Less is More: Efficient PET/CT Segmentation and Multimodal Prediction of Recurrence-Free Survival and HPV Status in Head and Neck Cancer

Lishan Cai[1,2], XingLong Liang[2,3], Tianyu Zhang[2,3], Jiaju Huang[4], Tao Tan[2,4], and Yunchao Yin[1]

[1] Department of Radiation Oncology, Amsterdam University Medical Center, De Boelelaan 1117, 1081 HV Amsterdam, the Netherlands
l.cai@amsterdamumc.nl
[2] Department of Radiology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX, Amsterdam, The Netherlands
[3] Department of Radiology and Nuclear Medicine, Radboud University Medical Centre, Geert Grooteplein 10, 6525 GA, Nijmegen, The Netherlands
[4] Faculty of Applied Sciences, Macao Polytechnic University, 999078, Macao Special Administrative Region of China

Abstract. Accurate delineation of primary head and neck tumors and metastatic lymph nodes on PET/CT is critical for radiotherapy planning and prognostic assessment. Building on this clinical need, the HECKTOR 2025 challenge uses a large multi-centric dataset to provide a comprehensive benchmark for multimodal methods that integrate imaging and clinical information across three key tasks: segmentation of the primary tumor and involved lymph nodes, recurrence-free survival prediction, and HPV status classification. In this study, we (Team MEDAI) present our solutions for all three challenge tasks. For automated tumor segmentation, we employed an ensemble of ten lightweight STU-Net (small) models, achieving efficient and precise delineation of both primary tumors and metastatic lymph nodes. For recurrence-free survival prediction and HPV status classification, we developed a multimodal framework that integrates volumetric PET/CT imaging, lesion masks derived from the segmentation models, and structured clinical variables. Code is available at https://github.com/Liiiii2101/HECKTOR2025-MEDAI. Team: MEDAI.

Keywords: Medical Image Analysis · HECKTOR challenge · Deep Learning · Segmentation · Recurrence-Free Survival · HPV Status

1 Introduction

Head and neck cancer (HNC), one of the most common malignancy worldwide, arises from the anatomical sites of the upper aerodigestive tract [5,11]. Positron

L. Cai and X. Liang contributed equally.

emission tomography /computed tomography (PET/CT), which provide complementary and synergistic information for HNC lesion segmentation and tumor characterization by highlighting metabolic and morphological tissue properties, is recommended for staging HNC, with additional roles in prognostication, treatment planning, and identifying unknown primary tumors [6]. Currently, radiotherapy is integral to the therapeutic management of HNC [2]. Precise tumor contouring is essential to ensure adequate dose delivery to the targeted tumor while sparing neighboring normal tissues. Manual segmentation of head and neck tumors and involved lymph nodes is time-consuming, labor-intensive, and prone to inter-observer variability. Accurate and automated segmentation of primary gross tumors and involved lymph nodes can support and streamline clinical workflows. Furthermore, when combined with clinical features, automated segmentation masks can be used to predict recurrence-free survival (RFS), offering valuable prognostic insight for HNC. Accurate RFS prediction helps identify patients at elevated risk of early recurrence, those who may benefit from treatment intensification, closer surveillance, or early referral for supportive interventions, while also supporting more individualized follow-up strategies for lower-risk patients. Additionally, accurate assessment of biomarkers such as human papillomavirus (HPV) status from PET/CT and clinical data can improve risk stratification and diagnostic accuracy.

In the past decade, deep learning (DL) has shown promising results in medical image analysis [1]. Notably, DL models have demonstrated strong performance in HNC tumor segmentation using PET/CT in previous HEad and neCK TumOR Lesion Segmentation, Diagnosis and Prognosis (HECKTOR) challenge 2022 [3]. HECKTOR 2025 [13] provides a platform for developing 3D algorithms to segment primary HNC tumors (GTVp) and nodal metastases tumors (GTVn) on PET/CT, as well as for recurrence survival prediction and HPV status classification using multimodal and multi-centric dataset.

In this article, we present our approaches for all three tasks: automatic segmentation of primary tumors and involved lymph nodes on FDG-PET/CT, prediction of recurrence-free survival using PET/CT and clinical data, and classification of HPV status based on PET/CT and clinical information.

2 Dataset

2.1 Overview

For HECKTOR 2025, the dataset comprises a large-scale, multi-centric collection of multimodal PET/CT scans and detailed clinical data from 1,123 patients with histologically confirmed HNC. The publicly available training set comprises approximately 700 cases collected from eight different centers. The hidden leaderboard validation set contains approximately 50 unseen cases, whereas the hidden final leaderboard test set comprises around 450 cases from three centers.

2.2 Preprocessing

Resampling and Cropping The data preprocessing is inspired by the first-place solution of HECKTOR 2022 by Myronenko et al. [12]. All images were resampled to isotropic voxel spacing of 1 × 1 × 1 $\text{mm}^{\mathbf{3}}$ using B-spline interpolation for CT/PET and nearest-neighbor interpolation for lesion masks. Overlapping bounding boxes between CT and PET volumes were computed to ensure spatial alignment and remove excess background to focus on the relevant anatomical region. A region-of-interest (ROI) was automatically determined using PET intensities. The top portion of the PET was analyzed to identify the largest high-intensity region. The crop center was defined as the centroid of this region. A fixed-size crop (200 × 200 × 310) around this center was extracted for CT, PET, and segmentation masks if available. This resampling and cropping were applied across all three tasks to reduce input image size, significantly accelerating training and minimizing network workload on irrelevant regions.

Normalization for Segmentation nnU-Net [9] CT normalization scheme was applied to all cropped CT images. Intensity values from all foreground classes (excluding background) across the training set were collected to compute the mean, standard deviation, and 0.5/99.5 percentiles. Values were clipped to these percentiles, normalized by subtracting the mean, and scaled by the standard deviation. For PET, Z-Score normalization was applied by subtracting the mean and dividing by the standard deviation.

RFS Prediction and HPV Status Classification Structured clinical variables $\mathbf{x}_{\text{clin}} \in \mathbb{R}^p$ included demographic and treatment-related information (age, sex, tobacco and alcohol use, performance status, M stage, treatment regimen), as well as quantitative imaging-derived biomarkers. The latter comprised conventional parameters such as MTV (primary tumor volume), NTV (positive lymph node volume), T-SUV (primary tumor SUV_{max}), N-SUV (nodal SUV_{max}), TLG (total lesion glycolysis), and NLG (nodal lesion glycolysis). For each patient, a $96 \times 96 \times 96$ 3D patch was extracted around the lesion centroid (computed from lesion masks; if absent, the geometric center was used). CT values were clipped to $[-1000, 3000]$ Hounsfield units and linearly normalized to $[0, 1]$. The final input tensor was stacked as $[\text{CT}, \text{PET}, \text{Mask}] \in \mathbb{R}^{3 \times D \times H \times W}$, where mask labels $\{0, 1, 2\}$ corresponded to background, primary tumor, and nodal disease, respectively. The mask channel was converted to one-hot and the $\{1, 2\}$ channels were used as explicit lesion guidance. Continuous variables were median-imputed for missing values and standardized using Z-Score normalization, while categorical variables were imputed with “Unknown” and one-hot encoded.

3 Method

3.1 GTVp and GTVn Segmentation

Architecture For GTVp and GTVn segmentation on CT/PET, the Scalable and Transferable U-Net (STU-Net) [8] was employed. STU-Net, an extension of the nnU-Net framework known for its strong performance, incorporates refined convolutional blocks to enhance scalability, see Figure 1. Specifically, the basic nnU-Net convolutional blocks were augmented with residual connections [7] to facilitate scaling of model depth. Inspired by Liang et al. [10], the STU-Net small (STU-Net-S) configuration was adopted, with a depth of (1, 1, 1, 1, 1, 1) and width of (16, 32, 64, 128, 256, 256), making it considerably lighter and smaller than the other variants. The preprocessed images (CT/PET) were concatenated as a two-channel input, and the model was trained to simultaneously segment both GTVp and GTVn.

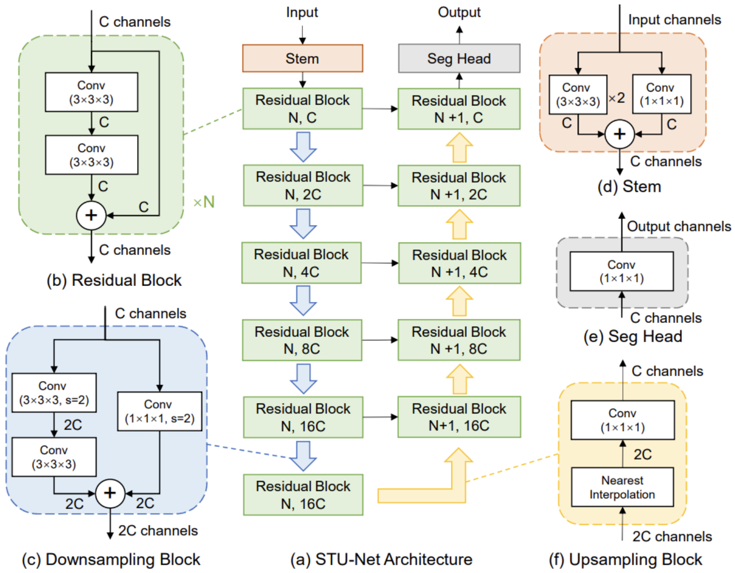

Fig. 1: STU-Net illustration by Huang et al. [8]. (a) architecture overview; (b) residual blocks for large-scale modeling; (c) downsampling in encoder stages; (d–e) stem and segmentation head for channel conversion; (f) weight-free interpolation for task-adaptive upsampling.

Loss The loss function is the combination of cross-entropy and Dice loss.

$$\mathcal{L} = \lambda_{CE}\left(-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=0}^{2} y_{i,c}\log(p_{i,c})\right) + \lambda_{Dice}\left(-\frac{1}{3}\sum_{c=0}^{2}\frac{2\sum_{i=1}^{N} p_{i,c}y_{i,c}}{\sum_{i=1}^{N} p_{i,c}^2 + \sum_{i=1}^{N} y_{i,c}^2 + \epsilon}\right).$$

$p_{i,c}$ is the predicted probability (softmax) that voxel i belongs to class c; $y_{i,c}$ is ground truth (1 if voxel i is class c, else 0). GTVp and GTVn were trained together. We empirically set $\lambda_{\text{CE}} = \lambda_{\text{Dice}} = 1$.

Data Split STU-Net-S was trained with 5-fold cross-validation on all training cases, and a separate 10-fold cross-validation was also trained on the same data.

Optimization Built on nnU-Net, STU-Net automatically configured all hyperparameters. Training for each fold used a batch size of 2, a patch size of (192, 112, 112), and 1,000 epochs on an NVIDIA RTX A6000 GPU. During training, diverse real-time augmentation strategies are utilized, such as geometric transformations (rotation and scaling), intensity adjustments (brightness, contrast, and gamma correction), noise and blur perturbations, low-resolution simulation, and image mirroring.

3.2 RFS Predication

We developed a multimodal survival prediction network that integrates volumetric PET/CT imaging, lesion masks, and structured clinical features, see Figure 2.

Image encoder with explicit lesion guidance. The image backbone was a 3D ResNet-18 [7], which processed the CT and PET channels. In parallel, a lightweight *Lesion-Guidance Module (LGM)* encoded the two lesion channels (GTVp and GTVn ground truth masks) into feature maps matching the backbone's first stage output. After the backbone's initial feature extraction layer, which captures low-level visual patterns, the two streams were fused by element-wise addition to incorporate explicit lesion guidance at the earliest stage.

$$\mathbf{f}_{\text{fuse}} = \phi_{\text{conv1}}([\text{CT, PET}]) + \text{LGM}(\text{Mask}),$$

which was subsequently propagated through the residual blocks, global average pooling, and flattened into an image-level representation $\mathbf{f}_{\text{img}} \in \mathbb{R}^d$.

Clinical Feature Processing. Structured clinical variables $\mathbf{x}_{\text{clin}} \in \mathbb{R}^p$ included continuous features (age, MTV, NTV, SUV, TLG, NLG) and categorical factors (sex, tobacco/alcohol use, performance status, M stage, treatment). Continuous features were median-imputed and Z-Scored, while categorical features were imputed with “Unknown” and one-hot encoded. A single-layer MLP produced a clinical embedding:

$$\mathbf{f}_{\text{clin}} = \text{MLP}(\mathbf{x}_{\text{clin}}) \in \mathbb{R}^{256}.$$

Multimodal fusion and survival head. The final representation was obtained by concatenating image and clinical embeddings,

$$\mathbf{f} = [\mathbf{f}_{\text{img}}; \mathbf{f}_{\text{clin}}],$$

followed by dropout and an MLP head to output logits $\mathbf{z} \in \mathbb{R}^K$ for K discrete time intervals (here $K = 4$; cut points at 769, 1357, and 2626 days). The cut points were chosen based on the empirical distributions of RFS times and ensuring sufficient samples per intervals for stable optimization.

Discrete-time survival modeling. We adopted a discrete-time hazard parameterization [14]:

$$h_k = \sigma(z_k), \quad k = 1, \dots, K,$$

$$S_k = \prod_{j=1}^{k} (1 - h_j), \quad S_0 = 1$$

where h_k is the hazard in interval k, and S_k the survival probability up to interval k. Given observed interval index $y \in \{0, \dots, K-1\}$ and censoring indicator $c \in \{0, 1\}$, the negative log-likelihood was

$$\ell_{\text{base}} = -(1-c)\big[\log S_{y-1} + \log h_y\big] - c \log S_y.$$

We further applied a convex combination with the uncensored component (weight $\alpha = 0.5$) to stabilize optimization:

$$\mathcal{L} = \sum_i \Big((1-\alpha)\,\ell_{\text{base},i} + \alpha\,\ell_{\text{uncens},i} \Big).$$

The patient-level risk score was defined as

$$r = -\sum_{k=1}^{K} S_k,$$

which is monotonically associated with recurrence risk.

Data Split A 5-fold cross-validation was performed on all training cases.

Optimization Optimization used AdamW (weight decay 10^{-2}) with learning rates 1×10^{-4} for the image (ResNet-18+LGM) and clinical+classifier branches (5×10^{-4}). Models were trained for 20 epochs with batch size 16 on an NVIDIA RTX A6000 GPU. Early stopping was guided by validation concordance index (C-index), computed in a Harrell-style fashion from comparable event pairs. The best model per fold was selected by maximal internal validation C-index.

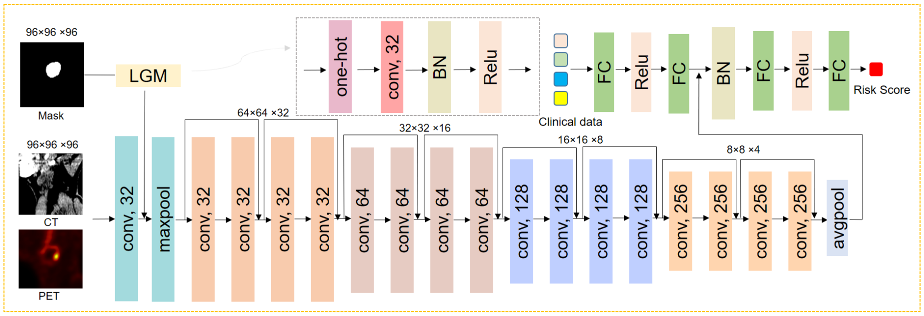

Fig. 2: The illustration of the proposed multimodal survival prediction network. PET/CT images and lesion masks are processed in parallel and fused through an early convolutional layer, while clinical variables are encoded through fully connected layers. The lesion mask is processed by a lesion-guidance module (LGM, grey dashed box, indicated by the grey arrow), which transforms the three-class mask (representing tumor, lymph nodes, and background) into a one-hot representation. This transformed mask is passed through convolution, batch normalization, and ReLU layers to generate lesion-aware feature maps, which enhance the focus on regions of interest. Image features and clinical features are then concatenated and passed through joint classifiers to predict the risk score.

3.3 HPV status classification Task

The HPV classification task employed the same multimodal network and preprocessing pipeline as described for survival prediction—combining volumetric PET/CT imaging, lesion masks via the *LGM*, and structured clinical features. The only architectural difference was the prediction head, which produced a single logit $z \in \mathbb{R}$ corresponding to HPV status ($y \in \{0, 1\}$), with probability $p = \sigma(z)$.

Loss function and calibration. To address class imbalance, training optimized a per-sample weighted binary cross-entropy loss,

$$\mathcal{L}_{\text{BCE-w}} = -\, w(y)\, [\, y \log p + (1-y) \log(1-p)\,],$$

where $w(y) = w_{\text{pos}}$ if $y = 1$ and $w(y) = w_{\text{neg}}$ if $y = 0$. Unless otherwise specified, we set $w_{\text{pos}} = 1$ and $w_{\text{neg}} = \frac{n_{\text{pos}}}{\max(n_{\text{neg}},1)}$ according to training-set class counts. Balanced accuracy at a reference threshold of 0.5 was additionally logged during training and testing.

Data Split A 5-fold cross-validation was performed on all training cases.

Optimization Optimization used AdamW (weight decay 10^{-2}) with learning rates 1×10^{-4} for the image (ResNet-18+LGM) and clinical+classifier branches

(5×10^{-4}). Models were trained for 30 epochs with batch size 16 on an NVIDIA RTX A6000 GPU. The best checkpoint per fold is selected by the maximal validation balanced accuracy.

4 Result

4.1 GTVp and GTVn Segmentation

Segmentation performance was evaluated using the Dice Similarity Coefficient (Dice) for GTVp and GTVn, and the F1 score computed from IoU ($\geq$30%). Segmentation results were first evaluated in cross-validation (Table 1, 2) , after which our algorithms were submitted to the validation leaderboard (Table 3) and final test leaderboard. On average, 10-fold cross-validation did not outperform 5-fold cross-validation; however, applying the 10-fold ensemble strategy yielded improvements on the validation leaderboard set (around 50 unseen cases), particularly for the F1 score of GTVn. Interestingly, although STU-Net-S (both 5-fold and 10-fold) achieved better GTVp segmentation in cross-validation, it showed stronger performance (Dice) on GTVn than GTVp in the leaderboard validation and final leaderboard test. Other STU-Net variants, including STU-Net Base (STU-Net-B), were also trained using 5-fold cross-validation. The cross-validation results showed no significant performance improvement (see Appendix). However, STU-Net-B has a larger number of trainable parameters, higher GPU memory requirements, and requires approximately twice the training time compared to STU-Net-S (see Table 4).

Table 1: Segmentation Performance Using 5-Fold Cross-validation (STU-Net-S)

Metrics	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Dice (GTVp)	0.7307	0.6761	0.6921	0.6934	0.6790	0.6943
Dice (GTVn)	0.6399	0.6621	0.6862	0.6595	0.6632	0.6622
F1 (GTVn)	0.5942	0.6590	0.6768	0.6581	0.6269	0.6430

Table 2: Segmentation Performance Using 10-Fold Cross-validation (STU-Net-S)

Metrics	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Average
Dice (GTVp)	0.6891	0.7295	0.6586	0.7384	0.6710	0.6813	0.6551	0.6292	0.7027	0.6777	0.6833
Dice (GTVn)	0.6523	0.6290	0.6713	0.6170	0.5558	0.6528	0.6532	0.6423	0.6617	0.6232	0.6359
F1 (GTVn)	0.6165	0.6085	0.6748	0.5371	0.5573	0.5990	0.6096	0.6466	0.6078	0.5495	0.6007

Table 3: Leaderboard External Validation of Segmentation Performance

Metrics	Ensemble (5-Fold)	Ensemble (10-Fold)	Final (10-Fold)
Dice (GTVp)	0.7626	0.7653	0.7418
Dice (GTVn)	0.7931	0.7932	0.7640
F1 (GTVn)	0.6385	0.6641	0.6472

Both Emsemble (5-Fold and 10-Fold) were validated using around 50 unseen cases
Final refers to the Final leaderboard test results, based on around 450 unseen cases.

Table 4: Computational Expenses

Network	Param (M)	FLOPs (T)	VRAM (G)	Train (s)	Infer (s)	Depth	Width
STU-Net-S	14.55	0.66	4.4	41	5	(1,1,1,1,1,1)	(16,32,64,128,256,256)
STU-Net-B	58.16	2.62	7.9	78	15	(1,1,1,1,1,1)	(32,64,128,256,512,512)

The comparison of computational cost of STU-Net-S and STU-Net-B
Train: averaged training time per epoch
Infer: averaged inference time per case

4.2 RFS Prediction

The C-index was used to evaluate RFS prediction. The leaderboard validation results were obtained from an ensemble of our 5-fold cross-validation models, showing similar C-index values to those observed in the 5-fold cross-validation. However, a substantial drop in performance was observed on the final leaderboard test set, with the C-index decreasing to 0.5281. see Table 5.

Table 5: RFS Prediction Results

Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average	Validation	Final
C-Index	0.7073	0.6383	0.7009	0.7160	0.7715	0.7068	0.7060	0.5281

Validation refers to the leaderboard validation results, based on around 50 unseen cases.
Final refers to the Final leaderboard test results, based on around 450 unseen cases.

4.3 HPV Status Classification

Balanced accuracy and specificity were used to evaluate HPV status classification. The leaderboard validation results, obtained from an ensemble of our 5-fold cross-validation models, revealed a notable performance gap: balanced accuracy was substantially lower on the leaderboard compared to cross-validation. This trend persisted on the final leaderboard test set, where the balanced accuracy further dropped to 0.5085, highlighting a discrepancy between cross-validation and unseen test performance. see Table 6.

Table 6: HPV Status Classification Results

Fold	**Balanced Accuracy**	**Specificity**
Fold 1	0.8532	1.0000
Fold 2	0.8183	0.6667
Fold 3	0.9361	0.9000
Fold 4	0.7837	0.6154
Fold 5	0.9541	1.0000
Average	0.8691	0.8364
Validation	0.6076	0.9048
Final	0.5085	-

Validation refers to the leaderboard validation results, based on around 50 unseen cases.
Final refers to the final leaderboard test results, based on around 450 unseen cases.
consist of 80% HPV-positive cases and 20% HPV-negative cases.

5 Discussion and Conclusion

In this study, we employed the small variant of STU-Net to efficiently perform simultaneous segmentation of GTVp and GTVn. Across the leaderboard validation, the 10-fold ensemble outperformed the 5-fold ensemble across all three metrics, with a notable 3% increase in the GTVn F1 score. Consequently, the 10-fold ensemble was selected for the final leaderboard test, where it achieved first place overall in the segmentation task. GTVn are smaller and more variable in shapes and locations. The 10-fold ensemble, benefiting from increased model diversity and reduced prediction variance, effectively detects and localizes GTVn. We also evaluated other STU-Net variants, including STU-Net Base (STU-Net-B). Cross-validation results showed no significant performance improvement despite the substantially higher computational cost. Due to submission constraints, evaluation of STU-Net-B on the leaderboard was not conducted at this stage.

Using a 5-fold ensemble of the multimodal survival prediction network, which integrates imaging, predicted lesion masks, and structured clinical features, the C-index on the leaderboard validation was comparable to cross-validation. However, a substantial performance gap was observed on the final leaderboard test, indicating limited model generalization. It is worth noting that RT-Dose and CT-planning scans were not included in our model, and their incorporation could potentially further enhance the performance of RFS prediction. The same multimodal framework was applied to the HPV status classification task. However, substantial overfitting was observed, with leaderboard validation and final leaderboard test metrics—particularly balanced accuracy—considerably worse than cross-validation, likely due to high model complexity and class imbalance. Future work could explore strategies to address class imbalance, such as advanced sampling techniques and to reduce model complexity through architecture simplification or regularization. The study by Celgla et al. [4] has shown that primary tumor features—SUVmax, TotalSUV, MTV, TLG, TLRmax, and

TLRTLG—are informative for predicting HPV status, particularly for identifying HPV-negative tumors. Future work could further explore these features to develop more accurate models for HPV classification.

In this study, we present our solutions for the three tasks of the HECKTOR 2025 challenge. A 10-fold ensemble of STU-Net-S achieved strong segmentation performance. Using the predicted lesion masks, we developed a multimodal network that integrates the original images with clinical features to predict recurrence-free survival (RFS) and classify HPV status. Future work will focus on improving the generalizability of the RFS prediction and HPV classification models.

References

1. Altaf, F., Islam, S.M., Akhtar, N., Janjua, N.K.: Going deep in medical image analysis: concepts, methods, challenges, and future directions. IEEE access **7**, 99540–99572 (2019)
2. Anderson, G., Ebadi, M., Vo, K., Novak, J., Govindarajan, A., Amini, A.: An updated review on head and neck cancer treatment with radiation therapy. Cancers **13**(19), 4912 (2021)
3. Andrearczyk, V., Oreiller, V., Abobakr, M., Akhavanallaf, A., Balermpas, P., Boughdad, S., Capriotti, L., Castelli, J., Cheze Le Rest, C., Decazes, P., 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, pp. 1–30. Springer (2022)
4. Cegla, P., Currie, G., Wroblewska, J.P., Kazmierska, J., Cholewinski, W., Jagiello, I., Matuszewski, K., Marszalek, A., Kubiak, A., Golusinski, P., et al.: [18f] fdg pet/ct imaging and hematological parameters can help predict hpv status in head and neck cancer. Nuklearmedizin-NuclearMedicine **64**(01), 22–31 (2025)
5. Dm, P.: Global cancer statistics, 2002. CA Cancer J Clin **55**, 74–108 (2005)
6. Escott, E.J.: Role of positron emission tomography/computed tomography (pet/ct) in head and neck cancer. Radiologic Clinics **51**(5), 881–893 (2013)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Huang, Z., Wang, H., Deng, Z., Ye, J., Su, Y., Sun, H., He, J., Gu, Y., Gu, L., Zhang, S., et al.: Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. arXiv preprint arXiv:2304.06716 (2023)
9. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
10. Liang, X., Huang, J., Han, L., Zhang, T., Wang, X., Gao, Y., Lu, C., Cai, L., Tan, T., Mann, R.: Dpdnet: An dual-prompt-driven network for universal pet-ct segmentation. arXiv preprint arXiv:2507.07126 (2025)
11. Mody, M.D., Rocco, J.W., Yom, S.S., Haddad, R.I., Saba, N.F.: Head and neck cancer. The Lancet **398**(10318), 2289–2299 (2021)
12. Myronenko, A., Siddiquee, M.M.R., Yang, D., He, Y., Xu, D.: Automated head and neck tumor segmentation from 3d pet/ct hecktor 2022 challenge report. In: 3D

Head and Neck Tumor Segmentation in PET/CT Challenge, pp. 31–37. Springer (2022)

13. Saeed, N., Hassan, S., Hardan, S., Aly, A., Taratynova, D., Nawaz, U., Khan, U., Ridzuan, M., Andrearczyk, V., Depeursinge, A., Xie, Y., Eugene, T., Metz, R., Dore, M., Delpon, G., Papineni, V.R.K., Wahid, K., Dede, C., Ali, A.M.S., Sjogreen, C., Naser, M., Fuller, C.D., Oreiller, V., Jreige, M., Prior, J.O., Rest, C.C.L., Tankyevych, O., Decazes, P., Ruan, S., Tanadini-Lang, S., Vallières, M., Elhalawani, H., Abgral, R., Floch, R., Kerleguer, K., Schick, U., Mauguen, M., Bourhis, D., Leclere, J.C., Sambourg, A., Rahmim, A., Hatt, M., Yaqub, M.: A multimodal and multi-centric head and neck cancer dataset for segmentation, diagnosis, and outcome prediction (2025), https://arxiv.org/abs/2509.00367
14. Zhang, Y., Xu, Y., Chen, J., Xie, F., Chen, H.: Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. arXiv preprint arXiv:2401.01646 (2024)

Appendix

Segmentation Performance Using 5-Fold Cross-validation (STU-Net-B)

Metris	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Dice (GTVp)	0.7254	0.6744	0.6899	0.6974	0.6693	0.6913
Dice (GTVn)	0.6323	0.6885	0.6870	0.6831	0.6464	0.6675
F1 (GTVn)	0.5843	0.6751	0.6789	0.6481	0.6258	0.6424

The experiment setup is the same as 5-fold cross validation using STU-Net-S