# Molecular-Information-Guided Framework for Head and Neck Tumor and Lymph Node Segmentation in PET/CT Images

Jee Yoon Jung[1 [0009−0008−5604−2193]], Min Jeong Cho[1,2,3], and Jae Sung Lee[1,2,3,4 [0000−0001−7623−053X]]

[1] Department of Nuclear Medicine, Seoul National University College of Medicine, Seoul, South Korea
[2] Interdisciplinary Program in Bioengineering, Seoul National University Graduate School, Seoul, South Korea
[3] Integrated Major in Innovative Medical Science, Seoul National University College of Medicine, Seoul, South Korea
[4] Brightonix Imaging Inc., Seoul, South Korea

**Abstract.** Accurate segmentation of Head and Neck (H&N) tumor and lymph node in PET/CT images is essential for diagnosis and treatment planning, yet remains challenging due to heterogeneous lesion morphology and variable physiological uptake. In this paper, we propose a 3D U-Net-based architecture with a PET-guided Spatial Attention Module (PSAM). PSAM consumes the PET image to generate a spatial attention map that gates encoder skip features. Furthermore, we use Squeeze-and-Excitation (SE) Normalization, which dynamically recalibrates channel responses and improves multimodal fusion. Notably, we introduce a novel molecular-information-guided preprocessing pipeline, which reduces the input volume size. This design enables modality-aware modeling and aims for robust, generalizable segmentation of both primary Gross Tumor Volume (GTVp) and nodal Gross Tumor Volumes (GTVn) across multi-center PET/CT data. We obtain a mean Dice score of 0.6073 with class-wise Dice scores 0.7133 (GTVp) and 0.5013 (GTVn) on the validation set. On the official test set, it obtains a Dice of 0.3803 (GTVp) and 0.3733 (GTVn). Further studies are required to improve generalization performance across multi-center datasets.

**Keywords:** HECKTOR2025 · MICCAI2025 · Segmentation · Challenge · Deep Learning · 3D CT · 3D PET

## 1 Introduction

In recent years, H&N tumor segmentation has become a critical issue in cancer research, clinical diagnosis, and surgical planning. Medical images such as CT and PET play an important role in radiotherapy and treatment. However, traditional manual segmentation methods are time-consuming, labor-intensive, and subject to inter-expert variability. To overcome these challenges, significant

progress has been made in automated segmentation methods, particularly those based on convolutional neural networks (CNNs).

One of the most successful models in this domain is U-Net [1, 2]. Its encoder-decoder architecture with skip connections has demonstrated exceptional performance in medical image segmentation by effectively combining low-level and high-level features. Despite its success, U-Net can face limitations in extracting features for segmenting complex and diverse lesion morphologies. For instance, segmenting complex H&N tumors or subtle lymph nodes can be challenging with a single-scale feature approach. Furthermore, the performance of any model is highly dependent on the quality of its input data and the preprocessing methods used, which is amplified in HECKTOR challenges [3, 4].

To address these limitations, we augment the U-Net architecture with a PSAM network. This attention module [5] allows the model to simultaneously consider PET features, enabling it to better recognize complex structures and subtle lesions within H&N tumor PET/CT images. We use SE Normalization [6] rather than Instance Normalization [7] which maximizes representational capacity by assigning dynamic weights to each channel of the feature maps. We propose a molecular-information-guided preprocessing pipeline that standardizes image geometry, localizes the field-of-view, and applies a novel cranial "brain-peak cut" to suppress non-relevant uptake. This pipeline reduces variability across centers and provides cleaner, more consistent inputs for robust model training.

## 2 Method

### 2.1 Dataset

The HECKTOR 2025 challenge provides 680 paired PET/CT patient cases from seven centers, with no bounding boxes supplied. Each case includes one 3D PET and one 3D CT image, together with a label volume delineating Gross Tumor Volume (GTVp) and nodal Gross Tumor Volumes (GTVn). We used 600 cases for training and 80 cases for validation. The center-wise distribution is summarized in Table 1. This multi-center PET/CT images are based on the recently released HECKTOR dataset described in [8].

**Table 1.** Number of patients for train/validation/total in seven centers

| Center | Patients (train/val/total) | Center | Patients (train/val/total) |
|---|---|---|---|
| CHUM | 44/12/56 | HMR | 14/4/18 |
| CHUP | 64/8/72 | MDA | 364/32/396 |
| CHUS | 60/12/72 | USZ | 7/4/11 |
| HGJ | 47/8/55 | Total | 600/80/680 |

The dataset spans seven institutions with unequal per-center case counts and heterogeneous axial coverage—ranging from head-only to whole-body acquisitions—leading to center-specific differences in field-of-view and uptake patterns

(Fig. 1). To ensure that model selection reflects performance across institutions, we constructed the validation split in a *center-proportional* manner.

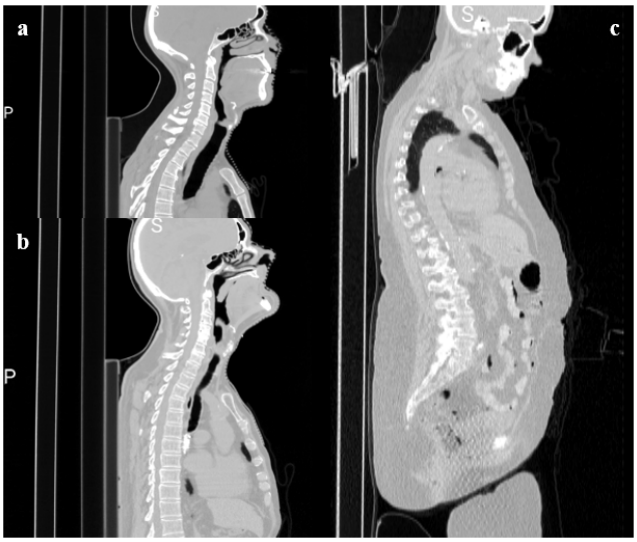


**Fig. 1.** Variation of data size across seven centers: (a) head only; (b) head and upper body; (c) whole body

### 2.2 Data Preprocessing

All CT/PET images were pre-registered. We introduce a molecular-information-guided pipeline consisting of geometric resampling, cropping, PET-based localization, brain peak cut, normalization, and restoration. The network input is a two-channel volume (CT, PET) of fixed size $256 \times 256 \times 112$ (x–y–z).

**Geometric resampling and voxelwise alignment** CT volumes were resampled to $\Delta = (1.0, 1.0, 3.0)$ mm using B-spline interpolation; PET volumes were resampled linearly and then regridded onto the CT lattice so that spacing, size, origin, and direction match exactly (voxelwise alignment).

**Image cropping** We swept several fixed crop sizes and recorded a binary outcome per size— **Satisfied (✔)** if both targets (GTVp=1, GTVn=2) were fully contained for all cases, and **Not satisfied (✖)** otherwise. Based on this criterion, we selected the smallest size, $\mathbf{256 \times 256 \times 112}$; the detailed pass/fail outcomes for all candidates are summarized in Table 2.

**Table 2.** Crop-size sweep with a pass/fail criterion.

| Crop size (H×W×D) | Satisfied |
|---|---|
| 192×192×112 | ✖ |
| 208×208×112 | ✖ |
| 224×224×112 | ✖ |
| 256×256×64 | ✖ |
| 256×256×96 | ✖ |
| 256×256×112 | ✔ |
| 128×128×128 | ✖ |

For each candidate crop size, the "Satisfied/Not satisfied" label in Table 2 was determined strictly on a per-case basis: a crop was marked as "Satisfied" only if *both* GTVp and GTVn were fully contained for *every* case in the dataset. No exceptions were allowed.

**Molecular-information-guided in-plane localization** Let $P \in R^{Z\times Y\times X}$ denote the resampled PET. We detect the superior cranial boundary by scanning from the top slice for suprathreshold uptake with $\tau = 0.5$:

$$z_{\text{top}} = \max\left\{ z \in \{0,\dots,Z-1\} \;\middle|\; \max_{x,y} P[z,y,x] > \tau \right\}. \tag{1}$$

From the 20-slice slab immediately inferior to $z_{\text{top}}$,

$$\mathcal{S} = \{\max(0, z_{\text{top}} - 19), \dots, z_{\text{top}}\}. \tag{2}$$

We compute the activity centroid over the suprathreshold set

$$\Omega = \big\{(z,y,x) \in \mathcal{S} \times \{0,\dots,Y-1\} \times \{0,\dots,X-1\} \;\big|\; P[z,y,x] > \tau\big\}. \tag{3}$$

$$\bar{x} = \frac{1}{|\Omega|} \sum_{(z,y,x)\in\Omega} x, \qquad \bar{y} = \frac{1}{|\Omega|} \sum_{(z,y,x)\in\Omega} y. \tag{4}$$

We then crop a $256 \times 256$ in-plane window centered at $(\bar{x}, \bar{y})$. Defining half-widths $r_x = r_y = 128$, the coordinates are

$$x_1 = \min\big(\max(\lfloor \bar{x} \rfloor - r_x,\, 0),\, \max(0,\, X - 2r_x)\big), \quad x_2 = x_1 + 2r_x, \tag{5}$$

$$y_1 = \min\big(\max(\lfloor \bar{y} \rfloor - r_y,\, 0),\, \max(0,\, Y - 2r_y)\big), \quad y_2 = y_1 + 2r_y. \tag{6}$$

Boundary clamping and zero padding were applied if $X < 256$ or $Y < 256$.

**PET profile-guided brain peak cut** Within the in-plane crop we compute the PET $z$-profile:

$$s(z) = \sum_{x,y} P[z,y,x]. \tag{7}$$

Its peak index is

$$z_{\text{peak}} = \arg\max_{z} s(z). \tag{8}$$

We anchor the $z$-window to the top if strong uptake persists on the most superior slice or the peak is shallow:

$$\text{mean}\big(P[Z-1,:,:]\big) > 0.1 \quad \text{or} \quad z_{\text{peak}} \leq 50. \tag{9}$$

In both cases we extract a contiguous depth of 112 slices:

$$[z_1, z_2) = \begin{cases} [Z-112,\ Z), & \text{top-anchored}, \\ [\max(0,\ z_{\text{peak}}-111),\ z_{\text{peak}}+1), & \text{peak-anchored}. \end{cases} \tag{10}$$

If fewer than 112 slices are available, all available slices are kept and padding is applied inferiorly. Fig. 2 illustrates peak-anchored scenario, which results in the cranial brain peak cut. For top-anchored scenario, the window is extracted from the 112 most superior slices, meaning the superior boundary remains $Z$, and thus top-anchored figure is absent.

Although the $z$-window selection is heuristic and could require dataset-specific tuning in principle, the cranial PET uptake profile in HECKTOR was remarkably stable across institutions and scanners.

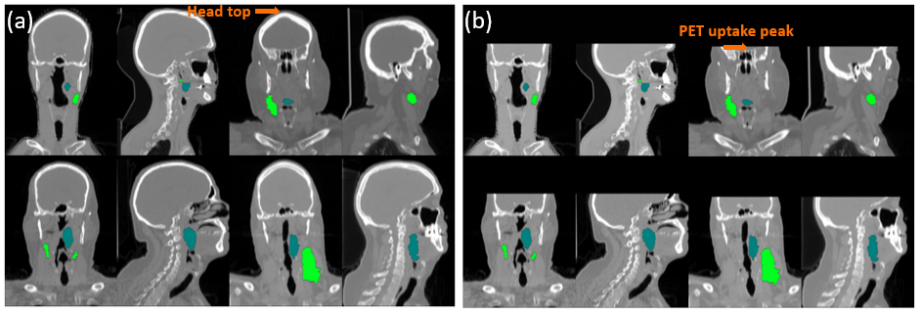


**Fig. 2.** Brain peak cut. (a) CT volumes before applying the cranial brain peak cut, showing strong cranial uptake; (b) cropped volumes after applying the proposed brain peak cut along the $z$-axis (peak-anchored).

**Final crop volume** Thus, the final PET/CT volume is cropped to a fixed field-of-view of

$$256 \times 256 \times 112,$$

centered in-plane at $(\bar{x}, \bar{y})$ and along $z$ according to Eq. 10. The fixed field-of-view of $256 \times 256 \times 112$ was chosen after confirming that, across all HECKTOR

training and validation cases from multiple centers, both GTVp and GTVn consistently fell within this depth range after PET-guided localization. This empirically verified coverage supports the robustness of the selected window size despite its heuristic construction.

**Intensity normalization and padding** We applied modality-specific normalization to account for the inherently different dynamic ranges of CT (Hounsfield units) and PET (SUV). Let $\sigma(\cdot)$ denote the logistic sigmoid.

For CT, intensities were clipped to the clinically relevant range of $[-1000, 1000]$ HU, linearly scaled to $[0, 1]$, and then passed through a sigmoid:

$$I_{\mathrm{CT}}^{\mathrm{norm}} = \sigma\left(\frac{\mathrm{clip}(I_{\mathrm{CT}}, -1000, 1000) - (-1000)}{1000 - (-1000)}\right).$$

For PET, voxel values were standardized by global z-score within the cropped volume and subsequently squashed:

$$I_{\mathrm{PET}}^{\mathrm{norm}} = \sigma\left(\frac{I_{\mathrm{PET}} - \mu}{\sigma}\right),$$

where $\mu$ and $\sigma$ denote the mean and standard deviation of the PET intensities.

To ensure that sigmoid squashing does not saturate lesion contrast at high SUVs, we empirically verified that all hot lesions in our dataset fall within the approximately linear region of the z-scored sigmoid input, resulting in no measurable loss of contrast.

The two normalized volumes were finally stacked as input channels, yielding a tensor of shape $(C, Z, Y, X)$ with $C = 2$.

**Provenance logging and restoration to native space** For each case we store a JSON sidecar with crop coordinates $(x_1, x_2, y_1, y_2, z_1, z_2)$, resampled geometry (size, spacing, origin, direction), and the original CT geometry. After inference, the predicted mask $\hat{M}$ on the cropped, resampled grid is pasted into a zero-initialized canvas of the resampled CT size at indices $[z_1{:}z_2,\ y_1{:}y_2,\ x_1{:}x_2]$, then resampled with nearest-neighbor interpolation onto the original CT grid to produce $\hat{M}_{\mathrm{orig}}$ with native spacing, orientation, and dimensions.

### 2.3 Network Architecture

The proposed network adopts a 3D U-Net backbone and is organized around two key modules: a *PET-guided spatial attention module* (PSAM), and *squeeze-and-excitation normalization* (SE Norm). Co-registered CT and PET images are concatenated along the channel axis and fed to the backbone, while the PET volume alone is processed by PSAM to produce a spatial attention map that modulates encoder features and guides decoding. SE Norm is used as the normalization within backbone blocks.

**PET-Guided Spatial Attention Module (PSAM)** Motivated by the higher lesion conspicuity on PET, PSAM takes the PET volume $\mathbf{P} \in R^{1\times D\times H\times W}$ and generates a single-channel attention map passed through a *sigmoid*. PSAM is a lightweight U-Net-style subnetwork producing a full-resolution map $\mathbf{A}$; multi-scale maps $\{\mathbf{A}^{(s)}\}$ for skip levels are obtained by average pooling with factors $\{1, 2, 4\}$. At each encoder stage $s$, the encoder feature $\mathbf{F}^{(s)}$ is *gated* by the corresponding PSAM map *before* it is concatenated into the decoder, i.e.,

$$\tilde{\mathbf{F}}^{(s)} = \mathbf{F}^{(s)} \odot \mathbf{A}^{(s)}, \tag{11}$$

where $\odot$ denotes element-wise multiplication. Thus PSAM influences the network *only* via the skip pathway, suppressing physiologic uptake outside lesions (e.g., salivary glands, brain, bladder) and highlighting metabolically active tumor regions that are fused with the anatomical context from CT.

**Squeeze-and-Excitation Normalization (SE Norm)** To counter residual false-positives from PET-only cues and to strengthen CT/PET fusion, we replace instance normalization (IN) with SE Norm in the convolutional blocks. Given an input feature $\mathbf{x}$, we first compute instance-normalized features $\mathbf{x}' = \mathrm{IN}(\mathbf{x})$. Global channel descriptors are then passed through a squeeze–excitation operator $f_{\mathrm{SE}}$ to produce per-channel affine parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta})$:

$$(\boldsymbol{\gamma}, \boldsymbol{\beta}) \;=\; f_{\mathrm{SE}}\big(\mathrm{GAP}(\mathbf{x}')\big), \qquad \mathbf{y}_c \;=\; \gamma_c\, \mathbf{x}'_c \;+\; \beta_c,\ \forall c. \tag{12}$$

This modulation performs context-aware reweighting of channels, enabling the network to *down-scale* channels dominated by physiological PET uptake while *up-scaling* channels aligned with tumor morphology on CT. Practically, SE Norm stabilizes optimization like IN yet introduces data-dependent channel calibration, yielding cleaner skip fusion and more reliable decoding.

**Loss Function** Given the substantial class imbalance inherent to PET/CT tumor delineation, we optimize the *soft Dice* objective. For prediction $\hat{\mathbf{y}}$ and reference $\mathbf{y}$,

$$\mathcal{L}_{\mathrm{Dice}}(\hat{\mathbf{y}}, \mathbf{y}) \;=\; 1 \;-\; \frac{2\sum_i \hat{y}_i\, y_i + \epsilon}{\sum_i \hat{y}_i^2 \;+\; \sum_i y_i^2 \;+\; \epsilon}, \tag{13}$$

where the sum runs over voxels (and classes for multi-class outputs) and $\epsilon$ is a small constant for numerical stability. The entire model is trained end-to-end using this objective without auxiliary constraints.

### 2.4 Training Details

All models are trained from scratch on the center-proportional split using the AdamW optimizer (learning rate $1\times10^{-3}$, weight decay $1\times10^{-5}$). A cosine–annealing schedule with warm restarts ($T_0$=25, $\eta_{\min}$=$1 \times 10^{-5}$) is used, and training runs for 200 epochs with a batch size of 1. The soft Dice loss (Section 2.3) is optimized,

and the checkpoint with the highest mean validation Dice over GTVp/GTVn is selected.

To ensure fair comparison, all baselines (U-Net, UNETR, U-Mamba) are trained with the same preprocessing, data split, optimizer, scheduler, and loss. No additional geometric augmentations are applied.

## 3 Results

We evaluate on the center-proportional validation split of HECKTOR 2025 (600 train / 80 validation cases). We report the Dice coefficient for the primary tumor (GTVp) and nodal disease (GTVn), together with their average. All metrics are computed per case and then averaged over the set. We compare against U-Net, UNETR [9] and U-Mamba [10].

For a fair comparison, all baseline models were trained using exactly the same preprocessed PET/CT volumes. As summarized in Table 3, our method attains a Dice of **0.7133** on GTVp and **0.5013** on GTVn, yielding a mean of **0.6073**. This improves the mean Dice over U-Net (0.5949), UNETR (0.5802) and U-Mamba (0.5932). The higher GTVp score suggests that PET-guided skip gating helps sharpen boundaries in metabolically active primaries. For GTVn, although our method achieves the best score among compared approaches, the absolute Dice remains lower than for GTVp, reflecting the inherent difficulty of segmenting small, low-contrast nodes.

**Table 3.** Validation Dice scores on the center-proportional split (80 cases)

| Method | GTVp | GTVn | Mean (GTVp/GTVn) |
|---|---|---|---|
| **Ours** | **0.7133** | **0.5013** | **0.6073** |
| UNETR | 0.6940 | 0.4664 | 0.5802 |
| U-Net | 0.7070 | 0.4828 | 0.5949 |
| U-Mamba | 0.7053 | 0.4811 | 0.5932 |

*Test-set Results.* On the official HECKTOR 2025 test set, our method achieved a GTVp Dice of **0.3803**, a GTVn aggregate Dice of **0.3733**, and an aggregate GTVn F1-score of **0.1982**. These scores were obtained through the challenge evaluation server after submitting our final model checkpoint.

The performance drop compared to the validation split is expected in a multi-center test environment, where differences in scanner types, reconstruction protocols, and annotation styles typically reduce absolute Dice scores. This behavior has been consistently observed in previous HECKTOR challenges. Despite this variance, our model maintained stable segmentation performance across centers without requiring external domain-specific tuning, supporting the practical robustness of the proposed PET-guided preprocessing and lightweight fusion design in real deployment scenarios.

## 4 Conclusion

We presented a PET-guided 3D U-Net for H&N tumor and lymph-node segmentation that (i) gates encoder skip features via PSAM, (ii) employs SE-Norm for stronger channel-wise fusion of CT and PET, and (iii) standardizes inputs with a PET-guided brain-peak-cut pipeline. On the HECKTOR 2025 validation split, the method achieves a mean Dice of 0.6073 without external data or test-time augmentation. Further studies are necessary to strengthen generalization capability across diverse center datasets. Future work will explore tighter PET-to-CT coupling (e.g., cross-modal attention in the encoder), uncertainty-aware post-processing, and automated hyperparameter selection for center-aware generalization.

## References

1. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: MICCAI (LNCS 9351), pp. 234–241. Springer (2015). https://doi.org/10.1007/978-3-319-24574-4_28
2. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In: MICCAI (LNCS 9901), pp. 424–432. Springer (2016). https://doi.org/10.1007/978-3-319-46723-8_49
3. Oreiller, V., Andrearczyk, V., Jreige, M., *et al.*: Head and neck tumor segmentation in PET/CT: The HECKTOR challenge. Med. Image Anal. **77**, 102336 (2022). https://doi.org/10.1016/j.media.2021.102336
4. Andrearczyk, V., Oreiller, V., Hatt, M., Depeursinge, A., *et al.*: Overview of the HECKTOR Challenge at MICCAI 2022: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT. In: LNCS, pp. 1–30 (2023). https://doi.org/10.1007/978-3-031-27420-6_1
5. Oktay, O., Schlemper, J., Le Folgoc, L., *et al.*: Attention U-Net: Learning Where to Look for the Pancreas. arXiv:1804.03999 (2018).
6. Hu, J., Shen, L., Sun, G.: Squeeze-and-Excitation Networks. In: CVPR, pp. 7132–7141 (2018). https://doi.org/10.1109/CVPR.2018.00745
7. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance Normalization: The Missing Ingredient for Fast Stylization. arXiv:1607.08022 (2016).
8. Saeed, N., Hassan, S., Hardan, S., Aly, A., Taratynova, D., Nawaz, U., *et al.*: A Multimodal and Multi-centric Head and Neck Cancer Dataset for Segmentation, Diagnosis, and Outcome Prediction. arXiv:2509.00367 (2025).
9. Hatamizadeh, A., Tang, Y., Nath, V., *et al.*: UNETR: Transformers for 3D Medical Image Segmentation. In: WACV, pp. 1748–1758 (2022). https://doi.org/10.1109/WACV51458.2022.00181
10. Ma, J., Li, Z., Xie, Y., *et al.*: U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation. arXiv:2401.04722 (2024).
11. Cho, M.J., Hwang, D., Yie, S.Y., Lee, J.S.: Multi-modal co-learning with attention mechanism for head and neck tumor segmentation on $^{18}$F-FDG PET-CT. EJNMMI Phys. **11**, 67 (2024). https://doi.org/10.1186/s40658-024-00670-y