A Two-Stage Coarse-to-Fine Ensembling Segmentation Framework with Multi-Channel CT Enhancement for Head and Neck Tumor and Lymph Segmentation in PET and CT Image

Chuanyi Huang[1][0009−0009−3223−0082] and Lisheng Wang[1*][0000−0003−3234−7511]

School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai 20024 , China
lswang@sjtu.edu.cn

Abstract. Head and neck (H&N) cancer segmentation from PET/CT is challenging due to heterogeneous imaging protocols across centers and the small proportion of tumor and lymph node volumes relative to the full field-of-view. We propose a two-stage coarse-to-fine framework for automatic segmentation of primary tumors (GTVp) and metastatic lymph nodes (GTVn) in the HECKTOR 2025 challenge. The framework first applies a head localization stage using an nnUNet to extract a coarse region-of-interest (ROI). In the fine segmentation stage, we integrate predictions from two complementary backbones: nnUNetResEncUNetLarge and MONAI-based SegResNet, both trained with five-fold cross-validation. To further enhance tumor delineation, especially on CT modality and across centers, we introduce multi-channel CT representations by concatenating raw CT, its squared intensity, cubic-root intensity, and PET as four input channels for nnUNetResEncUNetLarge. This design improves sensitivity to tumor intensity patterns and robustness against inter-center heterogeneity. The framework was evaluated quantitatively on the official test set of task 1 for the HECKTOR2025 challenge, achieving a GTVp DSC of 0.7341, a GTVn aggregated DSC of 0.7312, and a GTVn aggregated F1 score of 0.7260 as team SJTU_lab426.

Keywords: Head an Neck Cancer Automatic segmentation · Coarse-to-Fine Segmentation · Multi-Channel CT Enhancement · nnUNet · SegResNet

1 Introduction

Head and Neck (H&N) cancer ranks among the top five most prevalent cancer types globally [1]. Accurate segmentation of the primary gross tumor volume (GTVp) and involved lymph nodes (GTVn) is critical for radiotherapy treatment planning and outcome prediction[2, 3]. Traditionally, tumor delineation is performed manually by clinicians, which is labor-intensive, time-consuming, and prone to variability.

With the availability of 3D multimodal imaging modalities such as CT and PET, automatic segmentation using deep learning has become a promising approach. PET highlights metabolic activity but at relatively low resolution, whereas CT provides high-resolution anatomical details. Leveraging both modalities has shown improved performance in H&N tumor delineation.

The HEad and neCK TumOR (HECKTOR) challenge, organized in three editions in conjunction with MICCAI, serves as a benchmark for automated head and neck (H&N) cancer analysis using multimodal PET/CT imaging. In the latest edition, HECKTOR 2025, the evaluation protocol now emphasizes not only Dice similarity coefficient (DSC) but also lymph detection performance and cross-center robustness.

Previous editions, such as HECKTOR 2022, highlighted the difficulty of accurate segmentation under heterogeneous imaging conditions. The winning entry[4] (Myronenko et al.) employed an ensemble of MONAI-based SegResNet[5] models trained with five-fold cross-validation, achieving an aggregated DSC of 0.788. The runner-up solution leveraged nnUNet[6] with coarse-to-fine processing and multi-model ensembling, demonstrating the effectiveness of localization strategies and strong baselines. These top solutions reveal two essential insights: (1) accurate ROI extraction and model ensembling are key to achieving strong performance; (2) although PET provides reliable localization, its tumor boundaries are often indistinct. In contrast, CT offers richer boundary information but suffers from limited contrast between lesions and surrounding tissues, as well as inter-center intensity variations that reduce generalization. Therefore, enhancing CT-based representations is crucial to fully exploit its structural advantages while mitigating these limitations.

Motivated by these observations, our method adopts a two-stage coarse-to-fine strategy with nnUNet for robust head localization, followed by an ensemble of nnUNetResEncUNetLarge and MONAI SegResNet for fine segmentation. To strengthen CT-driven delineation and improve robustness across heterogeneous multi-center data, we introduce CT enhancement by augmenting the raw CT with squared and cubic-root intensity transforms, concatenated with PET as four-channel input. This design explicitly enriches CT representation, facilitating sharper lesion boundary detection and stronger cross-center generalization.[7]

2 Method

2.1 Dataset

The HECKTOR 2025 challenge[8] provides an extended multimodal dataset[9] for head and neck (H&N) tumor lesion segmentation and diagnosis which contains 680 cases from 7 centers.

Each case comprises a co-registered 3D PET/CT pair and a voxel-wise binary mask delineating the primary tumor and any metastatic lymph nodes. Scanning coverage differs across centers,as shown in Fig 1: some studies are restricted to the head-and-neck region, others extend to the thorax or upper abdomen, and

a subset provides whole-body acquisitions, introducing substantial anatomical-coverage heterogeneity.

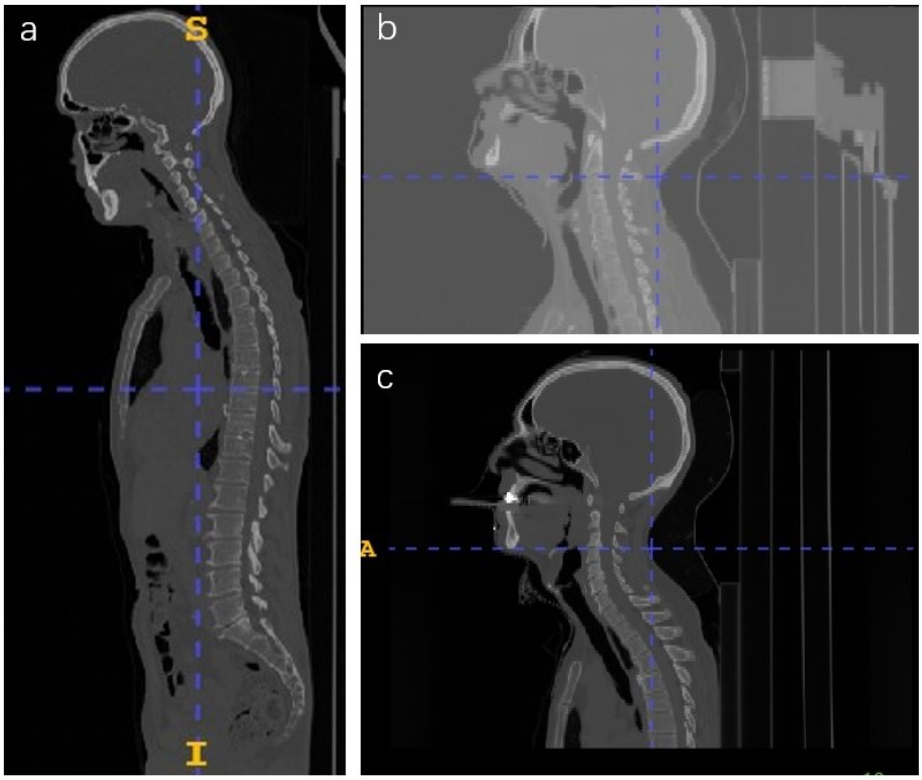

Fig. 1. Variation of image size from 7 centers. (a) whole upper body. (b) only head region. (c) head and upper body.

2.2 Overview of the Proposed Framework

Our method follows a two-stage coarse-to-fine pipeline designed to efficiently localize the head-and-neck (H&N) region and perform accurate tumor and segmentation. The overall workflow can be seen in Fig2 and can be summarized as follows:

1. **Head Localization (Coarse Stage):** An nnUNet model is trained to detect the H&N region from the full-body CT scan. A bounding box is generated to crop the relevant region-of-interest (ROI), thereby reducing input size and eliminating irrelevant anatomy.

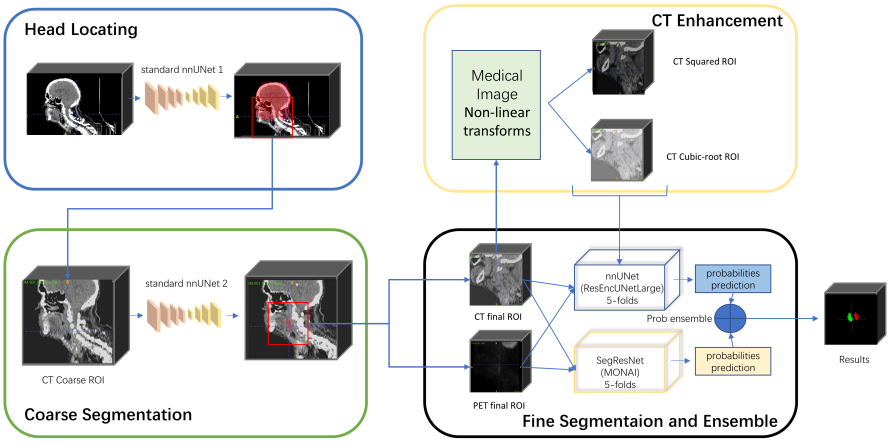

Fig. 2. two-stage coarse-to-fine pipeline of the whole framework

2. **Coarse Segmentation Stage:** We define the coarse ROI as the entire head-and-neck volume within which tumors and metastatic lymph nodes are expected to reside, and we obtain this initial region using a standard nnUNet model trained for coarse segmentation.
3. **Fine Segmentation Stage:** Within the localized ROI, two complementary segmentation backbones are applied:
   – *nnUNetResEncUNetLarge* – a residual encoder-decoder variant of nnUNet.
   – *MONAI SegResNet* – a widely adopted encoder-decoder architecture with deep supervision.

   Both models are trained with five-fold cross-validation. Their predictions are ensembled to produce robust final outputs.
4. **Multi-Channel CT Enhancement:** To strengthen CT-driven lesion delineation and improve robustness across heterogeneous multi-center datasets, we augment the CT modality by generating squared (x^2) and cubic-root ($x^{1/3}$) intensity channels in addition to the raw CT (x). Together with PET (y), these form a four-channel input $[x, x^2, x^{1/3}, y]$. Then we used it as multi-channel inputs for nnUNetResEncUNetLarge.
5. **Inference and Post-processing:** During inference, predictions from the ensemble are averaged. Small isolated false positives are removed via connected-component analysis.

2.3 Network Architecture and Details

Head Localization Model To cope with the large variability along the z-axis (ranging from a few dozen to several hundred slices) while the transverse plane remains 512 × 512, we first isolated a dedicated head-and-neck ROI. Because

heads always appear near the top but may be left- or right-shifted (due to different acquisition centers), an automatically defined bounding box is unreliable. We therefore adopted a coarse-to-fine strategy: the initial step simply localizes the approximate head region rather than producing an exact segmentation. For this localization task we trained a standard nnUNet using CT images only. Each volume is resampled—starting from the top of the original scan—to a uniform grid of $256 \times 256 \times 128$ voxels at $2 \times 2 \times 3$ mm spacing, employing B-spline interpolation. Forty CT scans were randomly selected from the training set and manually labeled for the head region in just a few hours. Prior to training, all intensities were clipped to [–110, 190] HU and normalized to [–1, 1].

Although isotropic spacing is often recommended, we followed nnU-Net's default preprocessing strategy, which determines anisotropic voxel spacing based on the median physical resolution of the dataset. This choice preserves the native resolution characteristics of head-and-neck scans and avoids unnecessary resampling artifacts.

The detailed parameters of the head-locating model are listed in the table1 below.

Table 1. nnU-Net 3D Full Resolution Configuration for Head Localization Model

Property	**Configuration**
Preprocessor	DefaultPreprocessor
Batch Size	2
Patch Size	$96 \times 160 \times 160$
Median Image Size	$128 \times 256 \times 256$
Spacing	(3.0, 2.0, 2.0) mm
Normalization	Z-Score normalization (no mask)
Resampling	B-spline (order=3, data),
	Nearest-neighbor (order=1, segmentation)
Network Class	PlainConvUNet (3D)
Number of Stages	6
Features per Stage	[32, 64, 128, 256, 320, 320]
Kernel Sizes	$3 \times 3 \times 3$ (all stages)
Strides	{(1,1,1), (2,2,2), (2,2,2), (2,2,2), (2,2,2), (1,2,2)}
Convolutions per Stage (Encoder)	[2, 2, 2, 2, 2, 2]
Convolutions per Stage (Decoder)	[2, 2, 2, 2, 2]
Convolution Bias	True
Normalization Layer	InstanceNorm3d ($\epsilon = 10^{-5}$, affine=True)
Nonlinearity	LeakyReLU (inplace=True)
Dropout	None
Loss Function	Cross-entropy + Dice (batch Dice = False)

Coarse Segmentation Model To define the coarse region that encompasses both primary tumors and metastatic lymph nodes within the head-and-neck area, we employ a straightforward nnU-Net model. CT and PET volumes are first cropped according to this coarse ROI. Intensities inside the CT crop are clipped to [–110, 190] HU and then linearly scaled to [–1, 1], whereas PET values are simply normalized to [0, 1] at the same coarse level. Once the coarse ROI is available, we derive the subsequent fine ROI from the union of all tumor and lymph-node labels. The bounding box of the fine ROI is centered on the centroid of these labels (blue square in Fig. 3). This cube is resampled to 144 × 144 × 144 voxels at 1 × 1 × 1 mm spacing via B-spline interpolation—identical to last year's setup—yielding a 144 mm isotropic volume comfortably larger than an average neck or chin diameter and thus sufficient to enclose every target. The specific parameters of the coarse-segmentation model are provided in the table2 below.

Table 2. nnU-Net 3D Full Resolution Configuration for Coarse Segmentation Model

Property	**Configuration**
Preprocessor	DefaultPreprocessor
Batch Size	2
Patch Size	$128 \times 128 \times 128$
Median Image Size	$144 \times 144 \times 144$
Spacing	(2.0, 2.0, 2.0) mm
Normalization	Z-Score normalization (2 channels, no mask)
Resampling	B-spline (order=3, data),
	Nearest-neighbor (order=1, segmentation)
Network Class	PlainConvUNet (3D)
Number of Stages	6
Features per Stage	[32, 64, 128, 256, 320, 320]
Kernel Sizes	$3 \times 3 \times 3$ (all stages)
Strides	{(1,1,1), (2,2,2), (2,2,2), (2,2,2), (2,2,2), (2,2,2)}
Convolutions per Stage (Encoder)	[2, 2, 2, 2, 2, 2]
Convolutions per Stage (Decoder)	[2, 2, 2, 2, 2]
Convolution Bias	True
Normalization Layer	InstanceNorm3d ($\epsilon = 10^{-5}$, affine=True)
Nonlinearity	LeakyReLU (inplace=True)
Dropout	None
Loss Function	Cross-entropy + Dice (batch Dice = False)

Fine Segmentation Models The fine ROI tightly encloses tumors and metastatic lymph nodes. CT and PET volumes are resampled from the original study using

this ROI; CT intensities are clipped to $[-110, 190]$ HU and linearly scaled to $[-1, 1]$, whereas PET values are normalized to $[0, 1]$.

Fine segmentation proceeds in two stages: distributed training followed by simple probability averaging.

The specific parameters of the coarse-segmentation model are provided in table3 and table4

Table 3. nnU-Net ResUNet 3D Full Resolution Configuration for Fine Segmentation Models

Property	**Configuration**
Preprocessor	DefaultPreprocessor
Batch Size	4
Patch Size	$160 \times 160 \times 160$
Median Image Size	$144 \times 144 \times 144$
Spacing	(1.0, 1.0, 1.0) mm
Normalization	CT Normalization + Z-Score (2 channels, no mask)
Resampling	B-spline (order=3, data),
	Nearest-neighbor (order=1, segmentation)
Network Class	ResidualEncoderUNet (3D)
Number of Stages	6
Features per Stage	[32, 64, 128, 256, 320, 320]
Kernel Sizes	$3 \times 3 \times 3$ (all stages)
Strides	{(1,1,1), (2,2,2), (2,2,2), (2,2,2), (2,2,2), (2,2,2)}
Residual Blocks per Stage	[1, 3, 4, 6, 6, 6]
Convolutions per Stage (Decoder)	[1, 1, 1, 1, 1]
Convolution Bias	True
Normalization Layer	InstanceNorm3d ($\epsilon = 10^{-5}$, affine=True)
Nonlinearity	LeakyReLU (inplace=True)
Dropout	None
Loss Function	Cross-entropy + Dice (batch Dice = False)

1. **Distributed training.** Two architectures are trained with 5-fold cross-validation each, producing 10 models in total:
   (1)nnUNet Large ResUNet: 4-channel input

$$\big[x,\ x^2,\ x^{1/3},\ y\big],$$

   where x denotes CT and y denotes PET. The additional x^2 and $x^{1/3}$ channels strengthen CT-driven lesion delineation and improve robustness across heterogeneous multi-center data.

(2)MONAI SegResNet: 2-channel input [CT, PET]. Every model outputs two independent binary probability maps (tumor foreground and lymph-node foreground).

2. **Probability averaging.** The 20 resulting probability maps (2 classes × 10 models) are averaged class-wise; the two averaged volumes are then thresholded to obtain the final binary labels.

Table 4. AutoSeg3D (SegResNetDS) Configuration for Fine Segmentation

Property	**Configuration**
Input Modality	CT (2 channels, plus optional PET)
Output Classes	3 (tumor, lymph_node, background)
ROI Size	$144 \times 144 \times 144$
Spacing (Median / Lower / Upper)	(1.0, 1.0, 1.0) mm
Normalization	Range normalization, crop foreground
Batch Size	1
Optimizer	AdamW (weight decay 1e−5)
Loss Function	Dice + Cross-Entropy Loss
	(include background, squared pred, softmax)
AMP	True
Network Architecture	SegResNetDS
Initial Filters	32
Downsampling Blocks	[1, 2, 2, 4, 4]
Residual Depth	4
Normalization	Instance Normalization (INSTANCE_NVFUSER)
Input Channels	2
Output Channels	3
Nonlinearity	LeakyReLU (in MONAI default)
Channels Last	True
Data Augmentation	Random crop, crop foreground,
	auto scale ROI/filter allowed
Resampling	Disabled (native resolution $1mm^3$)
Intensity Bounds	[0.3204, 0.6318]

2.4 Training Details

All models were trained on two NVIDIA RTX 3090 GPUs with 24 GB memory each. The training configurations in terms of number of epochs and loss functions are summarized below:

– **Head locating nnU-Net:** 500 epochs, using Dice + Cross-Entropy loss.

- **Coarse segmentation nnU-Net:** 1000 epochs, using Dice + Cross-Entropy loss.
- **Fine segmentation nnU-Net (Residual Encoder UNet):** 1000 epochs, using Dice + Cross-Entropy loss.
- **Fine segmentation SegResNet:** 300 epochs, using Dice + Cross-Entropy loss.

The loss functions are defined as follows:

Dice Loss For a single class c, the Dice coefficient is

$$\mathrm{Dice}_c = \frac{2\sum_i p_{i,c} g_{i,c}}{\sum_i p_{i,c}^2 + \sum_i g_{i,c}^2},$$

where $p_{i,c}$ is the predicted probability for voxel i and class c, $g_{i,c}$ is the ground truth (one-hot), and i sums over all voxels. The Dice loss is then

$$\mathcal{L}_{Dice} = 1 - \frac{1}{C}\sum_{c=1}^{C} \mathrm{Dice}_c.$$

Cross-Entropy Loss

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} g_{i,c} \log p_{i,c},$$

where N is the total number of voxels and C is the number of classes.

The loss formulation used in AutoSeg3D is mathematically identical to that of nnU-Net, both combining Dice and Cross-Entropy losses in equal proportion. AutoSeg3D adopts the same Dice + CE structure through MONAI's implementation, which serves as a unified loss interface across multiple sub-models. Therefore, while AutoSeg3D provides a framework-level wrapper for ensemble training and logging, it does not alter the underlying loss definition or weighting scheme.

3 Results and Discussion

Sanity-check Leaderboard

Table 5 summarizes our four preliminary submissions on the public sanity-check set (Task 1). Models are listed in the order they were uploaded; the ranking therefore reflects earlier rather than final performance.

Validation Leaderboard

Table 6 reports our two submissions on the hidden validation set (Task 1). The first row employs the full pipeline (nnUNet Large ResUNet + MONAI SegResNet + CT enhancement), while the second row uses the same two backbones without the additional CT channels.

Table 5. Sanity-check results (Task 1). The column *GTVp Dice* denotes primary-tumor Dice; *GTVn Aggr. Dice / F1* denote lymph-node metrics. Abbreviations: **nL-RU** = nnUNet Large ResUNet, **SRN** = SegResNet (MONAI), **CTE** = CT Enhancement. The first row corresponds to the full ensemble and enhancement pipeline, subsequent rows progressively ablate components.

Model	GTVp Dice	GTVn Aggr. Dice	GTVn Aggr. F1
nL-RU + SRN + CTE	0.8851	0.7973	0.8000
SRN + CTE	0.8778	0.7755	0.8000
nL-RU + CTE	0.8727	0.7503	0.8000
nL-RU (baseline)	0.8718	0.7222	0.7500

Table 6. Validation results (Task 1). Metrics are computed on the hidden validation set. Abbreviations: **nL-RU** = nnUNet Large ResUNet, **SRN** = SegResNet (MONAI), **CTE** = CT Enhancement.

Model	GTVp Dice	GTVn Aggr. Dice	GTVn Aggr. F1
nL-RU + SRN + CTE	0.7547	0.7743	0.6235
nL-RU + SRN	0.7217	0.7820	0.5975

Test Leaderboard

Table 7 reports our official submission on the hidden test set (Task 1). The results are for the full pipeline (nnUNet Large ResUNet + MONAI SegResNet + CT Enhancement).

Table 7. Test set results (Task 1). Metrics are computed on the official hidden test set. Abbreviations: **nL-RU** = nnUNet Large ResUNet, **SRN** = SegResNet (MONAI), **CTE** = CT Enhancement.

Model	GTVp Dice	GTVn Aggr. Dice	GTVn Aggr. F1
nL-RU + SRN + CTE	0.7341	0.7312	0.7260

Discussion

The experimental results highlight several important findings. First, the coarse-to-fine design proves effective: the initial head-localization and coarse segmentation stages substantially reduce irrelevant anatomy, simplifying the fine segmentation task. This strategy not only accelerates training and inference but also mitigates false positives in distant regions. Second, integrating complementary backbones (nnU-Net Large ResUNet and MONAI SegResNet) consistently improves robustness. The sanity-check leaderboard (Table 5) shows that the ensemble with CT enhancement achieves the best Dice scores for both GTVp and

GTVn, confirming the benefit of multi-model fusion. Importantly, this ensemble-based strategy was used throughout all final submissions, ensuring consistency between model description and reported results.

Another key observation lies in the role of CT enhancement. By concatenating squared and cubic-root intensity transforms with raw CT, the model becomes more sensitive to subtle tumor-intensity variations while remaining robust to inter-center heterogeneity. This robustness is further supported by the model's stable validation performance across institutions included in the official split, suggesting improved generalization to varying scanner characteristics and acquisition protocols.

Nevertheless, the effect of CT enhancement differs between tasks. As shown in Tables 5 and 6, the enhanced CT input yields higher GTVp Dice but slightly lower GTVn Dice. This discrepancy can be attributed to the intrinsic contrast differences between primary tumors and metastatic lymph nodes. While intensity-based transformations enhance tumor boundaries, they may amplify CT noise or intensity bias near small lymph nodes, which exhibit lower contrast and greater inter-center variability. Consequently, CT enhancement tends to favor GTVp delineation but may compromise the consistency of GTVn predictions. Future work could address this trade-off via adaptive channel selection or loss reweighting between tumor and lymph-node classes.

Our approach ranked 5th overall on the HECKTOR2025 test set and demonstrated exceptional performance in metastatic lymph-node detection—the clinically critical task of identifying GTVn lesions. Specifically, our method achieved the highest GTVn Aggregated F1 score (0.7260), surpassing the next best result by 9.4% (0.7260 vs 0.6638). This significant improvement in detection F1—which jointly reflects precision and recall—suggests that our multi-model ensemble combined with CT enhancement provides superior sensitivity for localizing small metastatic lesions in heterogeneous PET/CT data.

Despite these gains, several limitations remain. First, although the ensemble improves stability, it increases computational cost during training and inference. Second, the reliance on intensity-based CT transformations may limit generalization to unseen imaging protocols. More sophisticated methods, such as self-supervised pretraining or radiomics-inspired feature augmentation, could further strengthen cross-center robustness. Finally, our current post-processing is limited to connected-component filtering; incorporating uncertainty estimation or anatomical priors could more effectively reduce false positives.

Overall, the results demonstrate that combining coarse-to-fine localization, backbone ensembling, and CT enhancement provides a robust and generalizable solution for multi-center head-and-neck cancer segmentation. Future work will focus on improving lymph-node delineation and reducing the computational burden of ensemble inference.

References

1. Parkin, D.M., Bray, F., Ferlay, J., Pisani, P.: Global cancer statistics, 2002. CA: a cancer journal for clinicians **55**(2), 74–108 (2005)
2. Vallieres, M., Kay-Rivest, E., Perrin, L.J., Liem, X., Furstoss, C., Aerts, H.J., Khaouam, N., Nguyen-Tan, P.F., Wang, C.S., Sultanem, K., et al.: Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Scientific reports **7**(1), 10117 (2017)
3. Bogowicz, M., Riesterer, O., Stark, L.S., Studer, G., Unkelbach, J., Guckenberger, M., Tanadini-Lang, S.: Comparison of pet and ct radiomics for prediction of local tumor control in head and neck squamous cell carcinoma. Acta oncologica **56**(11), 1531–1536 (2017)
4. 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)
5. Myronenko, A.: 3d mri brain tumor segmentation using autoencoder regularization. In: International MICCAI brainlesion workshop. pp. 311–320. Springer (2018)
6. 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)
7. Chang, Q., Yan, Z., Zhou, M., Qu, H., He, X., Zhang, H., Baskaran, L., Al'Aref, S., Li, H., Zhang, S., et al.: Mining multi-center heterogeneous medical data with distributed synthetic learning. Nature communications **14**(1), 5510 (2023)
8. Saeed, N., Hassan, S., Hardan, S., Aly, A., Taratynova, D., Nawaz, U., Khan, U., Ridzuan, M., Andrearczyk, V., Depeursinge, A., et al.: A multimodal and multi-centric head and neck cancer dataset for segmentation, diagnosis and outcome prediction. arXiv preprint arXiv:2509.00367 (2025)
9. 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