nnU-Net v2 for Head-and-Neck PET/CT Primary Tumor and Lymph Node Segmentation: A Simple, Strong Baseline

Yansong Bu[1,2⋆], Zihao Wang[1,2⋆], Yuwan Wang[1,2], Jiexin Jiang[1], Yuhan Chen[1], and Mengye Lyu[1,2]

[1] College of Health Science and Environmental Engineering, Shenzhen Technology University, Shenzhen, China
[2] College of Applied Sciences, Shenzhen University, Shenzhen, China
Corresponding author: Mengye Lyu
lvmengye@sztu.edu.cn

Abstract. Accurate segmentation of primary tumors (GTVp) and nodal lesions (GTVn) is essential for radiotherapy planning in head and neck cancer (HNC). The HECKTOR 2025 challenge provides a large-scale multi-center PET/CT dataset for benchmarking automated tumor segmentation. In this study, we developed a segmentation pipeline based on the nnU-Net v2 framework without architectural modifications. The model was trained using 5-fold cross-validation on the complete training cohort and deployed in a Docker container for submission. On the official validation leaderboard, our approach achieved a GTVp Dice of 0.7444, a GTVn Dice of 0.7956, and a GTVn F1-score of 0.5868, securing seventh place overall in Task 1. These results demonstrate that nnU-Net v2 remains a competitive baseline for multi-center PET/CT segmentation tasks, providing robust tumor delineation performance across heterogeneous datasets. We ranked third overall in the HECKTOR 2025 Challenge (Task 1).

Keywords: Medical image segmentation · nnU-Net· tumor segmentation· MRI-guided radiation therapy.

1 Introduction

Head and neck cancer (HNC) remains one of the most challenging malignancies to treat due to its complex anatomical structures and the close proximity of tumors to critical organs at risk [1]. Radiation therapy (RT) plays a central role in the management of HNC, and accurate delineation of gross tumor volumes (GTVs) is essential for effective treatment planning and outcome optimization, especially with the increasing adoption of MRI-guided RT, which offers superior soft-tissue contrast [2]. In recent years, the HEad and neCK TumOR Lesion Segmentation (HECKTOR) challenge series has provided large-scale, multi-center

⋆ These authors contributed equally to this work.

datasets with standardized evaluation protocols, enabling the benchmarking of automated tumor segmentation methods in PET/CT imaging [3]. The multi-institutional nature and heterogeneous data distributions of HECKTOR make it a valuable testbed for developing clinically robust algorithms.

Manual delineation of primary tumors (GTVp) and involved lymph nodes (GTVn) from PET/CT is time-consuming, labor-intensive, and prone to inter-observer variability [4]. Deep learning methods, especially fully convolutional neural networks, have shown promise in automating this task and alleviating the burden on clinicians [5]. Among them, nnU-Net has emerged as a widely adopted framework due to its self-configuring design and strong performance across diverse medical image segmentation challenges [6]. Its ability to automatically adapt preprocessing, architecture, and training parameters to the target dataset makes it a strong and reproducible baseline. Therefore, we present a simple, fully documented nnU-Net v2 baseline for the HECKTOR 2025 challenge, aiming to maximize reproducibility and provide a strong reference for PET/CT-based HNC segmentation.

We evaluate this baseline on the HECKTOR 2025 Task 1 segmentation benchmark. Without introducing architectural modifications or large-scale pre-training, we demonstrate that the default nnU-Net v2 pipeline achieves competitive performance on this multi-center dataset. On the official validation leaderboard, our approach reached an overall rank of seventh place, highlighting that nnU-Net v2 remains a strong baseline for PET/CT-based HNC tumor segmentation. In the final test phase, our method ranked third overall in Task 1, further reinforcing this message.

2 Dataset

The dataset used in this study originates from the HEad and NeCK TumOR Lesion Segmentation (HECKTOR) 2025 challenge [3]. It is a large-scale, multi-institutional collection of FDG-PET/CT scans from over 1,100 patients with histologically confirmed head and neck cancer, acquired across 10 international medical centers. The training cohort consists of approximately 700 cases from 8 centers, while the test cohort includes about 450 previously unseen cases from 3 centers. Notably, a subset of the 2022 test set was incorporated into the 2025 training cohort. The estimated HPV status distribution in the test set is approximately 80 HPV-positive and 20 HPV-negative.

Each patient case includes co-registered CT and PET images. PET images were normalized to standardized uptake values (SUVs), and CT scans were provided as low-dose non-contrast-enhanced images. Ground-truth segmentation masks, available only for Task 1, delineate primary tumors (GTVp) and involved lymph nodes (GTVn). All images underwent standardized preprocessing to ensure cross-center comparability and facilitate model training.

3 Method

3.1 Model Architecture

For Task 1, we selected the nnU-Net v2 [6] framework as our segmentation model and trained it on the HECKTOR 2025 dataset. nnU-Net is a widely adopted, self-configuring deep learning framework for medical image segmentation, which automatically adapts preprocessing, network architecture, data augmentation, and training hyperparameters to the target dataset. This flexibility allows it to serve as a robust and reproducible baseline across a variety of imaging tasks.

In our study, we used the 3D full-resolution configuration, which determines patch size, network depth, loss function (a combination of Dice and cross-entropy), and learning rate schedule automatically according to the dataset properties. No manual modifications were made to the network architecture. For multi-modal input, CT and PET images were concatenated as separate channels, following the default nnU-Net v2 implementation.

3.2 Pre-processing

The raw dataset was provided in a patient-wise folder structure, where each folder contained a CT scan, a PET scan, and the corresponding segmentation mask. To comply with the nnU-Net v2 input format, the data were reorganized into a standardized directory layout, including a dataset description file and modality-indexed inputs (with CT assigned to channel 0000 and PET to channel 0001). During this step, we identified a small number of invalid cases where the CT and label volumes were not aligned in size; these cases were excluded from training. Furthermore, discrepancies in voxel dimensions between CT and PET images were corrected by resampling the PET volumes using the SimpleITK library to ensure voxel-wise correspondence. The final dataset consisted of paired CT and PET inputs with corresponding labels, ready for training with nnU-Net v2.

3.3 Training and Inference

Model training was performed with 5-fold cross-validation to maximize the use of the available training cohort and assess robustness across different partitions. Each fold was trained independently for 1000 epochs using mixed-precision training. All other hyperparameters were kept as default in nnU-Net v2. During inference, the framework's standard strategy was employed, namely sliding-window prediction with overlapping patches, followed by Gaussian weighting to fuse patch-level outputs into the final segmentation map. For the official submission, results from the cross-validation models were aggregated and evaluated on the challenge validation set.

3.4 Implementation Details

All experiments were implemented using Python and the PyTorch deep learning library, and training was conducted on a GPU server. The nnU-Net v2 (3D full-resolution) framework was used for both training and inference. During inference, single-case predictions were performed using a sliding-window strategy with Gaussian weighting, and the predicted segmentation masks were resampled to the original CT geometry to preserve label integrity. The final outputs were saved in the required format according to HECKTOR challenge guidelines. Training protocols are summarized in Table 1, and hardware configuration and development environments are presented in Table 2.

Table 1. Training protocols.

Network initialization	normal initialization
Batch size	2
Patch size	$96 \times 160 \times 160$
Total epochs	1000
Optimizer	SGD with Nesterov momentum ($\mu = 0.99$)
Weight decay	3e-5
Loss function	Dice + cross-entropy loss
Initial learning rate (lr)	0.001

Table 2. Development environments and requirements.

System	Ubuntu 22.04
CPU	Intel(R) Xeon(R) CPU E5-2680 v4
RAM	256 GB
GPU (number and type)	8× Tesla P40
CUDA version	12.0
Programming language	Python 3.10.18
Deep learning framework	PyTorch 2.7.1; torchvision 0.22.1

4 Results

The performance of the proposed pipeline based on nnU-Net v2 was evaluated on the HECKTOR 2025 Task 1 segmentation benchmark. To ensure robustness, we conducted a 5-fold cross-validation on the training cohort. This strategy allowed us to assess the variability of the model across different data partitions and reduce the risk of overfitting to a single split. The cross-validation results demonstrated stable segmentation performance for both primary tumors (GTVp) and lymph nodes (GTVn), highlighting the reliability of the nnU-Net v2 framework

in a multi-center PET/CT setting. As illustrated in Fig. 1, the predicted segmentations show good agreement with the ground-truth annotations, further confirming the effectiveness of nnU-Net v2.

For the official submission, the trained nnU-Net v2 model was applied to the challenge validation set. On the official leaderboard, our method achieved a GTVp Dice of 0.7444 (rank 9), a GTVn aggregated Dice of 0.7956 (rank 2), and a GTVn F1-score of 0.5868 (rank 14), resulting in a mean position of 8.3 and an overall rank of 7th among all participating teams. These results indicate that the default nnU-Net v2 configuration, without any architectural modifications or external pre-training, remains a strong and competitive baseline for automatic segmentation of head and neck tumors in PET/CT images.

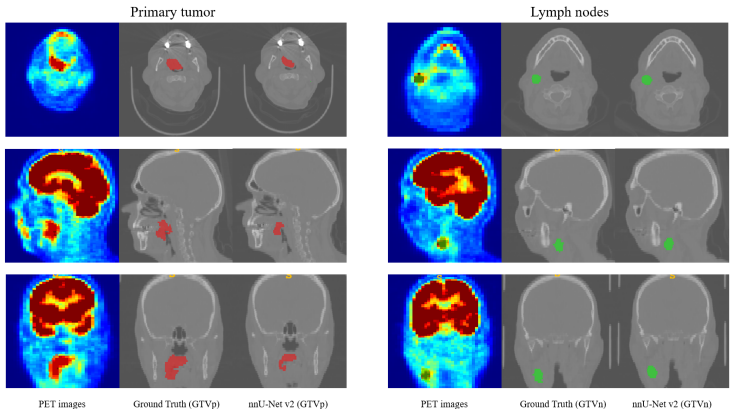

Fig. 1. Qualitative visualization of the segmentation results. For ease of comparison, only representative cropped regions of the original images are shown.

Table 3. Cross-validation Dice scores (training cohort, 5-fold).

Fold	Mean Validation Dice
0	0.6913
1	0.6958
2	0.6933
3	0.6800
4	0.6650

Table 4 summarizes the quantitative performance of our method on the official validation leaderboard.

Table 5 reports the official test results, where our team ranked third overall in Task 1.

Table 4. Official validation leaderboard results (team `sztu_bme2025`, HECKTOR 2025 Task 1).

Metric	Score	Rank	Description
GTVp Dice	0.7444	9	Primary tumor segmentation
GTVn Dice (agg)	0.7956	2	Nodal tumor segmentation
GTVn F1-score	0.5868	14	Detection sensitivity
Mean Position	8.3	–	Averaged ranking across metrics
Overall Rank	–	7	Leaderboard standing

Table 5. Official test leaderboard results (team `sztu_bme2025`, HECKTOR 2025 Task 1).

Metric	Score	Rank	Description
GTVp Dice	0.7308	–	Primary tumor segmentation
GTVn Dice (agg)	0.7641	–	Nodal tumor segmentation
GTVn F1-score (agg)	0.6320	–	Detection sensitivity
Mean Position	–	–	Averaged ranking across metrics
Overall Rank	–	**3**	Leaderboard standing

5 Discussion

In this study, we investigated the performance of nnU-Net v2 on the HECKTOR 2025 Task 1 segmentation benchmark. Without introducing architectural modifications or external pre-training, the standard 3D full-resolution configuration of nnU-Net v2 achieved a competitive 7th place on the official validation leaderboard and maintained comparable accuracy on the hidden test set, ultimately placing third overall in the final Task 1 ranking.

These results highlight the robustness of nnU-Net v2 as a strong baseline for head and neck tumor segmentation in multi-center PET/CT data, particularly reflected in the high DSCagg score for nodal disease (GTVn).

Despite these promising results, several limitations should be noted. First, the relatively low F1-score for GTVn suggests that the model tends to miss smaller or less distinct lesions, indicating limited sensitivity in challenging cases. This may be partly due to the absence of advanced post-processing strategies, such as threshold optimization or connected-component filtering, which could improve recall. Second, unlike recent large-scale models that leverage transfer learning, our approach did not incorporate external pre-training or domain adaptation techniques. As a result, the model may be less effective at handling inter-institutional variability across unseen test centers.In addition, our internal experiments indicated that using STU-Net [7, 8] might further improve the F1-score, especially for nodal disease. However, due to time constraints, we were not able to submit this approach for official evaluation.

Future research can address these limitations in several ways. Semi-supervised learning strategies could leverage the unannotated test-center data to improve generalization. Domain adaptation methods may further reduce the performance

gap across centers with heterogeneous imaging protocols. A two-stage pipeline combining lesion detection and fine-grained segmentation could help capture small-volume nodal disease more effectively. [9,10] In addition, uncertainty-aware post-processing may provide more reliable outputs for clinical use. Finally, training with larger patch sizes or more complex architectures could be explored on hardware with higher memory capacity, such as TPUs or high-memory GPUs, to better capture long-range spatial context.

In summary, this study demonstrates that nnU-Net v2, even in its default configuration, provides a strong and reproducible baseline for automatic tumor segmentation in PET/CT images of head and neck cancer. These findings suggest that while more advanced models may further improve accuracy, nnU-Net v2 remains a reliable starting point for clinical and research applications in radiation therapy planning.

Acknowledgments. This work was in part supported by Shenzhen Science and Technology Program (Shenzhen Higher Education Stable Support Program, No. 20220716111838002) and a university–industry collaborative project (No. HT20241125001).

Disclosure of Interests. The authors have no competing interests.

References

1. G. Anderson, M. Ebadi, K. Vo, J. Novak, A. Govindarajan, and A. Amini, "An updated review on head and neck cancer treatment with radiation therapy," *Cancers*, vol. 13, no. 19, p. 4912, 2021.
2. D. Lavigne, S. P. Ng, B. O'Sullivan, P. F. Nguyen-Tan, E. Filion, and L. e. a. Létourneau-Guillon, "Magnetic resonance-guided radiation therapy for head and neck cancers," *Current Oncology*, vol. 29, no. 11, pp. 8302–8315, 2022.
3. N. Saeed, S. Hassan, S. Hardan, A. Aly, D. Taratynova, U. Nawaz, U. Khan, M. Ridzuan, V. Andrearczyk, A. Depeursinge, Y. Xie, T. Eugene, R. Metz, M. Dore, G. Delpon, V. R. K. Papineni, K. Wahid, C. Dede, A. M. S. Ali, C. Sjogreen, M. Naser, C. D. Fuller, V. Oreiller, M. Jreige, J. O. Prior, C. C. L. Rest, O. Tankyevych, P. Decazes, S. Ruan, S. Tanadini-Lang, M. Vallières, H. Elhalawani, R. Abgral, R. Floch, K. Kerleguer, U. Schick, M. Mauguen, D. Bourhis, J.-C. Leclere, A. Sambourg, A. Rahmim, M. Hatt, and M. Yaqub, "A multimodal and multi-centric head and neck cancer dataset for segmentation, diagnosis, and outcome prediction," 2025.
4. S. Gudi, A.-H. Le, P. Maingon, R. de Crevoisier, L. Henriques de Figueiredo, J.-C. Simon, F. Mornex, J.-P. Guichard, J.-L. Lefebvre, and M. Kok, "Inter-observer variation in head and neck gross tumour volume delineation on ct and pet–ct: Comparison between radiation oncologists and radiologists from the same institution," *Radiotherapy and Oncology*, vol. 122, no. 2, pp. 300–304, 2017.
5. M. A. Naser, J. R. Weir-McCall, G. Harding, R. Gnann, R. H. Vallejo, R. Stoian, J. Kalpathy-Cramer, A. Louie, W. T. Ticks, P. Nguyen, J. O. prior, and M. Hatt, "Tumor segmentation in patients with head and neck cancers using deep learning based on multi-modality pet/ct images," *Cancers*, vol. 13, no. 3, p. 615, 2021.

6. F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.
7. Z. Huang, H. Wang, Z. Deng, J. Ye, Y. Su, H. Sun, J. He, Y. Gu, L. Gu, S. Zhang, and Y. Qiao, "Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training," *arXiv preprint arXiv:2304.06716*, 2023.
8. Z. Wang and M. Lyu, "Head and neck tumor segmentation for mri-guided radiation therapy using pre-trained stu-net models," in *Head and Neck Tumor Segmentation for MR-Guided Applications* (K. A. Wahid, C. Dede, M. A. Naser, and C. D. Fuller, eds.), vol. 15273 of *Lecture Notes in Computer Science*, pp. 65–74, Springer, Cham, 2025.
9. S. Huang, L. Mei, J. Li, Z. Chen, Y. Zhang, T. Zhang, X. Nie, K. Deng, and M. Lyu, "Abdominal ct organ segmentation by accelerated nnunet with a coarse to fine strategy," in *Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation* (J. Ma and B. Wang, eds.), vol. 14221 of *Lecture Notes in Computer Science*, pp. 23–34, Springer, Cham, 2022.
10. S. Huang, H. Yang, L. Mei, T. Zhang, S. Liu, and M. Lyu, "From whole-body to abdomen: Streamlined segmentation of organs and tumors via semi-supervised learning and efficient coarse-to-fine inference," in *Fast, Low-resource, and Accurate Organ and Pan-cancer Segmentation in Abdomen CT* (J. Ma and B. Wang, eds.), vol. 14904 of *Lecture Notes in Computer Science*, pp. 283–292, Springer, Cham, 2024.