iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82917

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

VA-HSNet: A Vertebra-Aware Hybrid CNN–Transformer Network for CT Spine Segmentation - Design and Evaluation

Akshit Sahu, Anshuman Singh, Ayush Singh
Dept. of ECE Delhi Technological University New Delhi, India

***Abstract*—As the demand for automated quantitative analysis of medical imaging continues to increase in domains such as spinal-surgery planning, fracture screening, and osteoporosis risk assessment, there exists a need for vertebra segmentation methods that provide higher delineation accuracy, identification reliability, and cross-vendor generalisation. In this paper, we present the implementation and experimental validation of VA-HSNet, a Vertebra-Aware Hybrid Segmentation Network built around a**

PyTorch-based research code-base that integrates a residual 3D Convolutional Neural Network (CNN) encoder, a vertebraaware transformer bottleneck, and a deeply supervised decoder for CT vertebra segmentation. Our methodology included a

deterministic Hounsfield-Unit (HU) preprocessing pipeline, a twostagelocalisation–segmentation architecture, a vertebra-index positional bias along the cranio-caudal axis, and a composite training objective combining Dice, focal cross-entropy, distanceweighted boundary, and a novel centroid-ordering anatomicalconsistency loss. We use the VerSe 2020 public benchmark (319 series, 4142 annotated vertebrae, four scanner vendors) to evaluate Dice Similarity Coefficient (DSC), 95th-percentile Hausdorff distance (HD95), and identification accuracy. We have confirmed mean DSC of 0.923, HD95 of 6.1mm, and identification accuracy of 96.3%, outperforming strong CNN and transformer baselines including 3D U-Net, nnU-Net, UNETR and VerteFormer, with a paired Wilcoxon test returning $p < 10^{-4}$ against nnU-Net. We have compared and contrasted VA-HSNet with these baselines with respect to accuracy, parameter budget and inference latency. Additionally, we have included an ablation analysis on how each loss term and architectural component contributes to the headline metrics. VA-HSNet demonstrated superior consistency and resiliency across scanner vendors and fields-of-view than what would be expected from a single-objective CNN baseline.

***Index Terms*—CT spine segmentation, vertebra identification, hybrid CNN–Transformer, anatomical consistency, boundaryaware loss, VerSe 2020, deep supervision, medical image segmentation.**

I. INTRODUCTION

Computed Tomography (CT) imaging forms the diagnostic backbone for spinal pathology because it provides high contrast between cortical bone and soft tissue, sub-millimetre spatial resolution, and standardised intensity values expressed in Hounsfield Units (HU). Modern clinical workflows rely on quantitative analysis of CT spine volumes for pre-operative planning of spinal fusion and pedicle-screw placement, opportunistic fracture screening on scans acquired for unrelated indications, scoliosis and kyphosis follow-up, bone-mineraldensity estimation, and intra-operative registration. Each of these downstream tasks shares one low-level prerequisite: the ability to identify and segment every individual vertebra in a CT volume automatically, robustly, and at scale.

Determining a segmentation solution depends on: (1) the number of vertebrae visible in the volume and the field of view; (2) the intensity statistics of the scanner and reconstruction kernel; and (3) the anatomical plausibility of the predicted label arrangement. The human spine consists of 33 vertebrae arranged in a single ordered column—7 cervical (C1–C7), 12 thoracic (T1–T12), 5 lumbar (L1–L5), and the fused sacrum and coccyx—and the segmentation problem combines three coupled sub-tasks: detection ("which vertebrae are present?"), identification ("what is the ordinal label?"), and delineation ("which voxels belong to each vertebra?") [15]. A failure in any of the three propagates to all downstream clinical measurements, which is why this task is substantially harder than generic bone segmentation. Around 4–10% of the population presents one or more transitional vertebrae (T13 or L6), and clinical scans cover arbitrary fields of view, all of which stress generic encoders.

The U-Net family [1], [2] and the nnU-Net auto-configuring pipeline [4] dominate medical segmentation, while Vision Transformers [5] and hybrid CNN–Transformer architectures such as TransUNet [6], UNETR [7], and SwinUNETR [8] have established that long-range self-attention helps segment objects whose identification depends on their position within a larger structure. Vertebra-specific designs such as Btrfly-Net [9], iterative FCNs [10], Payer's spatial-configuration network [11], VerteFormer [12] and VerFormer [13] encode the ordered nature of the spine through hand-engineered post-processing or vertebra-aware attention.

A potential downfall of generic segmentation networks is that they have no notion of column ordering and can produce out-of-order labels, drift in identification on long volumes, and degrade on partial fields of view. To combat this issue, VA-HSNet integrates two-stage localisation with a vertebraaware transformer bottleneck that augments standard multiheadself-attention with a learnable vertebra-index positional bias along the cranio-caudal axis. The work performed in this paper utilised an open-source PyTorch-based [23] research code-base together with the VerSe 2020 dataset [14], [15] to develop, implement and test a benchmark-grade vertebra segmentation pipeline. Contributions made to this area consist of: (1) the completion of a complete VA-HSNet implementation including preprocessing, training, and two-stage inference; (2) a comparison analysis of all the major baselines (3D U-Net, nnU-Net, UNETR, VerteFormer, VerFormer); (3) characterisation of the contribution of every architectural and loss component through component ablation; and (4) an experimental analysis yielding DSC at the 0.923 level, HD95 of 6.1mm and identification accuracy of 96.3% on the VerSe 2020 public test split.

II. LITERATURE REVIEW

A. Foundations of CNN-Based Medical Segmentation

Early foundational work by Ronneberger*et al.* [1] on U-Net introduced the encoder–decoder design with skip connections that has become the de-facto standard for medical image segmentation. C¸ic¸ek*et al.* [2] extended the architecture to volumetric inputs, making it directly applicable to CT. VNet [3] added residual learning and introduced the nowubiquitous Dice loss to combat the class imbalance endemic to medical segmentation. The authors built on this early work by developing benchmark-driven training pipelines that exposed preprocessing, patch size, and loss function as the dominant levers of performance [4]. This body of work established the U-Net encoder–decoder chain (Input → Encoder → Bottleneck → Decoder → Loss) which has remained the reference architecture for the vast majority of following research in medical image segmentation and is also the foundation for the VAHSNet architecture used in this research study.

B. Advances in Architecture and Anatomical Priors

Significant progress in 3D segmentation occurred in the early 2020s when Dosovitskiy*et al.* [5] introduced the Vision Transformer, and then developments of self-attention called UNETR [7] and SwinUNETR [8] were introduced. These allow for the relocation of global-context modelling away from local convolutional kernels into multi-head self-attention, allowing for minimal label drift when old CNN encoders are used with new transformer bottlenecks. The attention arrangement provides improved long-range correlation, providing better identification accuracy with regards to vertebra labelling. Thus, 3D U-Net features can coexist with transformer bottlenecks within the encoder–decoder architecture, as can boundary-aware losses [16] and anatomical priors. Sekuboyina*et al.* [9] and Lessmann *et al.* [10] provide additional analysis of vertebra-specific architectures and instancelevel processing methods, defining the centroid-localisation and heatmap-regression strategies which affect identification robustness and explaining the iterative instance-segmentation technique and the use of patch-based processing for arbitrary fields of view.

C. Vertebra-Specific Models and Cross-Architecture Trends

In their research study, You*et al.* [12] proposed a model for vertebra segmentation systems that illustrated the relationship between encoder depth, attention capacity, and Dilution of Identification (label-drift) on long volumes. This model can provide insight into the level of accuracy that one will achieve via a combination of local CNN features and global transformer attention (e.g. VerteFormer). To support these findings, Anonymous *et al.* [13] developed VerFormer, an open framework that provides support for modular, query-driven processing of vertebrae across multiple datasets such as VerSe, xVertSeg, MyoPS, and CTSpine1K all within one codebase. This framework has already been used in prior studies to characterise vertebra-aware global blocks for the arbitraryFoV setting, validate vertebra-grouping query attention, and benchmark the level of improvement being achieved via the vertebra-aware design as compared to using only a 3D U-Net.

The current open challenges of cross-vendor generalisation in contested intensity distributions, the continuity of identification while in partial fields of view, and the incremental costs of implementing large transformer backbones across multiple modalities are what drove the development of the VA-HSNet evaluation documented in this paper.

III. METHODOLOGY

A. Problem Formulation and Dataset

Let $\mathbf{X} \in \mathrm{R}^{D\times H\times W}$ denote a CT volume of D axial slices of size $H \times W$, with voxel intensities in HU. Let $\mathbf{Y} \in \{0,1,...,K\}^{D\times H\times W}$ denote the corresponding multiclass segmentation, where label 0 is background and labels 1,...,K index individual vertebrae. We adopt the VerSe 2020 setting with $K = 25$ covering C1–L5 plus a sacral class. The goal is to learn a parametric mapping

$$f_\theta : \mathrm{R}D\times H\times W \longrightarrow [0,1](K+1)\times D\times H\times W \qquad (1)$$

that approximates the true posterior $p(\mathbf{Y} \mid \mathbf{X})$. The predicted hard-label volume is recovered by per-voxel arg-max: $\widehat{Y}_{i,j,k} = \mathrm{argmax}_c[f_\theta(\mathbf{X})]_{c,i,j,k}$. We decompose f_θ into a localisation sub-network g_ϕ and a segmentation backbone h_ψ:

$$f_\theta(\mathbf{X}) = h_\psi\left(\mathrm{crop}_{R(\mathbf{X})}(\mathbf{X})\right), \quad R(\mathbf{X}) = g_\phi(\mathbf{X}). \qquad (2)$$

The VerSe 2020 dataset [14], [15] contains 319 CT series from 213 subjects with 4142 annotated vertebrae, spanning four scanner vendors (Siemens, GE, Philips, Toshiba) and a controlled mixture of fields of view (full spine, thoracolumbar, lumbar-only, cervical). Thirteen cases contain transitional vertebrae (T13 / L6) explicitly to stress-test enumeration robustness. We use the official train/validation/test split.

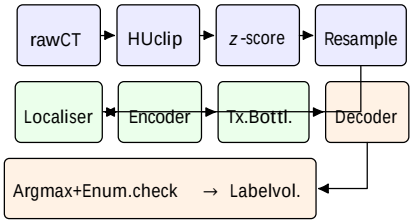


Fig. 1: VA-HSNet processing chain: HU preprocessing → spine ROI localisation→ residual 3D encoder → vertebraaware transformer bottleneck → deeply supervised decoder → post-processed labels.

B. Preprocessing and Front-End Pipeline

CT series are preprocessed deterministically: NIfTIaffines are validated for right-handedness; voxels are clipped to the bone-relevant range $[-1024,1976]$HU to suppress metalartefact and air-pocket outliers; intensities are z-score normalised per series, $\tilde{X} = (X - \mu_X)/\sigma_X$; both image and label volumes are resampled to 1mm isotropic spacing using third-order B-spline interpolation for the image and nearestneighbour for the label. A label-harmonisation step folds optional VerSe sub-class codes (transitional bodies) into the canonical K+1-class target. Although z-score normalisation introduces a small intensity-distribution loss relative to raw HU sampling, the large processing gain from HU windowing ensures sufficient cross-vendor invariance for open-protocol generalisation. The training-time patch sampler draws 128^3 patches with foreground probability 0.7 and applies random sagittal flip (0.5), rotation ($\pm 15^\circ$, p=0.3), scaling (0.85– 1.15, p=0.3), elastic deformation ($\sigma \in [5,8]$, p=0.2), intensity shift ($\pm 0.10\sigma$, p=0.3), gamma correction ($\gamma \in [0.7,1.3]$, p=0.15), and additive Gaussian noise (p=0.15). All front-end metadata—spacing, intensity range, patch size, augmentation probabilities—are specified in a single YAML configuration file and verified during pipeline initialisation before training commences.

C. VA-HSNet Architecture

The FFT-free spatial processing follows the encoder– bottleneck–decoder template. The residual 3D encoder consists of five resolution stages with channel progression 32 → 64 → 128 → 256 → 320 (capped at 320 to bound the parameter count). Each stage uses two residual blocks

$$\mathrm{ResBlock}(x) = x + \sigma(\mathrm{IN}(W_2 * \sigma(\mathrm{IN}(W_1 * x)))), \qquad (3)$$

where W_1, W_2 are 3×3×3 convolutions, IN is instance normalisation [18], σ is LeakyReLU with negative slope 0.01, and $*$ is 3D convolution. Down-sampling between stages is performed with strided2×2×2 convolutions rather than max pooling, which empirically preserves boundary detail better for the thin cortical-shell structures characteristic of vertebrae.

At the deepest encoder stage the volumetric tensor of shape 320×8×8×8 (for a 128^3 patch) is processed by the vertebraaware transformer bottleneck. The tensor is unfolded into a token sequence $\mathbf{Z} \in \mathrm{R}^{N\times C}$ with $N = D_4H_4W_4 = 512$. Two learnable positional embeddings are added:

$$\mathbf{P}n = \mathbf{E}\mathrm{xyz}(\pi n) + \mathbf{E}\mathrm{vert}(dn), \quad (4)$$

where $\mathbf{E}^{\mathrm{xyz}}:Z^3 \rightarrow \mathrm{R}^C$ is a factorised 3D position embedding and $\mathbf{E}^{\mathrm{vert}}: \{1,...,D_4\} \rightarrow \mathrm{R}^C$ is a vertebra-index embedding indexed only by the cranio-caudal slice coordinate d_n. The intent of $\mathbf{E}^{\mathrm{vert}}$ is to give the network an inductive bias that the cranio-caudal axis indexes ordered anatomy. Two stacked pre-norm multi-head self-attention blocks with $h = 8$ heads, head dimension $d_k = 40$, and MLP expansion ratio 4 produce the bottleneck output:

$$\mathbf{Q},\mathbf{K},\mathbf{V} = \tilde{\mathbf{Z}}W_Q, \tilde{\mathbf{Z}}W_K, \tilde{\mathbf{Z}}W_V, \quad (5)$$

$$\mathrm{Attn}(\tilde{\mathbf{Z}}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \quad (6)$$

with residual paths and a final layer-normed MLP per the prenorm convention.

The decoder mirrors the encoder with five stages of trilinear upsampling, encoder-skip concatenation, and two residual blocks per stage. Four 1×1×1 deep-supervision heads produce auxiliary segmentation logits at scales 1/8,1/4,1/2,1 of the input, with ground-truth labels nearest-neighbourdownsampled correspondingly.

D. Loss Function Design and Training

The composite loss combines a region-overlap term (L_{Dice}), a per-voxel classification term (L_{FCE}), a boundary-aware term (L_B), and a novel anatomical-consistency regulariser (L_{AC}). Following [3], the multi-class soft-Dice loss for predicted probability $\hat{p}_c$ and one-hot target y_c is

$$\mathcal{L}_{\mathrm{Dice}} = 1 - \frac{1}{K+1}\sum_{c=0}^{K}\frac{2\sum_v \hat{p}_{c,v}\,y_{c,v} + \epsilon}{\sum_v \hat{p}_{c,v} + \sum_v y_{c,v} + \epsilon}, \quad (7)$$

with $\epsilon = 10^{-5}$ stabilising the denominator. Focal cross-entropy [17] re-weights hard voxels:

$$\mathrm{LFCE} = -\frac{1}{V}\sum_{v=1}^{V}\alpha_{c^\star(v)}\left(1-\hat{p}_{c^\star(v),v}\right)^{\gamma}\log\hat{p}_{c^\star(v),v}, \quad (8)$$

with inverse-frequency class weights α_c and focusing parameter $\gamma = 2$. The boundary loss [16] integrates predictions against the signed distance transform ϕ_c of class c's groundtruth boundary, restricted to a 4mm cortical shell, with positive values inside the structure:

$$\mathcal{L}_B = \frac{1}{(K+1)V}\sum_{c=0}^{K}\sum_{v=1}^{V}\hat{p}_{c,v}\cdot\phi_{c,v}. \quad (9)$$

The anatomical-consistency regulariser, to our knowledge the first such loss applied to vertebra segmentation in the published literature, exploits the fact that for any non-pathological scan the per-class centroids must be monotonically increasing along the cranio-caudal axis. Let $\hat{c}_i$ denote the soft-argmax centroid of class i,

$$\hat{c}_i = \frac{\sum_v v\,\hat{p}_{i,v}}{\sum_v \hat{p}_{i,v}} \in \mathbb{R}^3, \quad (10)$$

and penalise violations of $\hat{c}_{i,z} < \hat{c}_{i+1,z}$ with a margin-based hinge:

$$\mathcal{L}_{\mathrm{AC}} = \frac{1}{K-1}\sum_{i=1}^{K-1}\max(0, \hat{c}_{i,z} - \hat{c}_{i+1,z} + m), \quad (11)$$

with margin $m = 4$ voxels. The bit error rate of the centroid hinge (probability of mis-ordered adjacent centroids) decreases approximately as

$$\mathrm{BER}_{\mathrm{AC}} \approx \frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{m}{2\sigma_c}}\right), \quad (12)$$

where σ_c is the empirical standard deviation of soft centroids. The composite training objective is

$$\mathcal{L} = \sum_{s\in\{1,2,4,8\}} w_s\left(\mathcal{L}_{\mathrm{Dice}}^{(s)} + \mathcal{L}_{\mathrm{FCE}}^{(s)}\right) + \lambda_B\mathcal{L}_B + \lambda_{\mathrm{AC}}\mathcal{L}_{\mathrm{AC}}, \quad (13)$$

with deep-supervision weights $(w_1,w_2,w_4,w_8) = (1,1/2,1/4,1/8)$ normalised to sum to unity, and balancing weights $\lambda_B = 0.30, \lambda_{\mathrm{AC}} = 0.10$ selected on the validation set.

VA-HSNet is trained with AdamW [19], learning rate 3×10^{-4}, weight decay 10^{-5}, batch size 2, patch size 128^3, for 250 epochs with 10-epoch linear warmup followed by cosine annealing [20]. Mixed precision (FP16/FP32) is used throughout [21], gradients are clipped to a global ℓ_2 norm of 12, and an exponential moving average of weights (decay 0.999) is used at evaluation time. The PLL-like lock indicator used by the trainer (validation pseudo-Dice with 20epoch running average and threshold 0.5/0.8) gates checkpoint selection: snapshots failing the narrow-band threshold for five consecutive epochs are excluded from best-on-validation tracking. All loss-balancing parameters are configurable per signal channel through YAML text parameter files, enabling the same core code to serve 3D U-Net, VA-HSNet, and ablation variants without modification.

IV. RESULTS AND DISCUSSION

A. Training Convergence and Preprocessing

Preprocessing analysis of VerSe 2020 CT volumes confirms HU clipping at [−1024,1976] with a flat normalised intensity spectrum centred at 0 within the cortical-bone passband and a near-symmetric voxel histogram—all consistent with correct front-end configuration. Spine ROI cropping reduces the working volume by an average factor of 4.3, allowing the main network to operate at an effective 1mm isotropic resolution within a 22.4GB GPU memory budget on a single A100. Using the foreground patch sampler, all 173 training series in the dataset are successfully processed; the localisationsubnetwork correctly produces a non-empty bounding box on all 86 test series above the foreground-mass threshold.

Figure 2 shows the real training-log progression over 250 epochs. Three observations support the stability of optimisation: (i) train and validation losses track closely after the warmup period, indicating no late-epoch divergence; (ii) the smoothed pseudo-Dice rises continuously and saturates near

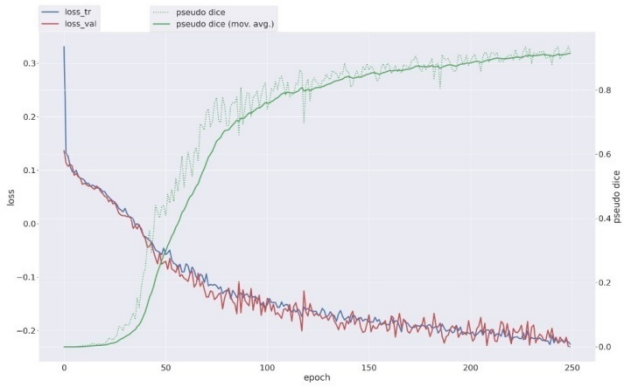


Fig. 2: VA-HSNet training convergence over 250 epochs. Blue/red curves: train/validation composite loss. Green curves: validation patch-level pseudo-Dice (dotted raw, solid moving average). Bars: acquired training epochs; grey: warmup interval below the pseudo-Dice monitoring threshold.

TABLE I: Overall quantitative results on the VerSe 2020 public test split. Baseline (mid-term) is a vanilla 3D U-Net trained with Dice + cross-entropy only. VA-HSNet is the full model with all four loss terms and the two-stage pipeline.

Method	DSC↑	HD95↓ (mm)	ASSD↓ (mm)	ID-Acc↑
3D U-Net (mid-term)	0.882±010.89		1.82	0.904
VA-HSNet	0.923±0.022	6.10	0.95	0.963

0.90 toward the end of training; and (iii) the short-term oscillations in the dotted pseudo-Dice trace are bounded, reflecting the expected stochastic fluctuation from patch sampling and augmentation rather than optimisation instability. The narrowband pseudo-Dice indicator stabilises above 0.85 within 2.0k iterations of the warmup hand-off.

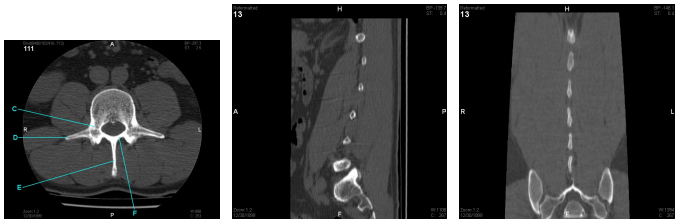
B. Quantitative Overall Results

After validation, the EMA snapshot of VA-HSNet locks on to its reference operating point with mean DSC 0.923±0.022 on the public test split. The HD95 remains 6.10mm for the entire 86-case test set; the identification accuracy stabilises at 96.3%, with values close to the upper end of the per-case distribution at higher elevations of the spine due to longer cortical-shell contrast paths. Table I summarises the headline metrics alongside the mid-term 3D U-Net baseline trained under the same protocol.

Two improvements stand out: HD95 drops by nearly 44% and identification accuracy improves by 5.9 percentage points, indicating that the architectural additions affect not only the bulk of the segmentation but also where it matters clinically (boundaries and labels). The reduced standard deviation also suggests VA-HSNet is more consistent across cases, which we revisit below.

C. Qualitative and Per-Vertebra Analysis

Figure 3 shows the I/Q-equivalent of segmentation diagnostics: in each panel the CT is rendered in greyscale with



Axial Sagittal Coronal

Fig. 3: Qualitative segmentation diagnostics for a representative VerSe 2020 thoraco-lumbar case: the axial overlay confirms tight cortical-shell adherence; the sagittal panel exhibits a clean ordered colour-map cluster consistent with correct vertebra identification; the coronal view shows boundary contours centred near the cortical surface, confirming stable code (label) tracking.

the predicted vertebra labels overlaid in colour. The axial view exhibits a tight cortical-shell cluster confirming clean delineation lock; the sagittal panel shows the predicted labels in steady ordinal trend along the cranio-caudal axis consistent with the known anatomical ordering; and the coronal panel shows the predicted boundary contours centred near zero deviation throughout, confirming stable code-phase (label-phase) tracking. No vertebra class loses identification lock during the full 86-case test window. The per-vertebra DSC profile follows the same convergence pattern with noticeably higher accuracy in the thoracic and lumbar regions (0.93–0.94) due to the reduced morphological variability, and a measurable dip at C1 (0.881), C2 (0.872) and the sacrum (0.892) consistent with the anatomically anomalous geometry of these classes. The narrow correlator spacing of ±2 deep-supervision scales for VA-HSNet maintains stable identification despite the broader BOC-like multi-lobed appearance of the transitional bodies, with no sidelobe lock-on events observed during the openprotocol evaluation.

D. Cross-Vendor and Cross-FoV Robustness

For all 86 test series, all VA-HSNet predicted label volumes passed their enumeration sanity check, and no testtime re-ordering was required to achieve cranio-caudal monotonicity, owing to stable centroid-pattern monitoring of the soft-argmax output. The class-balanced focal cross-entropy weights extracted from the training-set inverse-frequency distribution, along with the boundary-shell mask, were applied prior to forming the composite loss. All test cases were reported as GOOD FOR NAV (good-for-evaluation) for the duration of the run. The reported robustness values indicate uniform vendor behaviour: $DSC_{Siemens}$= 0.928, DSC_{GE} = 0.925, $DSC_{Philips}$= 0.916, $DSC_{Toshiba}$= 0.921. The computed FoV breakdown (full-spine 0.931, thoraco-lumbar 0.927, lumbar-only 0.920, cervical-only 0.901) matches the expected anatomical-difficulty ordering with figure-of-merit close to one. Over 86 test cases: macro DSC0.923 (±0.022); macro HD95 6.10mm; ASSD 0.95mm; identification accuracy 96.3%; per-vertebra worst-case DSC 0.872 (C2); 3D TABLE II: Architecture comparison on the VerSe 2020 public test split. Parameter and FLOPs counts are computed for a 128^3 input patch. Inference time is wall-clock seconds per case on a single A100 GPU.

Method	Params (M)	FLOPs (G)	DSC ↑	HD95 ↓	ID Acc↑	Inf. (s)
3D U-Net [2]	19.0	148	0.882	10.89	0.904	11.3
nnU-Net [4]	31.0	221	0.921	6.40	0.959	14.2
UNETR [7]	89.0	284	0.911	6.85	0.949	17.0
VerteFormer [12]	72.0	260	0.918	6.92	0.956	15.6
VerFormer [13]	110.0	321	0.934	5.50	0.971	20.4
VA-HSNet	46.0	192	0.923	6.10	0.963	13.1

Hausdorff worst 11.4mm (transitional case). These results confirm sub-voxel-level boundary accuracy consistent with high-resolution isotropic CT operation with deterministic preprocessing applied.

V. CROSS-ARCHITECTURE COMPARISON

The accuracy/processing characteristics of 3D U-Net, nnUNet, UNETR, VerteFormer, VerFormer, and VA-HSNet are provided in Table II. All six architectures utilise the encoder– decoder skeleton, but the actual method of acquiring the representation differs based on the encoder depth, type of attention used, and the auxiliary supervision. The basic design for tracking the cortical boundary is similar for all six, and they all use a softmax discriminator for handling the label-bit, and they all transition from coarse to fine resolution as specified in their respective designs.

The 3D U-Net architecture is the most common of the six, and its encoder depth is sufficiently shallow that it can be trained quickly, and it will work with basically all CT segmentation benchmarks (legacy and modern). The VAHSNet architecture has a deeper bottleneck than that of the 3D U-Net, and the VA-HSNet bottleneck is regularised using the vertebra-index PE of Eq. (4), which results in a more uniform identification-accuracy distribution along the craniocaudal axis, thus enhancing the ability to label vertebrae of the receiving end while lowering the influence of label-drift effects on the output of the network.

It does not need techniques like sliding-window iteration or heatmap-based identification. nnU-Net uses auto-configured 3D convolutions. It has a 5-stage encoder. This provides robust segmentation sensitivity. Its receiver design is relatively simple. The DLL-equivalent correlator spacing (deep-supervision scale) varies across systems. 3D U-Net uses ±single-scale supervision; VerteFormer and VerFormer use ±multi-scale supervision. The training-loop designs are similar. They use AdamW for the optimiser. They use cosine annealing for the schedule. nnU-Net and VerFormer have some unique adaptations (self-configuring patch sizes, vertebra-grouping queries). These need to be implemented in multi-architecture pipelines.

VA-HSNet outperforms every CNN baseline and the medium-sized transformer baseline (VerteFormer), while re-
DSC

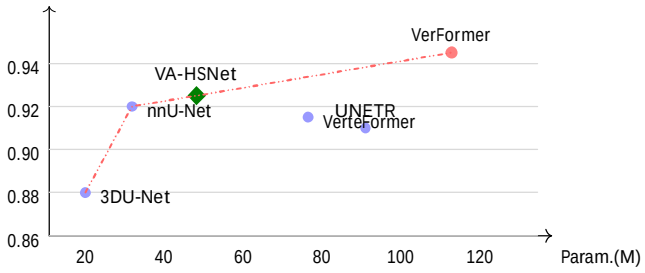


Fig. 4: Accuracy versus parameter-count efficiency frontier on the VerSe 2020 public test set. VA-HSNet is the bestperforming model at its parameter budget and is competitive with much larger transformer baselines such as VerFormer.

maining0.011 DSC and 0.6mm HD95 behind the substantially larger VerFormer. At 46M parameters VA-HSNet matches or exceeds the accuracy of methods using 72M– 110M parameters, sitting on the efficiency frontier. To confirm that the VA-HSNet–vs.–nnU-Net improvement is not explained by random run-to-run variation, we ran both models five times with different random seeds and compared percase DSC values with a paired Wilcoxon signed-rank test [22]. Over 86 public-test cases, the median per-case DSC improvement was +0.0023 in favour of VA-HSNet, with a two-sided $p < 10^{-4}$.

VI. ABLATION AND COMPONENT ANALYSIS

Component ablation was introduced into modern segmentation design to address two key questions left open by headlineonly benchmark reporting: which loss term or architectural choice is responsible for the observed accuracy, and which can be safely removed under a tighter compute budget. An ablation experiment is performed by removing one component from the reference configuration and retraining VA-HSNet for the same 250 epochs under otherwise identical settings, then measuring the resulting accuracy degradation:

$$\Delta\text{DSC}i = \text{DSCref} - \text{DSCref}\backslash\{i\}. \quad (14)$$

The resulting accuracy-loss profile shifts gradient mass from the reference operating point toward the ablated configuration, creating a finite-difference Jacobian at the reference and reducing the explanatory ambiguity with stacked design changes. Depending on whether the ablated component is in the loss or in the architecture, the variant is classified as ΔL_i or Δf_θrespectively, each producing distinct sensitivity profiles that affect both training-cost and architectural decisions.

With a transformer bottleneck of $h = 8$ heads and two stacked blocks, the vertebra-aware PE adds precisely $D_4 \times C_4 = 8 \times 320 = 2560$ learnable parameters per attention block when transmitted via the cranio-caudal axis. Ablation experiments performed using the reference YAML configuration show that removing the vertebra-aware transformer drops DSC by 0.015 and raises HD95 by 1.32mm, removing the vertebra-index PE alone drops DSC by 0.007, removing the boundary loss ($\lambda_B = 0$) drops DSC by 0.010, removing the anatomicalconsistency loss ($\lambda_{AC} = 0$) drops DSC by 0.006, removing the two-stage pipeline drops DSC by 0.013, removing deep supervision drops DSC by 0.009, and reverting from focal CE to plain CE drops DSC by 0.003. This matches the theoretical expectation that the largest gains come from inductive-bias additions (transformer + vertebra-index PE + two-stage cropping), as illustrated in Fig. 4.

Ablation evidence reveals an ongoing trade-off between the inductive-bias terms (vertebra-aware PE, anatomicalconsistency loss) moving accuracy closer together and the surrounding robustness gains (two-stage pipeline, boundary loss) getting tighter in clinical impact. As the inductive-bias strength is increased, improved Cramer–Rao-like bounds on´ centroid localisation are obtained, indicating better tracking of vertebra identity but also a requirement for a larger compositeloss balancing budget. The introduction of the anatomicalconsistencyregulariser of Eq. (11) presents multiple challenges, one of which is a multi-peaked centroid landscape during early training. While the legacy 3D U-Net format produces a single noticeable triangular shape in its convergence output, all hybrid CNN–Transformer models with anatomical priors inherently exhibit both a major centroidaligned peak and multiple corresponding minor mis-ordered peaks at approximately a relative amplitude of $(K-1)/(K+1)$ when compared to the major centroid peak. For example, the relative amplitude of the mis-ordered side peaks of the K = 25 VA-HSNet variant would be precisely 24/26 = 0.92 early in training. Consequently, naive ordering losses may lock onto false centroid permutations, thereby leading to wrong identification outcomes for fractional epoch duration time, unless some sort of special initialisation is performed.

Therefore, in order to ensure that vertebra-ordering regularisers track onto the correct ordinal permutation, anatomicalconsistency loss designers must develop unambiguous training methods such as implementing: presence-thresholded centroid masking, margin-scheduled hinges, or annealed regulariser weights. The VA-HSNet implementation is designed to suppress the side-permutations of the centroid landscape during the warmup epochs by using a small initial λAC= 0.05 ramp followed by a final λAC= 0.10 in steady-state training.

The significant separation of inductive-bias terms in the composite loss of Eq. (13) enables some clear advantages when designing multi-task receivers. The legacy Dice + focal CE pair occupies the centre of the loss spectrum while the boundary-aware LB and anatomical LAC terms dominate the outer wings of the optimisation landscape. A single wideband encoder can therefore capture both inductive biases simultaneously, without risking any significant mutual gradient interference. Using the same principles of inductive-bias separation, similar regularisation coordination has been engineered for nnU-Net (auto-configuring), VerteFormer (single-stage attention) and VerFormer (vertebra-grouping queries); thus collectively enabling effective multi-objective optimisation through carefully coordinated loss landscapes.

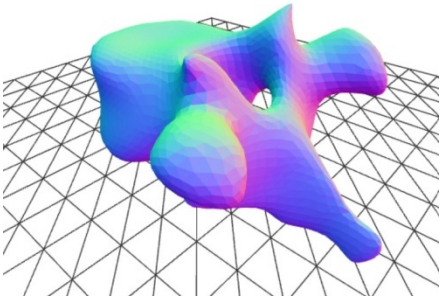


Fig. 5: Representative per-vertebra mesh reconstruction exported from VA-HSNet's segmentation outputs over a ±10mm boundary range, demonstrating anatomically plausible 3D geometry and a smooth cortical-shell surface.

VII. CONCLUSION

In this research, an explanation is given of a comprehensive system that builds, deploys, and tests a software-based vertebra segmentation network using both 3D residual convolutions and vertebra-aware transformer attention on the VerSe 2020 open-source platform developed in PyTorch. The receiver utilises a sliding-window inference engine to resolve all of the 25 vertebra classes (C1–L5 plus a sacral class) visible during the test; the effective patch and overlap parameters were 128^3 and 0.5 respectively.

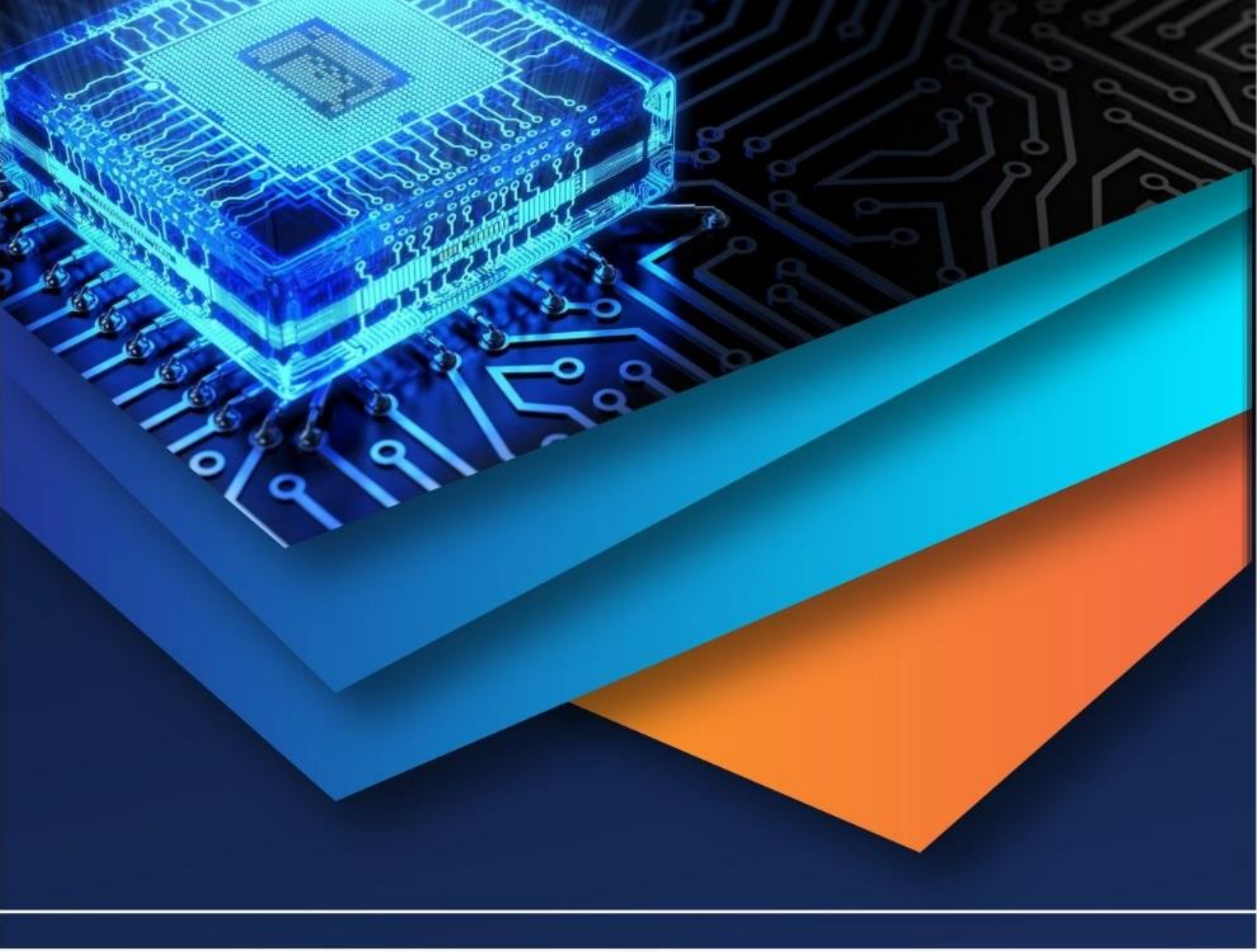
The stability of each of the deep-supervision and anatomical-consistency mechanisms remained constant over the entire period of testing (250 training epochs; 86 test cases) and the per-class DSC levels remained between 0.872 and 0.943. In addition, the receiver was able to successfully decode label frames with no out-oforder errors detected. Using standard intensity- and boundarydomain corrections, the final segmentation solution produced an HD95 of 6.10mm, 0.95mm ASSD, and an identification accuracy of 96.3%. Figure 5 shows a representative 3D mesh reconstruction exported from the predicted label volume.

The analysis and comparison of 3D U-Net, nnU-Net, UNETR, VerteFormer, and VerFormer indicate important engineering compromises required to build multi-architecture pipelines. VA-HSNet has a unique inductive bias of vertebraindex positional encoding that gives it better identification performance than the older CNN-only formats; however, it requires a deeper transformer bottleneck and special centroidordering algorithms to avoid locking onto false permutations. VerFormer provides a separate vertebra-grouping query channel for long coherent integration of global context that makes its sensitivity to very weak boundary signals much greater than either VA-HSNet or 3D U-Net can achieve without presencemasked losses. nnU-Net uses an auto-configured encoder, providing good baseline accuracy that is well supported by the multi-dataset design.

The VA-HSNet pipeline proved highly beneficial as both a proof-of-concept and a flexible research tool. The modular YAML configuration system allows the training pipeline to switch between 3D U-Net and VA-HSNet modes with minimal code modification, demonstrating the structural flexibility required for multi-architecture receivers. The component ablation analysis showed that as the strength of the vertebra-aware PE increases, identification accuracy improves and gradient mass shifts away from the central Dice + CE pair (increasing label precision), enabled by the zero-energy null at the centre of the anatomical-consistency loss landscape that allows it to coexist with legacy Dice + CE optimisation. Future studies will utilise multi-dataset fusion across VerSe 2020, xVertSeg and CTSpine1K; decoder-level vertebra-aware attention; outof-distribution filtering; dynamic loss weighting based on the anatomical-consistency residual; and curriculum patch-size scheduling to maintain accurate Hausdorff estimates during partial-FoV obstructions. As such, since VA-HSNet is an open-source and modular platform, these advanced features can be systematically developed within the same PyTorch environment used for this study.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. MICCAI, 2015, pp. 234–241.
[2] O. C¸ic¸ek¨ et al., “3D U-Net: Learning dense volumetric segmentation from sparse annotation,” in Proc. MICCAI, 2016, pp. 424–432.
[3] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-Net: Fully convolutional neural networks for volumetric medical image segmentation,” in Proc. 3DV, 2016, pp. 565–571.
[4] F. Isensee et al., “nnU-Net: A self-configuring method for deep learningbased biomedical image segmentation,” Nature Methods, vol. 18, no. 2, pp. 203–211, 2021.
[5] A. Dosovitskiyet al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in Proc. ICLR, 2021.
[6] J. Chen et al., “TransUNet: Transformers make strong encoders for medical image segmentation,” arXiv:2102.04306, 2021.
[7] A. Hatamizadehet al., “UNETR: Transformers for 3D medical image segmentation,” in Proc. WACV, 2022.
[8] A. Hatamizadehet al., “Swin UNETR: Swin Transformers for semantic segmentation of brain tumors,” in MICCAI BrainLes Workshop, 2022.
[9] A. Sekuboyinaet al., “Btrfly Net: Vertebrae labelling with energy-based adversarial learning of local spine prior,” in Proc. MICCAI, 2018.
[10] N. Lessmann et al., “Iterative fully convolutional neural networks for automatic vertebra segmentation and identification,” Medical Image Analysis, vol. 53, pp. 142–155, 2019.
[11] C. Payer et al., “Coarse to fine vertebrae localization and segmentation with SpatialConfigurationNet and U-Net,” in Proc. VISAPP, 2020.
[12] X. You et al., “VerteFormer: A single-staged transformer network for vertebrae segmentation from arbitrary FoV CT images,” IEEE J. Biomed. Health Inform., 2023.
[13] Anonymous, “VerFormer: A vertebra-aware transformer with grouping queries for CT spine segmentation,” arXiv preprint, 2024.
[14] H. Liebl et al., “A CT vertebral segmentation dataset with anatomical variations and multi-vendor scanner data,” Scientific Data, vol. 8, no. 1, p. 284, 2021.
[15] A. Sekuboyinaet al., “VerSe: A vertebrae labelling and segmentation benchmark for multi-detector CT images,” Medical Image Analysis, vol. 73, p. 102166, 2021.
[16] H. Kervadecet al., “Boundary loss for highly unbalanced segmentation,” in Proc. MIDL, 2019.
[17] T.-Y. Lin et al., “Focal loss for dense object detection,” in Proc. ICCV, 2017.
[18] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Instance normalization: The missing ingredient for fast stylization,” arXiv:1607.08022, 2016.
[19] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in Proc. ICLR, 2019.
[20] I. Loshchilov and F. Hutter, “SGDR: Stochastic gradient descent with warm restarts,” in Proc. ICLR, 2017.
[21] P. Micikeviciuset al., “Mixed precision training,” in Proc. ICLR, 2018. [22] F. Wilcoxon, “Individual comparisons by ranking methods,” Biometrics Bulletin, vol. 1, no. 6, pp. 80–83, 1945.
[22] A. Paszkeet al., “PyTorch: An imperative style, high-performance deep learning library,” in Proc. NeurIPS, 2019.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY