Modernizing nnU-Net with Convolutional SwiGLU for Segmentation, Staging, and Prognosis in HECKTOR 2026

Ting-Yi Chang[0009−0008−7396−6750]

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan 70101, Taiwan
p76144621@gs.ncku.edu.tw

Abstract. Head and neck cancer management relies on fluorodeoxyglucose positron emission tomography / computed tomography (FDG-PET/CT) to delineate the primary and nodal gross tumor volumes (GTVp/GTVn), assign the tumor and nodal (TN) stage, and estimate prognosis; HECKTOR 2026 casts these steps as one pipeline of segmentation, T/N staging, and recurrence-free survival (RFS) prediction. We present a three-stage, leakage-aware cascade whose downstream models consume only fold-aligned out-of-fold (OOF) upstream predictions. Its core, the **Efficient SwiGLU Residual Encoder U-Net**, modernizes nnU-Net v2 ResEnc-M by re-examining each borrowed component under controlled ablations: mean Dice improves from 0.6759 to 0.6948 while total parameters fall by 19% and decoder parameters by 94.5%. The submitted model attains a pooled OOF mean Dice of 0.6957, significantly above a final-setting ResEnc-M baseline; T/N balanced accuracies are 0.5662 and 0.5663, and the RFS ensemble reaches a Harrell C-index of 0.6762, versus 0.5982 for a clinical-only baseline. On the official test set, the weighted score is 0.5776 (Dice 0.6186; T/N 0.4107/0.5708; C-index 0.6280). Our results favor task-specific capacity reallocation over direct transfer of general-purpose design choices. **Team: NCKU Tingyi Solo.**

Keywords: Head and neck cancer · PET/CT · Medical image segmentation · SwiGLU · Tumor staging · Survival analysis

1 Introduction

Head and neck cancer is among the most common cancers worldwide. Radiotherapy planning requires delineating the primary gross tumor volume (GTVp) and all nodal gross tumor volumes (GTVn); manual contouring is time-consuming and observer-dependent. Combined fluorodeoxyglucose positron emission tomography / computed tomography (FDG-PET/CT) supports this task with complementary metabolic and anatomical information and further informs TN staging and recurrence-risk estimation, which directly influence treatment selection. The HECKTOR 2026 challenge casts these steps as a single clinical sequence: from PET/CT and clinical variables, participants must segment GTVp/GTVn, assign T and N categories, and predict recurrence-free survival (RFS) [13, 14].

The multi-center setting introduces acquisition variability, heterogeneous lesion sizes, and class imbalance, and all test centers are unseen during training [13]. Because the tasks form a cascade, upstream errors alter downstream features, while training downstream models on ground-truth masks would create an unrealistic shift at deployment; a valid system must address leakage and error propagation as well as task accuracy.

Although Transformers dominate elsewhere, nnU-Net [11, 5] and its residual-encoder revision [6]—architecturally a classical residual U-Net—remain the reference baselines and frequent top performers in 3D medical segmentation challenges, since the weaker inductive biases of Vision Transformers demand far larger training sets than multi-center medical cohorts provide [2]. The question is thus not whether to replace the convolutional backbone, but how to modernize it correctly.

MedNeXt [12] modernizes nnU-Net by porting ConvNeXt components [8], yet several premises behind that port do not necessarily hold in 3D medical imaging. In ConvNeXt, depthwise convolution *trades* accuracy for efficiency that is reinvested into larger width and kernels; if that reinvestment does not pay off, only the accuracy loss remains—and depthwise convolutions realize poor hardware efficiency in 3D [16]. Replacing ReLU with GELU is typically motivated by gains that are often at noise level, and in our measurements GELU is never faster than the simpler SiLU (Sect. 4.2). And MedNeXt's decoder mirrors its encoder, although EffiDec3D demonstrates that 3D decoders are heavily redundant [10]. More generally, many Transformer modifications fail to transfer across implementations [9]. We therefore re-derive the "modern recipe" premise by premise in controlled ablations.

From the large-language-model stack we retain one component: SwiGLU, which multiplies a SiLU-gated branch by a linear value branch [15, 17]. Implemented with dense convolutions, it can be read as a Fused-MBConv-style block [16] without squeeze-and-excitation but with multiplicative gating—a hardware-friendly formulation whose suitability for local 3D features we likewise validate by ablation.

Our main contributions are as follows:

1. A fixed physical region of interest (ROI) and the Efficient SwiGLU Residual Encoder U-Net: convolutional SwiGLU, SiLU, additive skips, a five-stage (16×) encoder, a 64-channel decoder, and AdamW, each validated component-wise, with a significant paired improvement over a final-setting ResEnc-M baseline.
2. A single-pass, 2.5-mm T/N classifier that jointly analyzes CT, PET, and predicted GTVp/GTVn masks with a shared pretrained SwiGLU encoder.
3. An RFS model combining clinical, mask-derived volumetric/metabolic, nodal-distribution, and probabilistic staging features (+0.078 C-index over a clinical-only baseline), with bootstrap uncertainty estimates throughout and a domain-shift analysis of the official test scores.

2 Methods

2.1 System Overview and Fixed Physical ROI

The cascade predicts GTVp/GTVn from CT and co-registered PET, estimates T1–T4 and N0–N3—per the American Joint Committee on Cancer (AJCC)/UICC 7th edition—from both images and predicted masks, and fuses upstream outputs with clinical data for RFS prediction. Validation uses fold-aligned out-of-fold (OOF) predictions throughout—each case is processed only by models that never saw it during training—while at test time the five fold-specific models are ensembled, without any test-time augmentation (TTA).

Training lesions span -120.6 to 126.5 mm left–right, -66.9 to 154.8 mm anterior–posterior, and at most 330.3 mm below the cranial end. PET is linearly resampled to the CT grid when their geometry differs, and both modalities are reoriented to a common right–anterior–superior (RAS) orientation. We then crop 160 mm about each in-plane CT center and 360 mm inferiorly from the cranial end, defining a nominal $320 \times 320 \times 360$ mm ROI. The crop is intersected with the acquired field of view, without padding or changing native spacing.

2.2 Efficient SwiGLU Residual Encoder U-Net

The baseline is nnU-Net v2 ResEnc-M [6]. Following the premise analysis above, we retain dense $3 \times 3 \times 3$ convolutions and InstanceNorm (IN), omitting depthwise large kernels, squeeze-and-excitation, and LayerScale; GroupNorm with 32 groups [18] is likewise rejected, as it reduces GTVn Dice in our ablation.

For input features x, the convolutional SwiGLU residual block is defined as

$$g = \mathrm{IN}(W_g * x), \quad v = \mathrm{IN}(W_v * x), \quad y = S(x) + \mathrm{IN}(W_o * [\mathrm{SiLU}(g) \odot v]), \quad (1)$$

where $*$ denotes convolution; W_g and W_v are dense $3 \times 3 \times 3$ convolutions, and W_o is a $1 \times 1 \times 1$ output projection. The output of W_o is normalized by InstanceNorm before being added to the identity or channel-projection shortcut $S(\cdot)$. This design translates SwiGLU gating from the Transformer feed-forward network [15] into local three-dimensional operations. Given stage input width C and expansion ratio $r = 2$, the hidden width is $h = 64\lceil(2/3)rC/64\rceil$. The five stages use widths $[32, 64, 128, 256, 320]$, repeats $[1, 3, 4, 6, 6]$, and hidden widths $[64, 128, 192, 384, 448]$. Following parameter-matched SwiGLU practice [15, 17], the 2/3 factor compensates for the two gated branches, and widths are rounded to multiples of 64.

The original ResEnc-M encoder reaches $32\times$ downsampling over six scales; following the analysis in Sect. 4.3, the final encoder removes the sixth scale and retains $16\times$ downsampling.

2.3 Lightweight Additive Decoder

Motivated by EffiDec3D [10]—and by the observation that the original nnU-Net already favored light decoders—decoder widths are $[64, 64, 64, 32]$ with one block

per scale. A $1 \times 1 \times 1$ projection aligns each encoder feature before addition:

$$d_i = \mathrm{Up}(d_{i+1}) + \mathrm{Conv}^{1\times1\times1}(e_i). \tag{2}$$

Addition avoids expanding the following convolution from C to $2C$, reducing parameters, multiply-adds, and fusion memory. We replace MedNeXt's GELU [12] with SiLU—the exact gate function of SwiGLU, so one nonlinear primitive serves the whole network—while additive skips measurably reduce runtime and memory (Sect. 4.2). Deep supervision is retained at every scale, as in nnU-Net and MedNeXt [5, 12]. Figure 1 summarizes the architecture and blocks.

2.4 Segmentation Training and Inference

The final model uses a target spacing of $[1.5, 0.9766, 0.9766]$ mm in (z, y, x) order, a $[96, 160, 160]$ patch, and batch size 2. CT follows nnU-Net normalization and PET uses per-case Z-scoring. Dice plus cross-entropy loss and nnU-Net v2 augmentation are used for 1,000 epochs [5, 6].

Optimization uses AdamW with learning rate 3×10^{-4}, $\beta = (0.9, 0.98)$, and weight decay 0.05 on 3D/transposed-convolution weights. Fifty warm-up epochs precede polynomial decay (power 0.9); gradients are clipped at 1.0.

At test time, the per-case inference time budget of the challenge cannot accommodate both fold ensembling and TTA, so we allocate it entirely to the five-fold ensemble: fold softmax outputs are averaged under nnU-Net-style sliding-window inference (0.5 overlap, Gaussian blending) [5] with TTA disabled (no mirroring), and predictions are restored to the CT grid by nearest-neighbor interpolation.

2.5 Segmentation-Guided T/N Staging

The T/N model receives CT, PET, and GTVp/GTVn OOF masks; using predicted masks for both training and validation prevents ground-truth leakage. At 2.5-mm isotropic spacing, one $[144, 128, 128]$ patch covers the complete ROI (batch size 4).

The classifier removes the decoder and initializes its five-stage encoder from the corresponding trained segmentation fold; the first convolution is expanded from two to four channels, preserving the learned CT/PET weights and zero-initializing the mask channels, while task heads are randomly initialized.

Global average pooling yields $p \in \mathbb{R}^{320}$. Separate T and N SwiGLU heads compute, for $j \in \{T, N\}$,

$$f_j = W_{o,j}[\mathrm{SiLU}(\mathrm{LN}(W_{g,j}p)) \odot \mathrm{LN}(W_{v,j}p)]\,, \tag{3}$$

where LayerNorm is denoted by LN. $W_{g,j}$ and $W_{v,j}$ map 320 to 448 dimensions; $W_{o,j}$ returns the gated product to 320 before a four-logit classifier. T0, Tx, Nx, and missing labels are excluded from the corresponding task. We minimize $\mathcal{L}_T + \mathcal{L}_N$ using fold-wise inverse-frequency-weighted cross-entropy and 0.1 label smoothing. AdamW training lasts 250 epochs with 25 warm-up epochs; five-model softmax probabilities are averaged at inference.

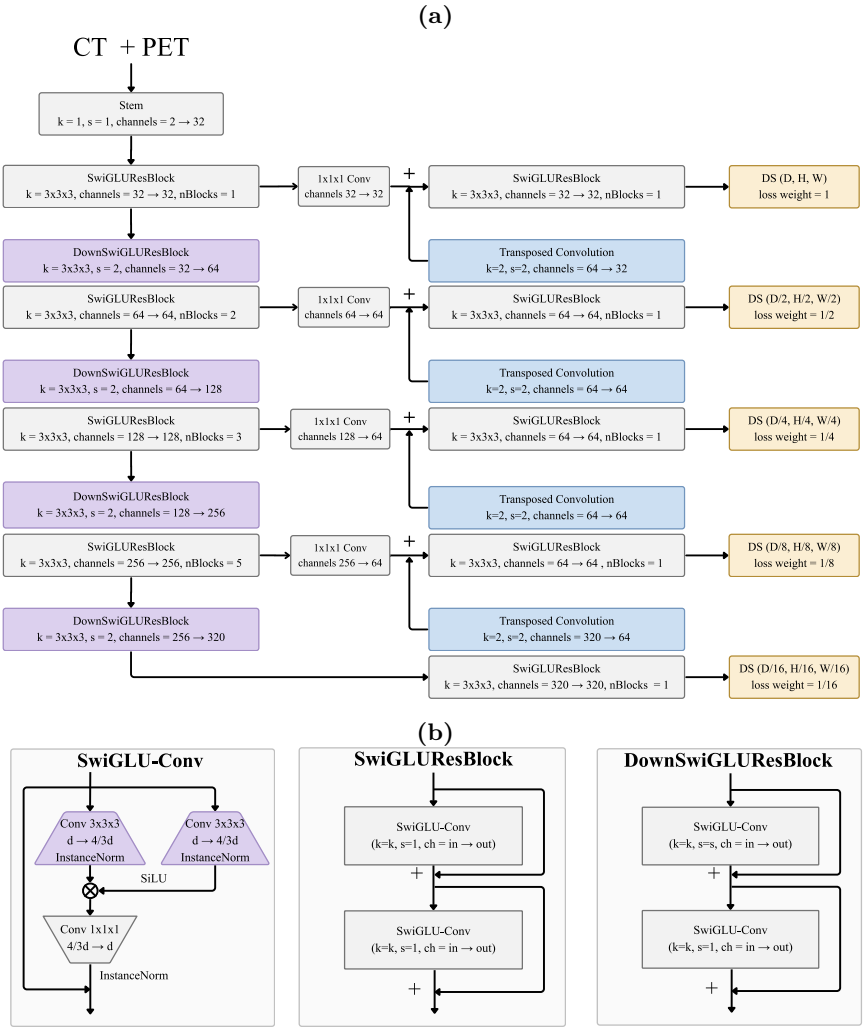

Fig. 1. Efficient SwiGLU Residual Encoder U-Net. (a) Five-stage encoder, lightweight additive decoder capped at 64 channels, and multi-scale supervision; DS denotes deep supervision. (b) SwiGLU-Conv, the resolution-preserving SwiGLUResBlock, and the downsampling DownSwiGLUResBlock.

2.6 Multi-Source Feature Fusion for RFS Prognosis

The RFS stage uses interpretable clinical, imaging, and staging features. Clinical variables are age, gender, human papillomavirus (HPV) status, tobacco and alcohol consumption, and performance status. For each of GTVp, GTVn, and their union, we compute five variables: volume, the maximum and mean standardized uptake values (SUVmax, SUVmean), metabolic tumor volume (MTV), and total lesion glycolysis (TLG).

PET feature nomenclature follows established procedural guidance [1]; MTV contains voxels at or above 40% of regional SUVmax, and TLG is their SUV sum times voxel volume. Two nodal features—the number of 26-connected GTVn components larger than 0.05 cm^3 and bilateral involvement across $x < -10$ and $x > 10$ mm—give 17 imaging features in total.

To retain staging uncertainty, complete T/N softmax distributions are summarized as expected stages:

$$T_{\text{exp}} = \sum_{i=0}^{3} (i+1)p(T_{i+1}), \qquad N_{\text{exp}} = \sum_{i=0}^{3} i\, p(N_i). \tag{4}$$

Unlike an argmax, these expected stages retain uncertainty between adjacent categories ($p(\cdot)$ denotes stage probability).

All upstream RFS features are fold-aligned OOF predictions. Volume, MTV, and TLG use $\log(1+x)$; continuous variables are fold-wise median-imputed and standardized, while categorical variables receive an Unknown category and one-hot encoding. We fit a random survival forest (RSF) [7] (300 trees, leaf size 20, square-root feature sampling) and a gradient-boosted Cox model (GBM) [3] (300 estimators, learning rate 0.05, depth 2, subsample 0.8) with scikit-survival. Within each validation fold, each model's risk scores are standardized using the mean and standard deviation of its predictions on the corresponding training fold, after which the two Z-scores are averaged equally. For deployment, the same statistics are recomputed after refitting on all training cases; the output direction follows the challenge specification.

The equal-weight ensemble was fixed *before* inspecting pooled OOF scores: picking the best single learner on the same OOF data would constitute selection on the validation signal, and the ensemble–best-member difference is far smaller than its bootstrap uncertainty (Sect. 4.1); two learners with different inductive biases also hedge against center shift.

3 Experimental Setup

3.1 Data, Splits, and Evaluation Metrics

We use 782 co-registered annotated PET/CT cases and the fixed five-fold split (157/157/156/156/156) [13]. Valid T, N, and RFS targets are available for 775, 782, and 727 cases, respectively; the latter include 148 events.

Metrics are GTVp/GTVn Dice and their mean, four-class balanced accuracy (BAcc) for each staging head, and Harrell's concordance index (C-index) [4] for RFS, pooled over five-fold OOF predictions. Class-specific Dice is computed on the original CT grid; cases in which both ground truth and prediction are empty for a class are excluded, while false positives on class-empty cases count as zero. Every pooled metric carries a case-level nonparametric bootstrap 95% confidence interval (CI; 2,000 resamples for segmentation and staging, 1,000 for RFS); paired differences resample the common case set.

Table 1. Pooled five-fold OOF validation with case-level bootstrap 95% CIs, and official test results (unseen centers).

Metric	OOF	95% CI	Test
GTVp Dice	0.7081	0.690–0.726	–
GTVn Dice	0.6833	0.664–0.701	–
Mean Dice	0.6957	0.682–0.708	0.6186
T balanced accuracy	0.5662	0.531–0.602	0.4107
N balanced accuracy	0.5663	0.523–0.611	0.5708
RFS C-index (Harrell)	0.6762	0.631–0.721	0.6280
Weighted challenge score	–	–	0.5776

3.2 Ablation Design

Architectural ablations use a $[96, 160, 160]$ patch and 0.97-mm through-plane spacing. Each step tests one premise of the transfer analysis in the Introduction: GroupNorm (32 groups) tests MedNeXt's normalization choice; SiLU tests the GELU cost–benefit claim; additive skips and the 64-channel decoder cap (Cap64) test decoder redundancy; removing the deepest stage (No5) tests the value of $32\times$ downsampling; AdamW isolates the optimizer (earlier rows use nnU-Net's default stochastic gradient descent, SGD); twofold expansion isolates capacity; and SwiGLU tests gating. Because the design is cumulative, each contribution is conditional on the components before it (order fixed a priori); full independent attribution would require a factorial study. A separate comparison of one non-final architecture at 0.97- versus 1.5-mm through-plane spacing evaluates cranio-caudal coverage; its values are not directly pooled with the architectural ablation.

4 Results

4.1 Main Validation and Official Test Results

Table 1 summarizes the submitted system across all three tasks. The final segmentation model attains a pooled OOF mean Dice of 0.6957. Under the identical ROI, spacing, patch, and evaluation protocol, a final-setting nnU-Net ResEnc-M baseline (default SGD recipe, 1,000 epochs) reaches 0.6847 (GTVp 0.7082, GTVn 0.6612); the case-paired difference of +0.0111 (95% CI 0.0036–0.0189) excludes zero, with the gain concentrated in GTVn (+0.022).

A majority-class staging predictor attains 0.25 balanced accuracy by construction, so both heads are clearly above chance. For prognosis, the same pipeline restricted to clinical variables yields a C-index of 0.5982 (95% CI 0.551–0.644); the mask-derived imaging and probabilistic staging features therefore contribute +0.078. Prior HECKTOR editions evaluated segmentation with an aggregated Dice on different cohorts [14], so their values are not directly comparable.

Per-task detail follows in Tables 2 and 3. N-stage raw accuracy is 0.6803, but its balanced accuracy reflects the N2 majority (471 of 782 cases). Among the

Table 2. Pooled OOF validation results for T/N staging (BAcc: balanced accuracy).

Head	Cases	Accuracy	BAcc	Per-class recall
T	775	0.5819	0.5662	T1 0.6772; T2 0.6316; T3 0.4645; T4 0.4915
N	782	0.6803	0.5663	N0 0.5287; N1 0.3816; N2 0.7941; N3 0.5608

Table 3. Pooled OOF validation of the RFS models.

Model	Harrell C-index	95% CI
Clinical-only ensemble	0.5982	0.551–0.644
Random survival forest	0.6635	0.618–0.708
Gradient-boosted Cox	**0.6779**	0.632–0.721
Z-score average ensemble	0.6762	0.631–0.721

RFS learners, GBM attains the best Harrell C-index (0.6779), but its paired advantage over the pre-specified Z-score ensemble is +0.0016 (95% CI −0.008 to +0.011): statistically indistinguishable, supporting the pre-specified choice (Sect. 2.6).

4.2 Segmentation Architecture Ablation

Table 4 shows the five-fold cumulative ablation, each row testing one transfer premise (Sect. 3.2). GroupNorm mainly degrades GTVn, so MedNeXt's normalization choice does not transfer here. SiLU and additive skips raise mean Dice from 0.67589 to 0.67843. Cap64 preserves performance despite cutting the decoder to 0.72M parameters, confirming decoder redundancy; No5 adds 0.01540 GTVn Dice. The final SwiGLU model gains 0.01891 mean Dice over baseline, primarily through GTVn (+0.03748), while cutting decoder parameters by 94.5%.

Table 4. Cumulative segmentation ablation (controlled 0.97-mm setting). Add: additive skips; Cap64: 64-channel decoder cap; No5: no sixth encoder scale (32×); Exp2: expansion ratio 2 (also in SwiGLU). Best value per column in bold; runtime and memory in Table 5.

Configuration	GTVp	GTVn	Mean	Δ Mean
ResEnc-M baseline	0.70795	0.64384	0.67589	–
+GroupNorm (32 groups)	0.70820	0.63799	0.67310	-0.00279
+SiLU	0.70983	0.64489	0.67736	+0.00147
+SiLU+Add	0.71047	0.64639	0.67843	+0.00254
+SiLU+Add+Cap64	**0.71389**	0.64326	0.67858	+0.00269
+SiLU+Add+Cap64+No5	0.71189	0.65866	0.68528	+0.00939
+SiLU+Add+Cap64+No5+AdamW	0.70849	0.67401	0.69125	+0.01536
+SiLU+Add+Cap64+No5+AdamW+Exp2	0.70861	0.67693	0.69277	+0.01688
+SwiGLU+Add+Cap64+No5+AdamW	0.70828	**0.68132**	**0.69480**	**+0.01891**

AdamW alone raises mean Dice by +0.00597, explaining 79.7% of the combined AdamW and Exp2 gain over No5; expansion adds 0.00152 and SwiGLU 0.00203.

Table 5. Runtime and memory of the configurations of Table 4, plus a GELU counterfactual of the baseline architecture (not part of the cumulative chain). Single RTX 5090, fp16, torch.compile on, 200 timed iterations after 50 warm-up; inference batch 1 on a $[96, 160, 160]$ patch; training batch 2 with the full loss and each recipe's optimizer, 250 steps per epoch, network compute only.

		Inference		Training	
Configuration	Params (M)	ms	GB	s/epoch	GB
ResEnc-M baseline	101.94	17.5	1.47	28.4	7.0
+GroupNorm (32 groups)	101.94	18.4	1.47	30.2	7.9
+GELU (counterfactual)	101.94	17.5	1.47	29.4	8.4
+SiLU	101.94	17.5	1.47	29.2	8.4
+SiLU+Add	96.83	15.7	1.11	26.5	7.4
+SiLU+Add+Cap64	91.02	15.5	1.09	26.6	7.3
+SiLU+Add+Cap64+No5	57.75	14.9	0.97	25.7	7.0
+SiLU+Add+Cap64+No5+AdamW	57.75	14.9	0.97	25.4	7.2
+SiLU+Add+Cap64+No5+AdamW+Exp2	58.21	14.4	1.12	25.9	7.0
+SwiGLU+Add+Cap64+No5+AdamW	82.34	21.7	1.50	39.0	8.4

Comparison only with the SGD baseline would therefore overstate the effect of expansion.

Table 5 tells the efficiency story of the recipe. Every step through Exp2 makes the network simultaneously lighter, faster, and more accurate than the baseline (-43% parameters, -18% inference latency, -9% epoch time, $+0.0169$ mean Dice): additive skips eliminate concatenation traffic, Cap64 strips the decoder from 11.64M to 0.72M parameters—removing capacity that invites overfitting in a 782-case cohort, at no cost in mean Dice—and No5 discards a redundant high-capacity stage. The counterfactual rows then validate our component choices. The GELU counterfactual offers no measured advantage in any column, whereas SiLU matches it at inference, trains marginally faster (29.2 vs. 29.4 s/epoch), and is the exact gate function of SwiGLU—a single, slightly faster nonlinear primitive serving the entire network. GroupNorm, in contrast, is the slowest configuration in both phases on top of the worst GTVn Dice, confirming its rejection. SwiGLU finally reinvests the banked savings where they matter clinically: $+24\%$ inference latency and $+37\%$ epoch time complete a cumulative $+0.037$ GTVn gain over the baseline (Table 4), and the 82M-parameter model still infers within 1.5 GB. One caveat: smooth activations store their inputs for the backward pass, adding $\sim$1.4 GB of training memory over LeakyReLU.

4.3 Assessing the 32× Downsampling Stage

ResEnc-M contains a sixth encoder scale with a maximum downsampling factor of 32; for the $[96, 160, 160]$ training patch, its deepest feature map spans only $[3, 5, 5]$ voxels. Such a coarse representation enlarges the receptive field but may discard spatial detail important for small nodes, so we asked whether it is complementary or redundant.

We first performed a diagnostic inference-time ablation on 32 cases without retraining, zeroing one encoder output at a time: for Stages 0–4 the skip feature

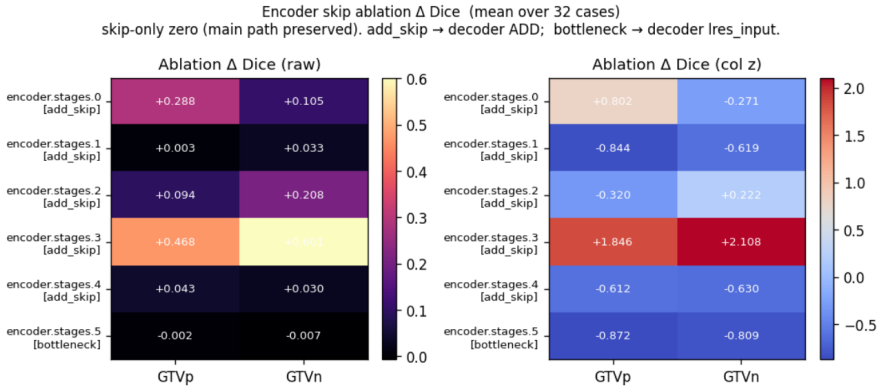

Fig. 2. Encoder-output ablation over 32 cases. The left panel shows raw Dice changes, and the right panel shows column-wise standardized values. Stage 5 is the bottleneck; its negative values indicate that Dice increases when its decoder input is suppressed.

supplied to the decoder, and for Stage 5 the lowest-resolution bottleneck input. Figure 2 defines $\Delta\text{Dice} = \text{Dice}_{baseline} - \text{Dice}_{ablated}$, so a positive value indicates that removing a feature is harmful. The large changes at Stage 3 (+0.468 for GTVp and +0.601 for GTVn) confirm that the perturbation detects features the decoder depends on; in contrast, Stage 5 yields −0.002 and −0.007: suppressing the bottleneck slightly *improves* Dice.

The inference-time result alone does not prove the deepest stage useless: it covers 32 cases, applies a post-hoc perturbation, and cannot distinguish an uninformative representation from a merely redundant one. We therefore removed the complete 32× stage and retrained all five folds: No5 trades 0.0020 GTVp Dice for +0.0154 GTVn (mean 0.6786 to 0.6853) while removing 33M parameters (Table 4). Both experiments motivate the final 16× encoder.

Through-plane spacing. Increasing the through-plane spacing from 0.97 to 1.5 mm expands 96-slice coverage from 93.1 to 144 mm. In a paired comparison of one non-final architecture, GTVp is unchanged, while GTVn rises from 0.64642 to 0.66088 and mean Dice from 0.67729 to 0.68445. The submitted model therefore uses 1.5-mm spacing; because the architectures differ, this supports the final choice but cannot be cross-attributed to Table 4.

5 Discussion

Most validation gains arise from GTVn lesions, which are smaller, vary in number, and are distributed across cervical levels [13, 14]; the improvements from No5, 1.5-mm spacing, and SwiGLU are consistent with greater reliance on spatial coverage and multi-scale features. These findings are implementation-specific, and

the cumulative design makes each contribution conditional on its predecessors (Sect. 3.2).

The runtime study adds a broader lesson: theoretical proxies predicted wall-clock behavior poorly in both directions—parameter count did not track latency (SwiGLU), and FLOP savings did not materialize as speed (Cap64). Measured premises, rather than imported intuitions, should drive component transfer between domains.

The official test results quantify the cost of unseen centers: N-stage balanced accuracy is essentially preserved (0.5663 to 0.5708), whereas T-stage drops most (0.5662 to 0.4107). In the AJCC system, N categorization is largely determined by nodal size, number, and laterality—geometric properties that transfer across acquisition protocols—while T categorization depends on finer, center-sensitive appearance cues such as invasion extent. Mean Dice falls to 0.6186 and the C-index to 0.6280, consistent with burden features that generalize only partially. Because all validation used fold-aligned OOF predictions, this gap reflects domain shift rather than optimistic leakage.

Using predicted masks matches deployment but propagates segmentation errors; low N1 recall may reflect rarity and subtle N1/N2 distinctions. The ensemble–GBM difference lies well within its bootstrap CI; we retain the pre-specified ensemble to avoid post-hoc selection, and only 148 events limit finer comparison.

Further limitations are the training-derived ROI boundaries and the nonidentical architectures in the spacing comparison. The runtime benchmarks cover a single GPU type, and the challenge timeline did not permit retraining external architectures such as MedNeXt or Transformer-based models under our protocol, so comparisons beyond the ResEnc-M family rest on premise-level analysis. External validation, calibration, lesion-level uncertainty, and joint training remain future work, and hand-crafted RFS features omit intratumoral heterogeneity.

6 Conclusion

We present a leakage-aware PET/CT cascade for HECKTOR 2026. Re-deriving the "modern convolutional recipe" premise by premise yields the Efficient SwiGLU Residual Encoder U-Net, which significantly outperforms a final-setting ResEnc-M baseline while cutting total parameters by 19% and decoder parameters by 94.5%, reinvesting a measured 24% of inference latency into the GTVn gains that dominate the improvement. OOF masks and staging probabilities connect the downstream tasks, adding +0.078 C-index over a clinical-only baseline, and the official test results document the remaining cost of center shift. Overall, the results favor task-specific capacity reallocation over wholesale transfer of general-purpose components.

Disclosure of Interests. The author declares no competing interests.

References

1. Boellaard, R., Delgado-Bolton, R., Oyen, W.J.G., et al.: FDG PET/CT: EANM procedure guidelines for tumour imaging: Version 2.0. Eur. J. Nucl. Med. Mol. Imaging **42**(2), 328–354 (2015)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
3. Friedman, J.H.: Greedy function approximation: A gradient boosting machine. Ann. Stat. **29**(5), 1189–1232 (2001)
4. Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A.: Evaluating the yield of medical tests. JAMA **247**(18), 2543–2546 (1982)
5. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**, 203–211 (2021)
6. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K.H., Jaeger, P.F.: nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. arXiv preprint arXiv:2404.09556 (2024)
7. Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S.: Random survival forests. Ann. Appl. Stat. **2**(3), 841–860 (2008)
8. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: CVPR. pp. 11976–11986 (2022)
9. Narang, S., Chung, H.W., Tay, Y., et al.: Do transformer modifications transfer across implementations and applications? In: EMNLP. pp. 5758–5773 (2021)
10. Rahman, M.M., Marculescu, R.: EffiDec3D: An optimized decoder for high-performance and efficient 3D medical image segmentation. In: CVPR. pp. 10435–10444 (2025)
11. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI 2015. LNCS, vol. 9351, pp. 234–241 (2015)
12. Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jäger, P.F., Maier-Hein, K.H.: MedNeXt: Transformer-Driven scaling of ConvNets for medical image segmentation. In: MICCAI 2023. LNCS, vol. 14223, pp. 405–415 (2023)
13. Saeed, N., Hassan, S., Hardan, S., et al.: A multimodal and multi-centric head and neck cancer dataset for segmentation, diagnosis and outcome prediction. arXiv preprint arXiv:2509.00367 (2025)
14. Saeed, N., Hassan, S., Hardan, S., et al.: HEad and neCK TumOR (HECKTOR) 2025: Benchmark of segmentation, diagnosis, and prognosis in multimodal PET/CT. arXiv preprint arXiv:2606.20143 (2026)
15. Shazeer, N.: GLU variants improve transformer. arXiv preprint arXiv:2002.05202 (2020)
16. Tan, M., Le, Q.V.: EfficientNetV2: Smaller models and faster training. In: ICML. PMLR, vol. 139, pp. 10096–10106 (2021)
17. Touvron, H., Lavril, T., Izacard, G., et al.: LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
18. Wu, Y., He, K.: Group normalization. In: ECCV 2018. LNCS, vol. 11217, pp. 3–19 (2018)