively.

The multi-class cross-entropy loss could be defined as

$$\mathcal{L}_{\text{CE}}(G, P) = -\frac{1}{I} \sum_{i=1}^I \sum_{c=1}^C G_{i,c} \log P_{i,c} \quad (1)$$

The soft Dice loss is defined as

$$\mathcal{L}_{\text{Dice}}(G, P) = 1 - \frac{2}{C} \sum_{c=1}^C \frac{\sum_{i=1}^I G_{i,c} P_{i,c}}{\sum_{i=1}^I G_{i,c}^2 + \sum_{i=1}^I P_{i,c}^2} \quad (2)$$

The final training loss combines these two terms as

$$\mathcal{L}(G, P) = \mathcal{L}_{\text{CE}}(G, P) + \mathcal{L}_{\text{Dice}}(G, P) \quad (3)$$

The combined loss penalises voxel-wise mis-classification through the cross-entropy term, while the Dice term directly optimises volumetric overlap between the predicted masks and the ground truth, particularly for the tumour class.

2.4. Evaluation Metrics

Following the recommendations of Metrics Reloaded (Maier-Hein et al., 2024), tumour segmentation performance was evaluated using three complementary metrics: Dice similarity coefficient (DSC), Hausdorff distance (HD) and normalised surface distance (NSD). Full discussion and mathematical definitions can be found in Appendix A.

2.5. Uncertainty Quantification

From a probabilistic perspective, segmentation uncertainty is typically attributed to two main sources: *aleatoric* uncertainty, arising from inherent imaging noise or ambiguity in the imaging data (such as low contrast or artefacts), and *epistemic* uncertainty, arising from limitations in the available data and model capacity. This study employs a combination of Monte Carlo dropout (Gal and Ghahramani, 2016) and test-time augmentation (TTA) (Wang et al., 2019) to obtain voxel-level probability maps and uncertainty maps.

Test-time augmentation was used as a mechanism to assess prediction stability under reasonable input transformations. For each test volume, a set of spatial and intensity augmentations, such as added Gaussian noise, flips or small rotations, was applied. The model was evaluated on each augmented volume and the resulting probability maps were mapped back to the original image space. The variability among these augmented predictions reflected the model’s sensitivity to input perturbations and was summarised into voxel-level uncertainty maps. Monte Carlo Dropout (MC Dropout) was additionally used as an approximate Bayesian treatment of the network parameters. During inference, dropout layers remained active, with the network evaluating the same input volume multiple times using a different random dropout mask. The voxel-wise means of these random probability maps serve as the final prediction, whilst their dispersion metrics (such as variance or entropy) are aggregated into an uncertainty map. Within this framework, MC Dropout primarily captured epistemic uncertainty related to the model parameters, while test-time augmentation detected aleatoric uncertainty by perturbing the input images.

In the uncertainty estimation configuration, dropout layers remained active during inference with a rate of 0.05, and one test-time augmentation (addition of Gaussian noise, flipping along one spatial axis or a small rotation of up to 5°) was randomly selected and applied at each forward pass. The network performed N predictions on augmented input volumes to obtain the probability and uncertainty maps. For a given voxel x and category c , let $p_{x,c}^n$ denote the predicted Softmax probability at the n_{th} forward pass, with $n = 1, \dots, N$. The prediction distribution can be approximated as:

$$P_{x,c} = \frac{1}{N} \sum_{n=1}^N p_{x,c}^n$$

and the corresponding voxel-wise class label obtained by choosing the class with maximum probability. Voxel-wise uncertainty H_x was derived from this approximate predictive

distribution using the predictive entropy:

$$H_x = - \sum_c P_{x,c} \log P_{x,c}$$

where higher entropy values corresponded to more uncertain predictions. As this configuration relies on repeated evaluations of a single model trained on a specific dataset, the uncertainty estimates provide only approximate and potentially an underestimated characterisation of predictive uncertainty compared to deep ensemble approaches.

3. Datasets and Experimental Setup

The proposed framework was developed using three datasets: the public HECKTOR2022 PET/CT dataset (Andrearczyk et al., 2022), the public Head and Neck Radiomics-HN1 (H&N1) cohort (Wee and Dekker, 2019) from The Cancer Imaging Archive (TCIA), and a private dataset collected from the Liverpool-Macarthur Cancer Therapy Centre, South Western Sydney Local Health District (SWSLHD). The HECKTOR and H&N1 datasets were used exclusively for network training to enhance anatomical diversity. Retrospective use of the private dataset was approved by the South Western Sydney Local Health District Human Research Ethics Committee as part of the Radiation Oncology Virtual Clinical Quality Registry (ethics ref: 2019/ETH04391). All quantitative comparisons and visualisation results reported in this manuscript were based on the SWSLHD dataset.

3.1. Datasets

The MICCAI HECKTOR2022 (Oreiller et al., 2022; Andrearczyk et al., 2022) dataset is a multi-centre collection of 524 pre-treatment FDG-PET/CT scans of patients with oropharyngeal squamous cell carcinoma. Each training case comprises co-registered PET and CT volume data with a ground-truth mask, where labels 1 and 2 correspond to primary tumour (GTVp) and pathological lymph nodes (GTVn) respectively.

The Head and Neck Radiomics-HN1 (H&N1) dataset on TCIA (Wee and Dekker, 2019) contains clinical data and CT scans from 137 patients with head and neck squamous cell carcinoma at multiple anatomical sites, treated with radiotherapy at Maastrro Clinic in the Netherlands. The inclusion of the H&N1 dataset enriched the representation of different head and neck anatomical sites within the training data. Target volumes for primary tumours and nodal metastasis were extracted from radiotherapy structure sets and mapped to the GTVp and GTVn label space for this study. Although PET images were available for 75 patients, the absence of body weight data precluded the calculation of Standardised Uptake Value normalised by body weight (SUVbw). Consequently, the PET scans for these patients were excluded from this study.

The private dataset was collected from the Liverpool-Macarthur Cancer Therapy Centre, South Western Sydney Local Health District (SWSLHD). A pre-treatment diagnostic whole-body FDG-PET/CT scan (SWSLHD_PRI) and a chest-up planning CT scan (SWSLHD_PLAN) for radiotherapy planning were available for each patient. Gross tumour volumes for the primary tumour (GTVp) and pathological lymph nodes (GTVn) were manually delineated by a radiologist according to departmental practice. All images

were de-identified prior to analysis and data collection was conducted under local ethics approval. In this study, the private cohort served as the evaluation data.

3.2. Pre-processing

The original voxel spacing in the CT images was maintained to avoid additional interpolation-related uncertainty. For each patient, the corresponding PET volume was resampled to the same voxel grid and size as the paired CT volume. This ensured voxel-wise alignment between modalities, which is consistent with the clinical practice of interpreting PET information within the anatomical context provided by CT. CT intensity values were clipped to the range of $[-1024, 3071]$ Hounsfield Units to remove extreme outliers while preserving the full range of clinically relevant tissue contrast. PET intensities were expressed as body-weight-normalised standardised uptake values (SUVbw) and clipped to $[0, 50]$ SUVbw to remove extreme out-of-field values while retaining explainability in standard quantitative PET units. These settings allowed the network outputs to be applied directly to the clinical images without further resampling, thereby achieving end-to-end implementation within the radiotherapy workflow and enhancing the explainability of the model outputs.

Following previously reported patching methodology (Chu et al., 2024), 3D training patches were extracted by combining random cropping with tumour-contained cropping around regions containing GTVp and/or GTVn to balance background contextual information with sufficient exposure to tumour voxels. Standard 3D data augmentation techniques were applied during training, including random intensity scaling, intensity shifting, addition of Gaussian noise, random rotations and random flips. These data augmentation techniques enhanced the model’s robustness to anatomical variations and differences in imaging protocols across these datasets. All training data were processed using consistent pre-processing and data augmentation protocols which simplified the workflow and facilitated more transparent comparisons between experiments.

3.3. Experimental Setup

All models were trained and evaluated using PyTorch 2.8.0 and CUDA 12.8 on two NVIDIA V100 GPUs with 32 GB of VRAM. A mini-batch size of 1 three-dimensional patch per GPU was used during training.

A 3D sliding-window strategy was employed at inference time. For each test volume, overlapping patches were processed with a 50% overlap in each spatial dimension. The final voxel label was obtained by combining voxel probabilities from overlapping windows through a Gaussian weighting scheme. This approach supported evaluation of volumetric data exceeding the input patch size, alleviating GPU memory constraints and ensuring consistency between the inference process and patch extraction during training. A patch size of $224 \times 224 \times 160$ was used throughout, with the same patch extraction strategy and sliding-window inference configuration applied in all experiments unless otherwise stated.

4. Results

In Table 1, segmentation performance on the SWSLHD staging PET/CT and radiotherapy planning (CT-only) cohorts is reported. All automatic segmentation baselines were

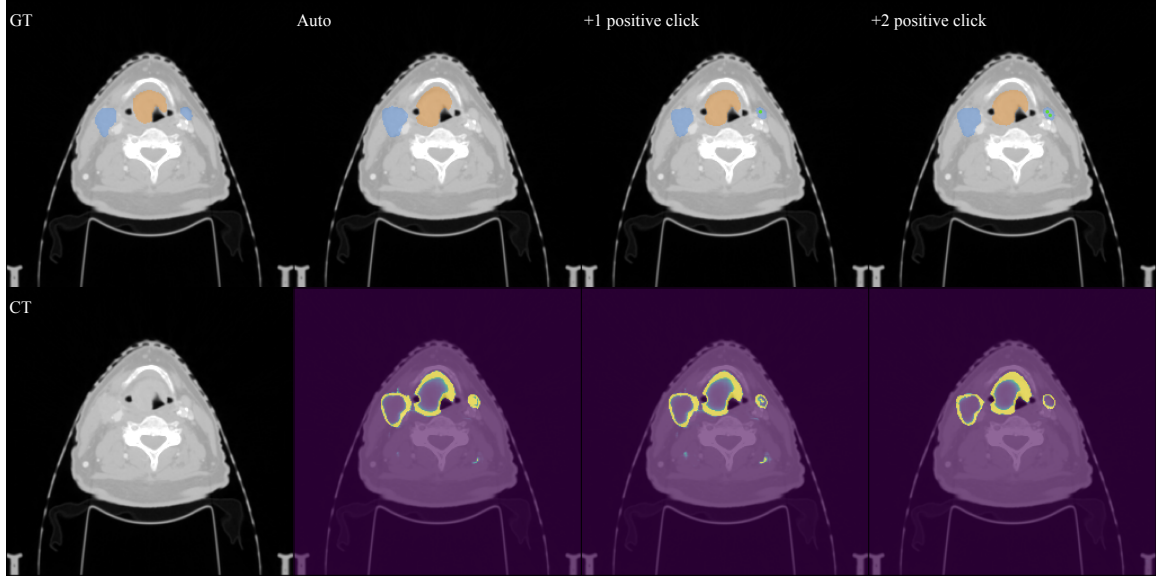


Figure 2: Refinement of automatic segmentation using positive simulated clicks on CT-only volume, with updated model uncertainty map displayed alongside.

trained and evaluated within a common pipeline, using identical training-validation splits, loss functions and training schedule, with nnU-Net run using the official implementation and all other architectures re-implemented under the same protocols. SAM-Med3D and LesionLocator, with and without local supervised fine-tuning, were evaluated on CT-only scans alongside the interactive branch of HN-UISEG. To ensure fair comparison with LesionLocator, only one click was provided for each lesion to refine the results of HN-UISEG. An extensive quantitative comparison with additional network architectures and HN-UISEG with multiple simulated clicks is provided in Appendix B.

Based on the results in Table 1, nnU-Net demonstrated the best segmentation performance across both cohorts among the automatic approaches, while the automatic branch of HN-UISEG achieved comparable results, particularly for nodal metastasis segmentation. Among the interactive approaches, SAM-Med3D underperformed, while LesionLocator and HN-UISEG demonstrated significant improvements over purely automated methods. Under identical simulated user interaction conditions, HN-UISEG demonstrated the best performance in the PET/CT primary-scan cohort, while LesionLocator with supervised fine-tuning performed better in the CT-only planning cohort, particularly for nodal metastasis segmentation.

In Figure 2, an example of refining the automatic segmentation from the planning (CT-only) cohort is shown. With two positive clicks placed in an uncertain region of nodal metastasis, the missing nodal volume was incorporated into the segmentation mask and the corresponding uncertainty map was updated accordingly. Another example of refining automatic segmentation from the primary cohort is provided in Appendix C.

Table 1: Quantitative comparison of segmentation performance on two SWSLHD cohorts. All automatic segmentation models were pre-trained on HECKTOR and H&N1 datasets and results are reported using a five-fold cross-validation on private cohorts. Dice and NSD are reported as percentage scores. HD metric values are in the Appendix B, due to space limitations here. Best results in each column are bolded.

Methods	Primary Scan				Plan Scan (CT-only)			
	Primary		Nodes		Primary		Nodes	
	Dice	NSD	Dice	NSD	Dice	NSD	Dice	NSD
Automatic approaches								
nnU-Net	63.17	61.64	62.23	59.15	58.36	58.39	56.93	55.10
TCTNet	55.52	54.25	50.73	59.25	53.23	54.93	52.26	50.62
SwinUNETR	56.59	57.55	52.73	54.55	53.87	53.73	52.73	58.36
HN-UISeg(Auto)	62.37	58.36	62.50	58.97	57.73	57.76	55.17	58.90
Interactive approaches								
SAM-Med3D	33.86	22.37	29.28	53.37	35.19	19.69	32.86	37.95
LesionLocator	49.51	47.66	64.72	69.76	51.33	53.92	67.64	77.01
LesionLocator + sft	65.45	61.37	68.17	70.85	63.82	60.14	71.69	83.96
HN-UISeg	67.06	64.72	69.83	75.97	60.74	65.92	70.48	72.84

5. Discussion and Conclusion

HN-UISeg demonstrates the potential of uncertainty-guided interactive refinement of head and neck tumour segmentation. However, several limitations remain. In the local SWSLHD cohorts, the automatic segmentation accuracy achieved was below that required for routine clinical application. As pre-training relied mainly on the HECKTOR dataset, which focuses on oropharyngeal squamous cell carcinoma, the model may exhibit bias and perform sub-optimally on other head and neck cancer types and sub-sites. The Monte Carlo dropout and test-time augmentation configuration produced overconfident probability and uncertainty maps in some cases. Because the interactive branch was trained on a relatively small dataset compared to foundation models, the model did not fully learn the semantics of user clicks. As a result, click operations were not consistently interpreted as strictly local corrections and could sometimes induce unintended changes in distant regions. Finally, the evaluation was limited to retrospective and single-centre data with simulated interactions rather than multi-centre cohorts with clinician input.

Despite these limitations, HN-UISeg demonstrates that combining uncertainty-aware automatic segmentation with an interactive refinement branch provides a unified framework for the segmentation of head and neck squamous cell carcinoma and pathological lymph nodes on both PET/CT and CT-only imaging. Future work will include multi-centre validation, more detailed anatomical subregion analysis, improved calibration of probability and uncertainty maps, and refined interaction mechanisms that restrict edits to user-selected regions.

Acknowledgments

Z. Chu gratefully acknowledges scholarship and research support from UNSW Sydney and the Radiation Oncology Department at South Western Sydney Local Health District. This research includes computations using the computational cluster Katana, which is supported by Research Technology Services at UNSW Sydney and the National Computational Infrastructure (NCI), which is supported by the Australian Government.

References

- Vincent Andrearczyk, Valentin Oreiller, Moamen Abobakr, Azadeh Akhavanallaf, Panagiotis Balermipas, Sarah Boughdad, Leo Capriotti, Joel Castelli, Catherine Cheze Le Rest, Pierre Decazes, et al. Overview of the HECKTOR challenge at MICCAI 2022: automatic head and neck tumor segmentation and outcome prediction in PET/CT. In *3D Head and Neck Tumor Segmentation in PET/CT Challenge*, pages 1–30. Springer, 2022.
- M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. MONAI: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.
- Ziping Chu, Sonit Singh, and Arcot Sowmya. TSDNet: A tumour segmentation network with 3d direction-wise convolution. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.
- Ziping Chu, Sonit Singh, and Arcot Sowmya. Robust automated tumour segmentation network using 3D direction-wise convolution and transformer. *Journal of Imaging Informatics in Medicine*, 37(5):2444–2453, 2024.
- Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.
- Alessia De Biase, Liv Ziegfeld, Nanna Maria Sijtsema, Roel Steenbakkens, Robin Wijsman, Lisanne V van Dijk, Johannes A Langendijk, Fokke Cnossen, and Peter van Ooijen. Probability maps for deep learning-based head and neck tumor segmentation: graphical user interface design and test. *Computers in biology and medicine*, 177:108675, 2024.
- Andres Diaz-Pinto, Pritesh Mehta, Sachidanand Alle, Muhammad Asad, Richard Brown, Vishwesh Nath, Alvin Ihsani, Michela Antonelli, Daniel Palkovics, Csaba Pinter, et al. DeepEdit: Deep editable learning for interactive segmentation of 3D medical images. In *MICCAI Workshop on Data Augmentation, Labelling, and Imperfections*, pages 11–21. Springer, 2022.
- Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

- Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.
- Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. SwinUNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.
- Fabian Isensee, Constantin Ulrich, Tassilo Wald, and Klaus H Maier-Hein. Extending nnU-Net is all you need. In *BVM workshop*, pages 12–17. Springer, 2023.
- Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger. nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 488–498. Springer, 2024.
- Fabian Isensee, Maximilian Rokuss, Lars Krämer, Stefan Dinkelacker, Ashis Ravindran, Florian Stritzke, Benjamin Hamm, Tassilo Wald, Moritz Langenberg, Constantin Ulrich, et al. nnInteractive: Redefining 3D promptable segmentation. *arXiv preprint arXiv:2503.08373*, 2025.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024a.
- Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024b.

- Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, et al. Metrics reloaded: recommendations for image analysis validation. *Nature methods*, 21(2):195–212, 2024.
- Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.
- Valentin Oreiller, Vincent Andrearczyk, Mario Jreige, Sarah Boughdad, Hesham Elhawalani, Joel Castelli, Martin Vallieres, Simeng Zhu, Juanying Xie, Ying Peng, et al. Head and neck tumor segmentation in PET/CT: the HECKTOR challenge. *Medical image analysis*, 77:102336, 2022.
- Maximilian Rokuss, Yannick Kirchhoff, Seval Akbal, Balint Kovacs, Saikat Roy, Constantin Ulrich, Tassilo Wald, Lukas T Rotkopf, Heinz-Peter Schlemmer, and Klaus Maier-Hein. LesionLocator: Zero-shot universal tumor segmentation and tracking in 3D whole-body imaging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30872–30885, 2025.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F Jaeger, and Klaus H Maier-Hein. Mednext: transformer-driven scaling of convnets for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 405–415. Springer, 2023.
- Tomas Sakinis, Fausto Milletari, Holger Roth, Panagiotis Korfiatis, Petro Kostandy, Kenneth Philbrick, Zeynettin Akkus, Ziyue Xu, Daguang Xu, and Bradley J Erickson. Interactive segmentation of medical images through fully convolutional neural networks. *arXiv preprint arXiv:1903.08205*, 2019.
- Mikko Saukkoriipi, Jaakko Sahlsten, Joel Jaskari, Lotta Orsmaa, Jari Kangas, Nastaran Rasouli, Roope Raisamo, Jussi Hirvonen, Helena Mehtonen, Jorma Järnstedt, et al. Interactive 3D segmentation for primary gross tumor volume in oropharyngeal cancer. *Scientific reports*, 15(1):28589, 2025.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyang Huang, Yiqing Shen, et al. Sam-med3d: A vision foundation model for general-purpose segmentation on volumetric medical images. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. TransBTS: Multimodal brain tumor segmentation using transformer. In *International conference on medical image computing and computer-assisted intervention*, pages 109–119. Springer, 2021.

Leonard Wee and Andre Dekker. Data from head-neck-radiomics-hn1. (*No Title*), 2019.

Xiaobao Wei, Jiajun Cao, Yizhu Jin, Ming Lu, Guangyu Wang, and Shanghang Zhang. I-medsam: Implicit medical image segmentation with Segment Anything. In *European Conference on Computer Vision*, pages 90–107. Springer, 2024.

Ziheng Zhao, Yao Zhang, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. One model to rule them all: Towards universal segmentation for medical images with text prompts. *arXiv preprint arXiv:2312.17183*, 2023.

Appendix A. Evaluation Metrics

Following the metrics literature (Maier-Hein et al., 2024), tumour segmentation performance was assessed using the Dice similarity coefficient (DSC), Hausdorff distance (HD) and normalised surface distance (NSD). DSC measures the overlap between prediction and ground truth, with higher values indicating better volumetric agreement. Let GT and PD denote the sets of foreground voxels in the ground-truth and predicted segmentation masks respectively. Then DSC may be defined as

$$\text{DSC} = \frac{2|GT \cap PD|}{|GT| + |PD|} \quad (4)$$

where $|\cdot|$ denotes the number of voxels within the set.

Normalised surface distance was used to quantify surface agreement within a fixed tolerance value, with higher values indicating a larger fraction of the contour surfaces lying within the specified tolerance. Let S_{GT} and S_{PD} denote the sets of surface voxels of the ground truth (GT) and predicted segmentation masks (PD) respectively, and let $d(x, S)$ denote the minimum Euclidean distance from a point x to the set S . For a given tolerance τ , the border regions around the two surfaces are defined as

$$\beta_{PD}^{(\tau)} = \{x : d(x, S_{PD}) \leq \tau\}, \quad \beta_{GT}^{(\tau)} = \{x : d(x, S_{GT}) \leq \tau\} \quad (5)$$

The NSD at tolerance τ is then defined as

$$\text{NSD}^{(\tau)}(PD, GT) = \frac{|S_{GT} \cap \beta_{PD}^{(\tau)}| + |S_{PD} \cap \beta_{GT}^{(\tau)}|}{|S_{GT}| + |S_{PD}|} \quad (6)$$

In this study, the tolerance was set to $\tau = 2$ mm, reflecting a typical clinically acceptable contour deviation in radiotherapy.

Hausdorff distance was computed between the predicted surfaces and ground truth masks to measure the worst-case boundary error. Using the same notation from NSD, the Hausdorff distance is given by

$$\text{HD}(PD, GT) = \max \left\{ \max_{x \in S_{PD}} d(x, S_{GT}), \max_{y \in S_{GT}} d(y, S_{PD}) \right\} \quad (7)$$

and is reported in millimetres. Lower HD values correspond to smaller worst-case boundary deviations between predicted and reference contours.

All metrics were computed separately for the primary tumour (GTVp) and pathological lymph nodes (GTVn).

Appendix B. Complete Quantitative Results

In Table 2, detailed segmentation performance on the SWSLHD staging PET/CT and radiotherapy planning (CT-only) cohorts is reported. Among automated methods, nnU-Net and HN-UISeg(Auto) demonstrated comparable performance and outperformed the other architectures. By comparison, U-Net variants based on Transformers and Mamba (U-mamba (Ma et al., 2024b), TCTNet and SwinUNETR) underperformed compared to the convolutional neural networks (CNN)-based models on this segmentation task. This could result from the relatively limited dataset size under the current training conditions, as well as the strong inductive bias inherent in convolutional architectures, which is better suited to handling the local high-frequency anatomical detail characteristic of head and neck tumours.

Among interactive methods, SAM-Med3D performed poorly, revealing notable domain shift when handling small, low-contrast head and neck lesions. LesionLocator exhibited robust overlap metrics under the zero-shot conditions and achieved the lowest Hausdorff distance value for both primary lesions and nodal metastasis on both staging and planning scans after supervised fine-tuning. HN-UISeg further improved segmentation performance when user interaction was incorporated. Using up to five clicks for the primary tumour and five for nodal metastases (10-clicks) further refined the segmentation masks and improved the delineation of both primary and nodal disease.

The advantage of fine-tuned LesionLocator over HN-UISeg in terms of Hausdorff distance revealed key design differences between the two branches. LesionLocator performs segmentation directly from user inputs and therefore rarely produced abnormal segments distant from lesions. In contrast, HN-UISeg refined automatic segmentation results and could inherit erroneous distant predictions, resulting in larger Hausdorff distances even when DSC and NSD are similar.

Appendix C. Qualitative Results for Primary PET/CT Scan

In Figure 3, interactive refinement of the automatic segmentation for a staging (PET/CT) scan is illustrated. The segmentation result was refined by placing a positive click in the uncertain region of the nodal metastasis and a negative click in a non-primary tumour region, with the uncertainty map adjusted accordingly.

Table 2: Segmentation performance on the SWSLHD cohorts. All models were pre-trained on HECKTOR and H&N1 datasets and reported using five-fold cross-validation. Dice and NSD are reported as percentage scores and HD is reported in millimetres. Best performance in each column is bolded.

Methods	Primary Scan						Plan Scan (CT-only)					
	Primary			Nodes			Primary			Nodes		
	Dice	NSD	HD	Dice	NSD	HD	Dice	NSD	HD	Dice	NSD	HD
Automatic approaches												
nnU-Net	63.17	61.64	54.82	62.23	59.15	71.05	58.36	58.39	58.89	56.93	55.10	67.80
TSDNet	59.70	58.61	64.47	54.99	53.51	76.63	53.76	54.36	57.75	55.52	48.17	76.55
U-mamba	55.98	56.36	65.45	51.34	52.37	105.79	52.57	50.85	63.22	50.61	63.36	93.68
TCTNet	55.52	54.25	59.21	50.73	59.25	111.07	53.23	54.93	62.15	52.26	50.62	107.5
SwinUNETR	56.59	57.55	62.72	52.73	54.55	91.23	53.87	53.73	65.32	52.73	58.36	101.62
HN-UISeg(Auto)	62.37	58.36	59.55	62.50	58.97	75.58	57.73	57.76	63.40	55.17	58.90	68.19
Interactive approaches												
SAM-Med3D	33.86	22.37	26.03	29.28	53.37	23.87	35.19	19.69	58.26	32.86	37.95	22.83
LesionLocator (zero-shot)	49.51	47.66	20.08	64.72	69.76	16.75	51.33	53.92	17.75	67.64	77.01	15.25
LesionLocator + sft	65.45	61.37	10.67	68.17	70.85	9.99	63.82	60.14	11.66	71.69	83.96	9.11
HN-UISeg (each-lesion)	67.06	64.72	58.82	69.83	75.97	66.75	60.74	65.92	47.79	70.48	72.84	51.56
HN-UISeg (10-clicks)	69.96	69.14	54.92	71.76	74.79	61.48	61.91	65.14	42.55	73.21	74.54	44.51

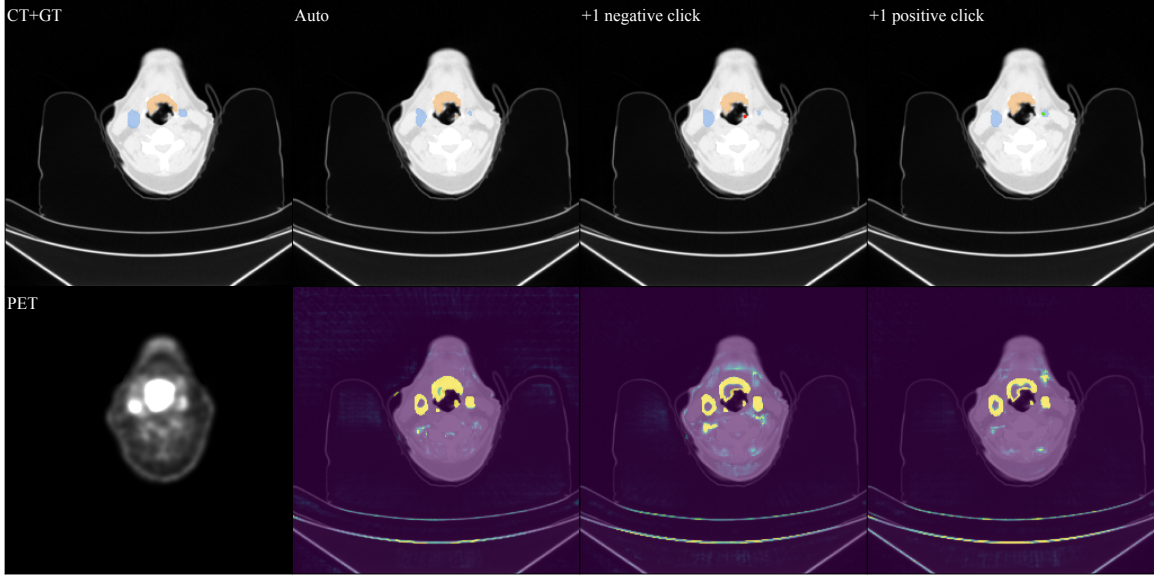


Figure 3: Refinement of automatic segmentation using one positive and one negative simulated clicks on PET/CT volume, with updated model uncertainty map displayed alongside.

Appendix D. Ablation Studies

D.1. Pre-processing Method

In Tables 3 and 4 the effect of CT and PET pre-processing schemes on segmentation performance is summarised. All metrics were obtained using the automatic branch of HN-UISeg, pre-trained on the HECKTOR and H&N1 datasets and evaluated with five-fold cross-validation on both SWSLHD cohorts.

For CT scans, using raw Hounsfield Unit (HU) values directly led to significantly worse segmentation performance than both windowed schemes, due to the presence of extreme outliers (such as a -16989 HU value). Both soft tissue window $[-200, 400]$ and the wide window $[-1024, 3071]$ achieved comparable Dice coefficients without significant difference. As a result, the $[-1024, 3071]$ HU window was adopted as the default CT pre-processing approach due to its clinical plug-and-play capability.

For PET, both radioactivity concentration (in Bq/mL) and un-windowed Standardised Uptake Value normalised by body weight (SUV_{bw}) are not as standardised as CT HUs and could be influenced by uptake time, reconstruction settings and patient physiology. As shown in Table 4, raw Bq/mL and un-windowed SUV_{bw} obtained significantly worse segmentation performance than SUV_{bw} with simple intensity standardisation. SUV_{bw} windowing to $[0, 50]$ and per-scan z-score normalisation performed comparably. Although liver-based normalisation is common in clinical practice, the liver lies outside the field-of-view in many chest-up scans. Therefore, SUV_{bw} windowing to $[0, 50]$ was adopted as a simple, plug-and-play setting.

Table 3: Ablation study on CT pre-processing methods. Results were obtained through five-fold cross-validation. Dice scores are reported as percentages. The Wilcoxon signed-rank test was used to assess statistical significance across all patients between methods, * indicates a p-value < 0.05 compared with the proposed method.

Methods	Primary Scan		Plan Scan (CT-only)	
	Primary	Nodes	Primary	Nodes
CT-HU (raw)	31.45*	34.44*	40.89*	40.29*
Soft Tissue Window [-200, 400]	61.76	63.23	57.07	55.93
Wide HU window [-1024, 3071]	62.37	62.50	57.73	55.15

Table 4: Ablation study on PET pre-processing methods on staging scans. Results were obtained through five-fold cross-validation. Dice scores are reported as percentages. The Wilcoxon signed-rank test was used to assess statistical significance across all patients between methods, * indicates a p-value < 0.05 compared with the proposed method.

Method	Primary	Nodes
PET-Bq (raw)	33.67*	37.92*
PET-SUVbw (raw)	34.42*	42.04*
z-score	61.78	63.18
PET-SUVbw window [0,50]	62.37	62.50

D.2. Number of Clicks

The relationship of the Dice scores for GTVp and GTVn with the number of user clicks in the staging PET/CT and planning CT-only cohorts is illustrated in Figure 4. In both cohorts, performance increased with the number of simulated clicks, particularly for pathological lymph nodes in the planning cohort. However, the clear trends indicate that beyond the average number of connected components, further increases in the number of clicks yield only marginal performance gains. Even with up to five clicks per case, segmentation accuracy remained below typical requirements for routine clinical application. This diminishing margin of benefit of additional clicks reflects the limitations of the current training strategy and highlights the need for improved training schemes and detailed case studies.

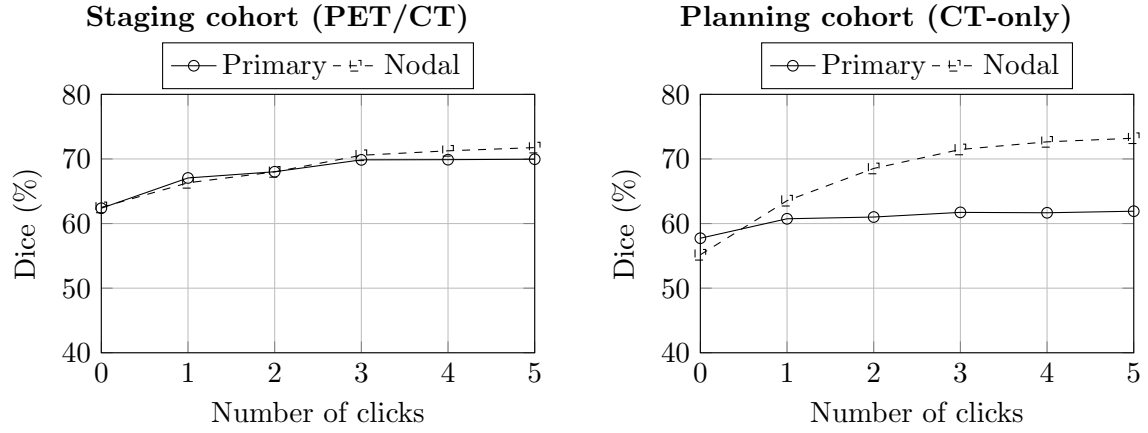


Figure 4: Dice score as a function of the number of user clicks for primary tumours and nodal metastases on the staging PET/CT cohort (left) and the planning CT-only cohort (right).