2026, 51(6) DOI:10.11817/j.issn.1672-7347.2026.260052

·APPLICATIONS OF ARTIFICIAL INTELLIGENCE IN CLINICAL MEDICINE·
·人工智能在临床医学中的应用·

HAR-Net: Automatic delineation of clinical target volumes for cervical cancer guided by hierarchical anatomical refinement

CHENG Tingting[1, 2, 3], DENG Liang[1, 2, 3], LIANG Zhan[1, 2, 3], TANG Du[1, 2, 3], ZHANG Zijian[1, 2, 3]

[1. Department of Radiation Oncology, Xiangya Hospital, Central South University, Changsha 410008; 2. National Clinical Research Center for Geriatric Disorders (Xiangya Hospital of Central South University), Changsha 410008; 3. Hunan Engineering Research Center for Intelligent Medical Imaging, Furong Laboratory, Changsha 410008, China]

ABSTRACT

Objective: In radiotherapy for cervical cancer, accurate delineation of the clinical target volume (CTV) is essential for ensuring prescription dose coverage while minimizing radiation exposure to organs at risk (OARs). In clinical practice, however, CTV boundaries are often poorly defined on standard medical images and exhibit substantial morphological variability. Consequently, automated methods based solely on image appearance or a single segmentation framework may fail to consistently capture the fine-grained anatomical structures and spatial coherence of the CTV. Existing methods are therefore prone to boundary displacement, morphological discontinuity, and inadequate generalizability across patients, which may compromise the reproducibility and clinical applicability of radiotherapy planning. This study aims to develop a hierarchical anatomical refinement network (HAR-Net) for automatic CTV delineation in cervical cancer. By explicitly integrating anatomical priors and spatial relationships in addition to image appearance information, HAR-Net was designed to improve segmentation accuracy and consistency.

Methods: Computed tomography images and manual delineations of the CTV and 5 OARs, the bladder, rectum, bilateral femoral heads, and small bowel, were retrospectively collected from 132 patients with cervical cancer who received radiotherapy at Xiangya Hospital, Central South University, between May 2022 and May 2025. All patients had stage IB2-IVA disease according to the 2018 International Federation of Gynecology and Obstetrics (FIGO) staging system. The delineations were completed by experienced radiation oncologists. Patients were randomly divided into a training set (n=92), validation set (n=15), and test set (n=25). HAR-Net used nnU-Net as its backbone and incorporated 2 collaborative modules. First, multi-scale OAR-CTV cross-attention (MOCCA) dynamically integrated the anatomical context of OARs into the CTV segmentation pathway through window-based three-dimensional cross-attention. Second, tri-axial coordinate attention

Date of reception: 2026-02-04
First author: CHENG Tingting, Email: chengtingting@csu.edu.cn, ORCID: 0000-0003-4198-079X
Corresponding author: ZHANG Zijian, associate professor, Email: wanzzj@csu.edu.cn, ORCID: 0000-0002-2702-8025
Foundation item: This work was supported by the Natural Science Foundation of Hunan Province (2026JJ80019, 2026JJ60323) and the Scientific Research Program of Furong Laboratory (the Program of the Department of Science and Technology of Hunan Province) (2024PT5110), China.
Open access: This is an open access article under Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) (https://creativecommons.org/licenses/by-nc-nd/4.0/).

©Journal of Central South University (Medical Science). All rights reserved.

(TCA) encoded global positional information along the depth, height, and width axes and spatially refined the features through adaptive reweighting. Segmentation performance was evaluated using the Dice similarity coefficient (DSC), 95th-percentile Hausdorff distance (HD95), and average symmetric surface distance (ASSD). HAR-Net was compared with several mainstream segmentation methods, including MultiResUNet, AMTA-Net, Swin UNETR, ResUNet-a, and nnU-Net.

Results: HAR-Net outperformed multiple state-of-the-art segmentation methods. It achieved a mean DSC of 82.420%, an HD95 of 8.390 mm, and an ASSD of 1.920 mm. Compared to the nnU-Net baseline, HAR-Net improved the DSC by 3.260 percentage points and reduced the HD95 by 4.080 mm. Ablation experiments further demonstrated that both the MOCCA and TCA modules contributed significantly to the improvement in model performance.

Conclusion: By hierarchically integrating multi-scale anatomical context from OARs with tri-axial spatial priors, HAR-Net effectively improves the accuracy and anatomical consistency of automatic CTV delineation for cervical cancer. This framework may reduce the need for repeated manual corrections during clinical delineation, decrease the risk of unnecessary radiation exposure to surrounding OARs, and enhance the overall quality and safety of radiotherapy planning. HAR-Net therefore represents a promising auxiliary tool for automated radiotherapy planning.

KEY WORDS cervical cancer; clinical target volume; deep learning; anatomical priors; medical image segmentation

HAR-Net：基于分层解剖精细化指导的宫颈癌临床靶区自动勾画

程婷婷[1,2,3]，邓俍[1,2,3]，梁瞻[1,2,3]，唐杜[1,2,3]，张子健[1,2,3]

[1. 中南大学湘雅医院肿瘤放疗科，长沙 410008；2. 国家老年疾病临床医学研究中心(中南大学湘雅医院)，长沙 410008；3. 智慧医学影像湖南省工程研究中心，芙蓉实验室，长沙 410008]

[摘 要] 目的：在宫颈癌的放射治疗中，临床靶区(clinical target volume，CTV)的准确勾画是保证处方剂量覆盖与降低危及器官(organs at risk，OARs)受照剂量的关键。然而在实际临床工作中，CTV边界往往在标准医学影像上呈现不清晰、形态差异大等特点，导致基于外观或单一分割框架的自动方法难以稳定捕捉CTV的解剖学细粒度结构与空间一致性。现有方法因此容易出现分割边界偏移、形态不连续及不同病例间泛化不足等问题，进而影响放疗计划的可重复性与临床可用性。为此，本研究旨在提出一种面向宫颈癌CTV自动勾画的分层解剖精细化网络(hierarchical anatomical refinement network，HAR-Net)，可以在影像外观信息之外，显式融合解剖学先验与空间关联，从而提升分割精度与一致性。**方法：**回顾性收集2022年5月至2025年5月于中南大学湘雅医院接受放疗的132例宫颈癌患者[国际妇产科联盟(International Federation of Gynecology and Obstetrics，FIGO)2018版分期IB2~IVA期]的CT影像，以及由资深放射肿瘤科医师完成的CTV和5个OARs(膀胱、直肠、双侧股骨头、小肠)的手工勾画数据，并将其随机分为训练集(92例)、验证集(15例)和测试集(25例)。HAR-Net以nnU-Net为骨干，引入2个协同模块：1)多尺度OAR-CTV交叉注意力(multi-scale OAR-CTV cross attention，MOCCA)，通过窗口式三维交叉注意力将OARs解剖上下文动态融合至CTV分割路径；2)三轴坐标注意力(tri-axial coordinate attention，TCA)，沿深度、高度、宽度三轴编码全局位置信息，对特征进行空间精细化重加权。采用Dice相似系数(Dice similarity coefficient，DSC)、95%豪斯多夫距离(95th percentile Hausdorff distance，HD95)和平均对称表面距离(average symmetric surface distance，ASSD)评估分割性能，并与MultiResUNet、AMTA-Net、Swin UNETR、ResUNet-a及nnU-Net等主流方法对比。**结果：**HAR-Net较多种当前先进分割方法表现更优。其平均DSC为82.420%，HD95为8.390 mm，ASSD为1.920 mm。与nnU-Net基线模型相

©Journal of Central South University (Medical Science). All rights reserved.

比，HAR-Net的DSC提高3.260个百分点，HD95减少4.080 mm。消融实验进一步证实，MOCCA和TCA模块对模型性能提升均具有显著贡献。**结论：**HAR-Net通过分层融合多尺度OAR解剖上下文与三轴空间先验，有效提升了宫颈癌CTV自动勾画的准确性和解剖一致性。该框架有望减少临床勾画中的重复修正工作，降低周围OARs受照风险，提高放疗计划的整体质量与安全性，为自动放疗计划提供了有前景的辅助工具。

[关键词] 宫颈癌；临床靶区；深度学习；解剖学先验；医学图像分割

Accurate delineation of the clinical target volume (CTV) is a cornerstone of effective radiotherapy planning for cervical cancer, directly influencing treatment efficacy and patient outcomes[1-2]. The CTV is conceptualized not merely as an expansion of the visible gross tumor volume but as a clinically defined volume encompassing regions presumed to contain microscopic disease. Unlike organs with visually distinct borders, CTV boundaries are often indistinct on standard imaging modalities like computed tomography (CT). Their delineation relies critically on the interpretation of complex anatomical relationships and strict adherence to consensus guidelines rather than direct grayscale variations in the image[3]. This fundamental characteristic—that CTV is a "virtual" volume defined by anatomical understanding and clinical protocols—poses a significant challenge for manual contouring, leading to inter-observer variability and consuming considerable clinical resources[4-5]. Consequently, developing robust and accurate automated CTV delineation methods is paramount for improving treatment consistency, efficiency, and potentially, patient prognosis by minimizing radiation toxicity to surrounding organs at risk (OARs) while ensuring adequate target coverage[6].

Deep learning, particularly convolutional neural networks (CNNs), has revolutionized computer vision since seminal works like Krizhevsky, et al[7] demonstrated their power on large-scale image classification, and studies such as Simonyan, et al[8] highlighted the importance of network depth. This success rapidly translated to medical image analysis. The application of CNNs to dense prediction tasks such as segmentation was significantly advanced by architectures like fully convolutional networks (FCNs)[9], paving the way for models like the U-Net[10], its 3D extension[11], and their subsequent variants, which have demonstrated considerable success in various medical image segmentation tasks[12-16]. However, the unique, guideline-driven nature of CTV delineation presents specific challenges that generic segmentation architectures, which primarily learn from visual patterns, often fail to adequately address. Because CTV boundaries are not always visually apparent and are defined by intricate anatomical context and clinical rules, standard models struggle to internalize and apply this crucial prior knowledge. While some approaches have attempted to incorporate anatomical priors, such as multi-task learning for OAR segmentation[17] or attention mechanisms focusing on specific regions[18], they often lack the mechanisms for fine-grained, multi-scale spatial reasoning or explicit enforcement of boundary adherence to clinical guidelines. The robust integration of complex anatomical knowledge—such as precise margin definitions relative to specific OARs or ensuring consistent delineation along key structural landmarks (e.g., the psoas muscles) as dictated by protocols—into end-to-end trainable networks remains an open research area. There is a pressing need for models that not only segment the CTV but also inherently understand and respect these critical anatomical constraints to produce clinically valid and reliable contours.

Deep learning approaches, particularly architectures based on U-Net and its variants, have demonstrated significant potential in medical image segmentation tasks. Key early architectures such as V-Net also paved the way for volumetric medical image segmentation[19]. However, CTV delineation for cervical cancer presents unique challenges compared to traditional organ segmentation. CTVs represent virtual volumes encompassing potential microscopic disease extensions, often with inherently ambiguous boundaries on standard imaging like CT, making them difficult to segment accurately[3]. Various research efforts have attempted to address these complexities through innovative architectural designs and methodological improvements.

Several studies have explored sophisticated network architectures specifically tailored for cervical cancer CTV segmentation. For instance, Ma, et al[20] developed a three-channel adaptive auto-segmentation

©Journal of Central South University (Medical Science). All rights reserved.

algorithm that leverages enhanced CT information to improve boundary delineation. To address the common challenge of limited annotated data, Peng, et al[21] introduced an innovative multi-decoder semi-supervised learning approach. The applicability of deep learning has also been demonstrated for magnetic resonance imaging (MRI) -based CTV segmentation, highlighting the potential for multi-modal imaging integration[3].

Advanced architectural innovations continue to emerge. The field has seen a progression toward transformer-based architectures and, more recently, state-space models. For example, Mamba-based frameworks have been investigated for rapid and precise high-risk CTV delineation in cervical cancer, showcasing the potential of these newer models[22]. These developments aim to better capture the complex spatial relationships inherent in CTVs.

The integration of anatomical prior knowledge is increasingly recognized as crucial for overcoming the limitations of purely visual pattern-based approaches. Khor, et al[23] developed anatomically constrained attention-guided networks that explicitly incorporate spatial relationships during training. Similarly, approaches like the prior-knowledge-guided U-Net (PK-UNet) aim to systematically integrate anatomical understanding into network architectures to guide the segmentation process, although its initial validation was in breast cancer[24].

Despite these promising developments, a fundamental challenge persists: CTVs are conceptually defined volumes whose boundaries depend heavily on anatomical interpretation and adherence to clinical guidelines rather than solely on direct visual cues in medical images. Current automated approaches can struggle with maintaining consistency and accuracy in regions where visual information is insufficient, potentially leading to clinically suboptimal contours. This underscores the continuing need for more sophisticated integration of anatomical knowledge and clinical reasoning into automated segmentation frameworks to ensure the generation of clinically valid and reliable CTV delineations.

Recognizing the limitations of purely image-driven segmentation, there has been a growing research emphasis on integrating anatomical prior knowledge into deep learning frameworks. This integration aims to address the inherent challenges of delineating structures that are guided more by anatomical understanding than by explicit image features[25]. Such anatomical intelligence can be incorporated at various stages, including data preprocessing, model architecture design, and post-processing refinement.

In the preprocessing stage, anatomical priors can inform the selection of regions of interest or enhance features critical for segmentation. For example, Zhang, et al[26] demonstrated a prior knowledge-guided deep learning approach where anatomical insights were used to construct improved input representations for segmenting complex abdominal structures, leading to better consistency.

During model design, several strategies allow for the direct embedding of anatomical constraints. Dalca, et al[27] utilized variational autoencoders to model expected anatomical configurations, guiding networks to maintain structural plausibility, particularly in ambiguous image regions. Multi-task learning presents another effective method; Boutillon, et al[28] showed that concurrently learning related anatomical segmentation tasks can enhance the primary segmentation by leveraging shared anatomical understanding. Similarly, Xu, et al[29] developed asymmetric multi-task attention network (AMTA-Net), an asymmetrical multi-task framework where a U-Net first segments foundational OARs (bladder and rectum). A cascaded attention mechanism then integrates multi-level OAR features and segmentation masks to guide prostate bed delineation, explicitly modeling the clinical reliance on OAR context for this challenging target volume. Furthermore, modifying loss functions to include anatomical constraints, such as those encouraging shape preservation, has shown considerable promise in maintaining the structural integrity of segmentations[30].

Post-processing techniques offer a final opportunity to enforce anatomical consistency. Mlynarski, et al[31] developed a method for CNN-based segmentation that applies topological constraints and shape regularization to ensure the final contours conform to expected anatomical properties. Hybrid approaches, which combine initial deep learning predictions with anatomically informed active contour models, have also been effective in refining boundary precision[32].

Despite these varied and advancing strategies for

©Journal of Central South University (Medical Science). All rights reserved.

incorporating anatomical priors, comprehensively modeling the complex, multi-scale anatomical relationships crucial for accurate CTV delineation remains a significant challenge. Effectively capturing the intricate spatial dependencies between the CTV and surrounding OARs, while simultaneously ensuring adherence to clinical margin definitions and consistency along key anatomical axes, is an ongoing pursuit[23]. These persisting limitations highlight the need for novel frameworks that can more deeply embed anatomical understanding directly within the network architecture to produce clinically robust and reliable CTV segmentations.

In summary, while deep learning, particularly U-Net architectures and their derivatives, has advanced automated CTV segmentation for cervical cancer, a critical gap remains. The inherent nature of CTV—defined by anatomical relationships and clinical guidelines rather than solely by image appearance—necessitates a more profound integration of anatomical intelligence than current standard approaches offer. Existing studies have highlighted various strategies for incorporating prior knowledge. However, challenges persist in comprehensively modeling multi-scale contextual interactions with OARs, ensuring consistency along key anatomical axes (such as psoas muscles), and explicitly enforcing adherence to clinical margin requirements and boundary sharpness.

To address these limitations, the hierarchical anatomical refinement network (HAR-Net) is proposed, a novel multi-stage framework designed for precise and clinically informed CTV delineation in cervical cancer. HAR-Net progressively refines the CTV segmentation by explicitly modeling and leveraging hierarchical anatomical relationships. It begins with a robust nnU-Net-based backbone for multi-scale 3D feature extraction. The core of this approach lies in 2 novel components: 1) Multi-scale OAR-CTV cross-attention (MOCCA) module, which dynamically incorporates OAR-specific contextual information at multiple spatial scales into the primary CTV segmentation pathway using windowed 3D cross-attention enhanced with learned relative positional biases. 2) Tri-axial coordinate attention (TCA) module further refines the segmentation by adaptively reweighting features based on learned tri-axial spatial priors, particularly relevant for consistent delineation along anatomical structures such as the psoas muscles.

1 Participants and Methods

1.1 Ethical statements

The study was approved by the Ethics Committee of the Xiangya Hospital, Central South University (Approval number: 2025060928) and was carried out according to the Declaration of Helsinki. According to the relevant guidelines and regulations of the retrospective study, the requirement of informed consent was waived.

1.2 Overall HAR-net framework: Integrating hierarchical anatomical knowledge

This study introduced the HAR-Net (Figure 1), a novel deep learning framework designed to learn a mapping H from an input 3D CT volume $I \in \mathbb{R}^{D \times H \times W}$ to a corresponding CTV segmentation mask $S_{\mathrm{CTV}} \in \{0, 1\}^{D \times H \times W}$, where D, H, W are the depth, height, and width of the volume. HAR-Net achieves this by progressively integrating hierarchical anatomical knowledge through a multi-stage refinement process. The core philosophy involves an initial stage of OAR-contextualization, followed by a refinement stage incorporating global structural priors. This hierarchical transformation is formulated as $S_{CTV} = \mathcal{H}(I) = f_{\mathrm{refine}}\left(f_{\mathrm{contextualize}}\left(f_{\mathrm{extract}}(I)\right)\right)$, where f_{extract} represents initial feature extraction, $f_{\mathrm{contextualize}}$ embodies OAR-CTV interaction modeling, and f_{refine} applies axis-aware refinements.

Mathematically, the end-to-end mapping of HAR-Net can be expressed in Equation (1) as:

$$S_{\mathrm{CTV}} = \mathrm{Seg}\left(\mathrm{TCA}\left(\mathrm{MOCCA}\left(f_{\mathrm{BB}}(I), f_{\mathrm{OAR}}(I)\right)\right)\right) \quad (1)$$

f_{BB}: NnU-Net-based backbone whose encoder-decoder structure provides the main CTV feature pathway. The MOCCA modules operate within its decoder stages. f_{OAR}: OAR feature pathway (sharing the encoder with f_{BB}, but with a dedicated OAR decoder) that provides stage-wise OAR contextual features to the MOCCA modules. The OAR decoder is trained to segment 5 pelvic OARs: Bladder, rectum, left femoral head, right femoral head, and small bowel. These organs were selected based on Radiation Therapy Oncology

©Journal of Central South University (Medical Science). All rights reserved.

Group (RTOG) guideline[33] as the primary dose-limiting structures with direct anatomical adjacency to the CTV—bladder anteriorly, rectum posteriorly, small bowel superiorly, and bilateral femoral heads laterally. MOCCA$\left(f_{\mathrm{BB}}(I),f_{\mathrm{OAR}}(I)\right)$: Output feature map from the final stage of the f_{BB} decoder, after OAR context from $f_{\mathrm{OAR}}(I)$ has been integrated at multiple scales by the MOCCA modules embedded within the f_{BB} decoder. TCA$(\cdot)$: TCA module refining the MOCCA-enhanced CTV features. Seg$(\cdot)$: Final segmentation head ($1\times1\times1$ convolution followed by sigmoid) producing the binary CTV mask.

HAR-Net implements this hierarchical strategy via a set of pivotal architectural modules. It begins with a robust nnU-Net-based backbone (responsible for f_{extract}) for rich multi-scale 3D feature extraction from I. This backbone employs an encoder-decoder architecture with skip connections, forming the basis of f_{BB} and the encoder part of f_{OAR}, following the principles of the nnU-Net framework[12], to effectively capture hierarchical features. Building upon this, the novel MOCCA module dynamically integrates OAR-specific contextual information from the f_{OAR} decoder into the f_{BB} decoder (a key part of $f_{\mathrm{contextualize}}$), producing OAR-aware features F_{MOCCA} (representing the output of the MOCCA-enhanced decoder). Following this, the TCA module (contributing to f_{refine}) further refines these features to yield F_{TCA}, by adaptively reweighting them based on learned tri-axial spatial priors, promoting consistency along critical anatomical axes. The final CTV segmentation S_{CTV} is derived from the refined features F_{TCA} after TCA processing.

The synergistic operation of these components facilitates the learning and application of anatomical intelligence at different levels of abstraction, from local OAR interactions to global structural rules.

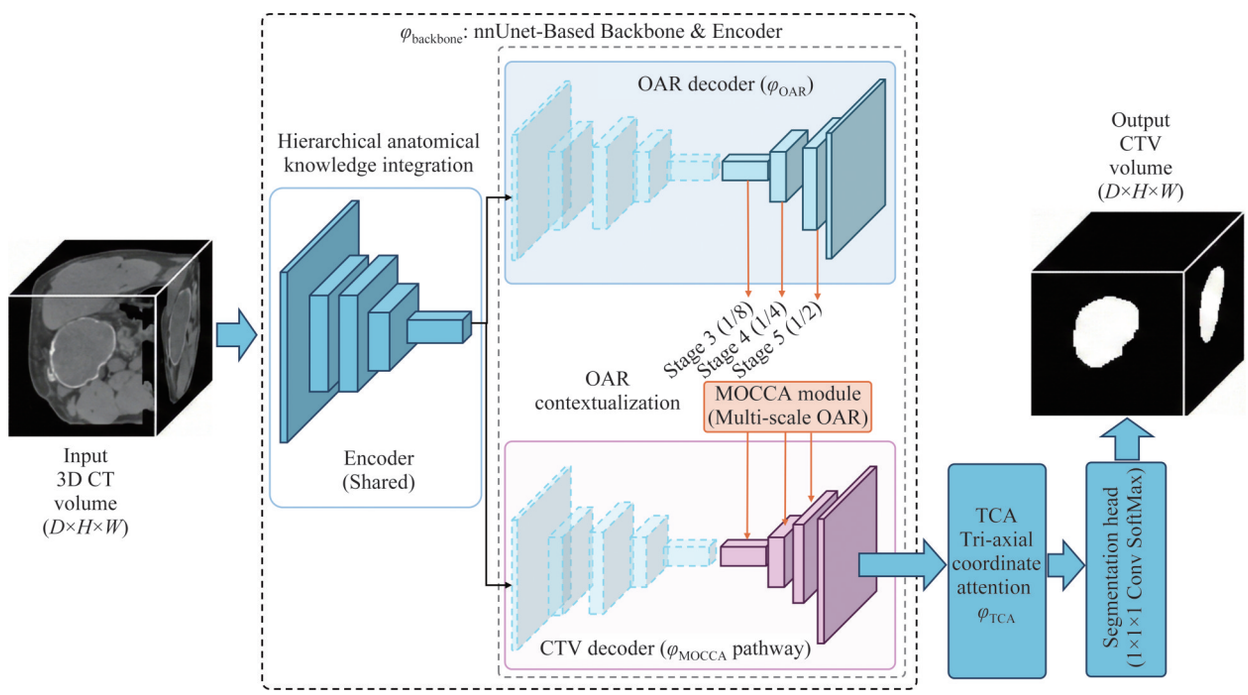


Figure 1 Proposed architecture

Framework processes a 3D input CT volume through a shared nnU-Net encoder $(\varphi_{\mathrm{backbone}})$ to extract multi-level pyramidal features. These features diverge into 2 parallel decoding pathways: An auxiliary OAR decoder (φ_{OAR}), which provides anatomical context, and the primary CTV decoder. The MOCCA module bridges these pathways at 3 hierarchical resolutions (e.g., 1/8, 1/4, 1/2), explicitly modeling the spatial relationship between OAR and the CTV via windowed 3D cross-attention. Finally, the aggregated features undergo global structural refinement via the TCA module (φ_{TCA}) to enforce axis-aware anatomical consistency before the final segmentation head generates the CTV probability map. CT: Computed tomography; CTV: Clinical target volume; OAR: Organs at risk; MOCCA: Multi-scale OAR-CTV cross-attention; TCA: Tri-axial coordinate attention.

©Journal of Central South University (Medical Science). All rights reserved.

1.3 MOCCA module: MOCCA with learned positional biases

CTV delineation is intrinsically linked to the localization and morphology of adjacent OARs. To model these dependencies, the MOCCA module is introduced, a component of HAR-Net (Figure 2). MOCCA dynamically integrates OAR-specific contextual information into the CTV segmentation pathway at multiple spatial scales within HAR-Net's decoder arm. This ensures learned relationships are sensitive to varying levels of feature abstraction.

Let $F_{\mathrm{CTV}}^{(s)} \in \mathrm{R}^{C_s \times D_s \times H_s \times W_s}$ denote the feature maps from the CTV decoding pathway at a given scale s, where C_s, D_s, H_s, W_s are the channels, depth, height, and width, respectively. Similarly, $F_{\mathrm{OAR}}^{(s)}$ represents corresponding feature map from a parallel OAR-focused decoding stream, providing salient information regarding the OARs at the same scale. The MOCCA module takes these 2 sets of feature maps as input at each relevant decoder stage. The OAR features $F_{\mathrm{OAR}}^{(s)}$ are obtained from a dedicated OAR decoder that shares the same encoder architecture as the main CTV pathway but is specifically trained to segment relevant OARs.

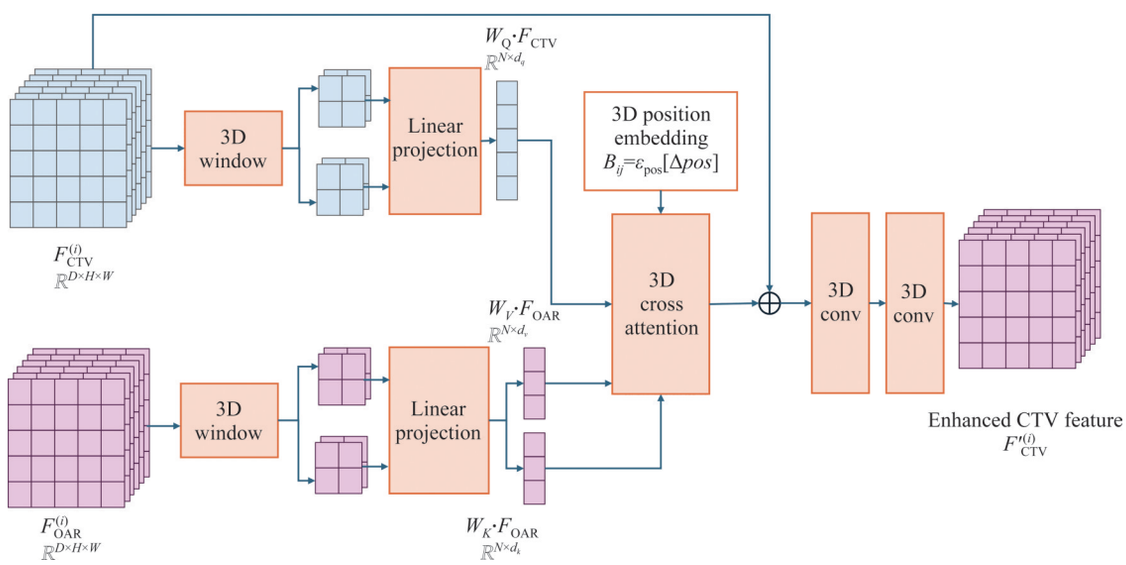


Figure 2 Detailed architecture of the MOCCA module

Input CTV features ($F_{CTV}^{(s)}$) and OAR features ($F_{OAR}^{(s)}$) undergo 3D windowing. Queries (Q) are derived from $F_{CTV}^{(s)}$ via linear projection. Keys (K) and Values (V) are similarly derived from $F_{OAR}^{(s)}$. Windowed 3D cross-attention is performed, incorporating learned 3D relative positional biases (B_{ij}). The attention output is added to the original $F_{CTV}^{(s)}$ features via a residual connection, followed by 2 3D convolutional layers to produce enhanced CTV features; conv: Convolutional layer.

The core mechanism of the MOCCA module is a windowed 3D cross-attention operation. Input feature maps $F_{\mathrm{CTV}}^{(s)}$ and $F_{\mathrm{OAR}}^{(s)}$, which are projected to the same channel dimension C_s', are partitioned into non-overlapping 3D windows of size $8 \times 8 \times 8$ voxels. Within each window, feature vectors are extracted. Specifically, queries ($Q \in \mathrm{R}^{N \times d_q}$) are derived via a linear projection from the windowed $F_{\mathrm{CTV}}^{(s)}$ tokens, while keys ($K \in \mathrm{R}^{M \times d_k}$) and values ($V \in \mathrm{R}^{M \times d_v}$) are derived via separate linear projections from the windowed $F_{\mathrm{OAR}}^{(s)}$ tokens. Here, $N = M = 8^3 = 512$ represents the number of tokens (voxels) per window. The query, key, and value dimensions (d_q, d_k, d_v) are set to C_s'. A single attention head is employed to balance model complexity and efficacy for these localized interactions. This localized attention strategy significantly reduces the quadratic complexity associated with global self-attention on large 3D volumes, making it computationally tractable while concentrating the learning on inter-structural relationships to spatially pertinent sub-volumes where direct OAR-CTV influence is most anatomically relevant. The window

©Journal of Central South University (Medical Science). All rights reserved.

size of 8 × 8 × 8 voxels was chosen based on anatomical scale matching and computational efficiency: Given a typical voxel spacing of approximately 0.98 mm × 0.98 mm × 3.00 mm, this window corresponds to a physical extent of roughly 7.8 mm × 7.8 mm × 24.0 mm, covering the critical OAR-CTV interface range (typically 10 − 30 mm). Smaller windows (e. g., 4 × 4 × 4, n=64) provide insufficient receptive fields, while larger windows (e. g., 16 × 16 × 16, n=4 096) incur approximately 64 × the computational cost with diminishing returns due to attention dilution. The attention output is computed using a formulation based on scaled dot-product attention[34] in Equation (2) as:

$$\text{Attention}\left(Q, K, V\right) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V \quad (2)$$

where B represents the learned 3D relative positional biases, and attention dropout with a probability of 0.1 is applied to the attention weights (output of Softmax) before multiplication with V for regularization.

A key aspect of the MOCCA module is the incorporation of learned 3D relative positional biases, B. These biases are added element-wise to the raw attention scores (logits) $QK^T/\sqrt{d_k}$ prior to Softmax normalization. The parameterization of B is a distinguishing feature: For each query-key pair i,j within an attention window, the bias value B_{ij} is retrieved from a learnable embedding table E_{pos}, which is initialized with zeros. If $\text{pos}_i = \left(d_i, \mathrm{h}_i, w_i\right)$ and $\text{pos}_j = \left(d_j, \mathrm{h}_j, w_j\right)$ denote the 3D coordinates of token i and token j respectively, and D_r, H_r, W_r are the window dimensions along depth, height, and width (set to $D_r = H_r = W_r = 8$ in this implementation), then B_{ij} is computed in Equation (3) as:

$$B_{ij} = \varepsilon_{\text{pos}}\left[\text{pos}_i - \text{pos}_j + \left(D_r - 1, H_r - 1, W_r - 1\right)\right] \quad (3)$$

The embedding table ε_{pos} has dimensions $\left(2D_r - 1\right) \times \left(2H_r - 1\right) \times \left(2W_r - 1\right)$ to accommodate all relative displacements. The indices, such as $d_i - d_j + D_r - 1$, map the relative displacements [which range from $-\left(D_r - 1\right)$ to $(D_r - 1)$] to non-negative values for table lookup. This mechanism enables the module to dynamically learn and encode anisotropic spatial configurations and specific margin relationships between OARs and CTV subregions directly from the training data.

The output of this attention mechanism—a set of OAR-contextualized value vectors—is then integrated with the original CTV features $F_{\text{CTV}}^{(s)}$ (from which Q was derived). This fusion is realized through a residual connection, which adds the attention output to $F_{\text{CTV}}^{(s)}$. This fused representation subsequently passes through a residual block comprising 2 sequential units of [3D Convolution (kernel 3 × 3 × 3, stride 1, same padding) → Instance Normalization → ReLU activation]. These subsequent convolutions further process and refine the integrated features, producing an enriched representation $F'^{(s)}_{\text{CTV}}$ that robustly combines the primary image features with the learned OAR-CTV spatial priors. The MOCCA module is implemented at 3 specific decoder stages, corresponding to resolution scales of 1/8, 1/4, and 1/2 of the input volume, namely decoder Stage 3 (1/8 resolution, 256 channels), decoder Stage 4 (1/4 resolution, 128 channels), and decoder Stage 5 (1/2 resolution, 64 channels), establishing a hierarchical refinement pattern that progressively integrates anatomical context from coarse to fine resolutions. Stage 3, with the largest receptive field, captures global spatial relationships between CTV and distant OARs (e.g., femoral heads). Stage 4 models mid-scale organ distance and orientation constraints (e. g., CTV relative to bladder and rectum). Stage 5, at near-original resolution, refines fine-grained boundary interactions. The earlier decoder stages (Stage 0–2) operate at 1/32 to 1/16 resolution, where excessive spatial compression renders precise spatial relationship modeling impractical. This multi-scale implementation facilitates the development of a resolution-dependent understanding of OAR-CTV spatial interdependencies, addressing anatomical constraints that define CTV boundaries. Encoding these structural relationships across multiple scales is designed to improve boundary precision and anatomical fidelity in the final CTV delineation.

Regarding how MOCCA learns clinically relevant spatial relationships, the cross-attention weights $softmax\left(QK^T/\sqrt{d_k} + B\right)$ encode the degree to which each CTV voxel attends to OAR voxels. During training, these weights are optimized via backpropagation and converge toward high attention at

©Journal of Central South University (Medical Science). All rights reserved.

OAR-CTV interfaces and low attention in distant regions. The 3D relative positional bias table ε_{pos} further encodes directional attention preferences—for instance, different biases along the anterior-posterior axis (bladder-CTV interface) versus the superior-inferior axis (small bowel-CTV interface), reflecting the model's learned directional anatomical constraints.

1.4 TCA module: TCA for global structural refinement

While the MOCCA module effectively captures local OAR-CTV contextual information, precise CTV delineation often necessitates adherence to global anatomical rules and consistent orientations along the principal body axes (e. g., superior-inferior, anterior-posterior, left-right). For instance, CTV boundaries may need to align with elongated structures like the psoas muscles or follow specific cranio-caudal extents. Standard convolutional layers, due to their inherently local receptive fields, often do not effectively capture these long-range, axis-aligned dependencies.

To address this, the TCA module is introduced, a 3D adaptation of the Coordinate Attention mechanism[35]. As illustrated in Figure 3, the TCA module is designed to explicitly model long-range dependencies and encode precise positional information along each spatial axis. It processes the feature maps that have been contextualized by OAR interactions through the preceding MOCCA modules within the decoder pathway. This allows the network to selectively emphasize informative features based on global, axis-aware anatomical context. The TCA module operates in 3 main steps: Axis-wise feature encoding, tri-axial attention generation, and feature re-weighting.

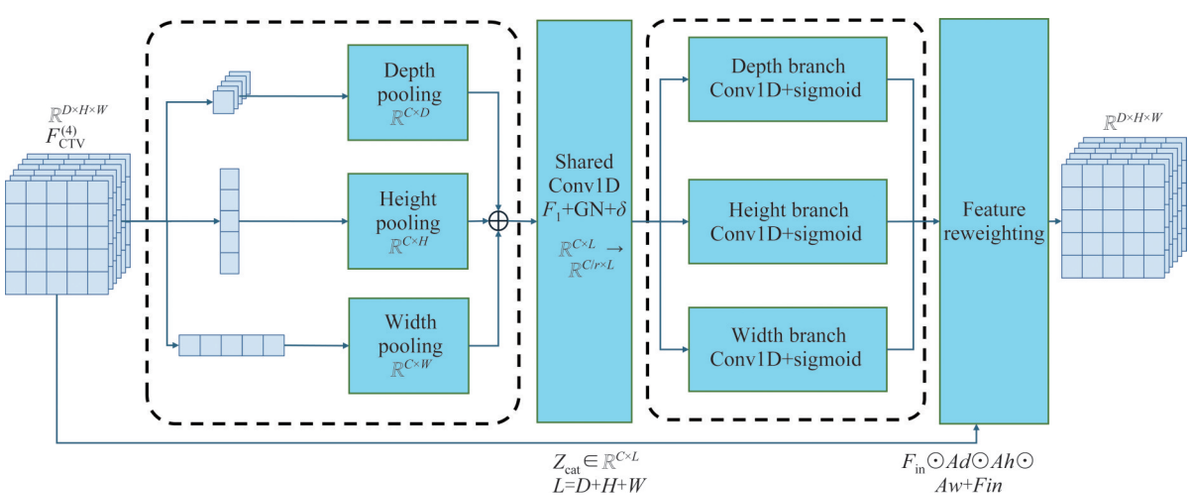


Figure 3 Architecture of the TCA module

Input feature map $F^{(4)}_{CTV}$, denoted as F_{in}, undergoes axis-wise average pooling (depth-wise, height-wise, width-wise) to produce 1D feature vectors. These are concatenated (Z_{cat}) and processed by a shared 1D convolution and non-linearity. The result is split and fed into axis-specific 1D convolutions followed by sigmoid activation to generate attention weights (Ad, Ah, Aw). Finally, the input features F_{in} are reweighted by these attention maps and combined with a residual connection. GN: Group normalization.

1.4.1 Axis-wise feature encoding

Given an input feature map $F_{in}\in \mathrm{R}^{C\times D\times H\times W}$, where C is the number of channels, and D, H, W are the depth, height, and width, respectively, the TCA module first captures global information along each axis. This is achieved by applying average pooling along the other 2 axes for each channel independently. Let P_a be an average-pooling operator that pools over all spatial axes except $a\in\{d, h, w\}$, where d, h, w correspond to depth, height, and width axes, respectively. For an input feature map $F_{in}\in \mathrm{R}^{C\times D\times H\times W}$, the pooled feature Z_c^a for channel c along axis a is given by Equations (4) to (6) as:

$$Z_c^d(i)=\frac{1}{HW}\sum_{j=0}^{H-1}\sum_{k=0}^{W-1}F_{in}(c,i,j,k) \tag{4}$$

$$Z_c^h(j)=\frac{1}{DW}\sum_{i=0}^{D-1}\sum_{k=0}^{W-1}F_{in}(c,i,j,k) \tag{5}$$

©Journal of Central South University (Medical Science). All rights reserved.

$$Z_c^w(k)=\frac{1}{DH}\sum_{i=0}^{D-1}\sum_{j=0}^{H-1}F_{in}(c,i,j,k) \tag{6}$$

where $i\in\{0,\cdots,D-1\}$; $j\in\{0,\cdots,H-1\}$; $k\in\{0,\cdots,W-1\}$ are the dimensions along depth, height, and width, and k_a is the coordinate along axis a, while k_b for $b\neq a$ are summed over. This generates 3 1D feature tensors: $Z^d\in\mathrm{R}^{C\times D}$, $Z^{\mathrm{h}}\in\mathrm{R}^{C\times H}$, and $Z^w\in\mathrm{R}^{C\times W}$. These operations embed the global receptive field from 2 dimensions into a single dimension, capturing long-range dependencies while retaining precise positional information for each respective axis.

1.4.2 Tri-axial attention generation

The encoded 1D feature tensors are concatenated along their spatial dimensions to form $Z_{cat}\in\mathrm{R}^{C\times L}$, where $L=D+H+W$. This tensor passes through a shared 1D convolution (kernel size=1), F_1: $\mathrm{R}^{C\times L}\rightarrow\mathrm{R}^{C/r\times L}$, with a reduction ratio $r=16$, followed by Group Normalization (2 groups, i. e. $\min(32,C/r)$ with $C=32$, $r=16$) and a non-linear activation δ (implemented as Hard-Swish, chosen because its smooth gradient transition near zero avoids the "dying ReLU" problem and facilitates fine-grained attention weight learning, while being computationally cheaper than GELU due to its piece-wise-linear form) in Equation (7) as:

$$f=\delta\left(\mathrm{GN}\left(F_1\left(Z_{cat}\right)\right)\right)\in\mathrm{R}^{C/r\times L}. \tag{7}$$

The tensor f is then split into 3 axis-specific tensors: $f^d\in\mathrm{R}^{C/r\times D}$, $f^{\mathrm{h}}\in\mathrm{R}^{C/r\times H}$, and $f^w\in\mathrm{R}^{C/r\times W}$. Each is processed by its own 1D convolution (kernel size=1), F_d,F_{h},F_w: $\mathrm{R}^{C/r\times *}\rightarrow\mathrm{R}^{C\times *}$, producing raw attention logits in Equation (8) as:

$$A_{raw}^d=F_d\left(f^d\right),\quad A_{raw}^{\mathrm{h}}=F_{\mathrm{h}}\left(f^{\mathrm{h}}\right),\quad A_{raw}^w=F_w\left(f^w\right) \tag{8}$$

1.4.3 Feature reweighting

The raw logits are normalized with a sigmoid $\sigma(\cdot)$, yielding attention maps A^d,A^{h},A^w. These 1D maps are reshaped for broadcasting: $A^d\in\mathrm{R}^{C\times D\times 1\times 1}$, $A^{\mathrm{h}}\in\mathrm{R}^{C\times 1\times H\times 1}$, $A^w\in\mathrm{R}^{C\times 1\times 1\times W}$. The final output F_{out} is then computed as follows, incorporating a residual skip for stability as in Equation (9):

$$F_{out}=F_{in}\odot A^d\odot A^{\mathrm{h}}\odot A^w+F_{in} \tag{9}$$

TCA adds only $O\left(C^2/r+(D+H+W)C\right)$ parameters and FLOPs, negligible compared to the backbone.

1.5 Loss function and training strategy

1.5.1 Loss Function

The network parameters are optimized by minimizing a composite loss function, L_{total}. The primary component of this loss, L_{CTV}, focuses on the final CTV segmentation output and combines a region-based loss and a voxel-wise loss for effective segmentation supervision, defined in Equation (10) as:

$$L_{\mathrm{CTV}}=L_{\mathrm{Dice}}\left(S_{\mathrm{CTV}},S_{\mathrm{GT}}\right)+\lambda_1 L_{\mathrm{BCE}}\left(S_{\mathrm{CTV}},S_{\mathrm{GT}}\right) \tag{10}$$

where S_{GT} is the ground truth CTV mask, L_{Dice} is the Dice loss (popularized by works like[19]), promoting overlap between prediction and ground truth, and L_{BCE} is the Binary Cross-Entropy loss, providing voxel-level supervision. λ_1 is a weighting factor to balance these 2 components. An auxiliary OAR segmentation head is also utilized. Its loss, $L_{\mathrm{OAR_aux}}$ (also computed as a combination of Dice and binary cross entropy loss for each relevant OAR), is incorporated to help in learning more discriminative OAR features for the MOCCA module. The overall loss L_{total} is then defined in Equation (11) as:

$$L_{total}=L_{\mathrm{CTV}}+\lambda_2 L_{OAR_\mathrm{aux}} \tag{11}$$

where $\lambda_2=0.5$ is a weighting factor that balances the contribution of the auxiliary OAR loss.

1.5.2 Training strategy

The HAR-Net framework, including the nnU-Net-based backbone, the MOCCA modules, the TCA module, and the final segmentation head, is trained in an end-to-end fashion. The training objective is to learn the complete mapping from an input 3D CT volume I to the final CTV segmentation mask S_{CTV}, by optimizing all network parameters simultaneously.

During the forward pass, an input CT volume I is processed through the following stages.

1) Feature extraction: I is fed into the nnU-Net-based backbone (f_{BB}), which extracts multi-scale hierarchical features through its encoder-decoder structure.

2) OAR context integration via MOCCA: At specific decoder stages (resolutions 1/8, 1/4, and 1/2 of the input volume, as detailed in Section 1.3), the CTV pathway features from f_{BB} are combined with corresponding OAR-specific features (derived from f_{OAR}). The MOCCA module at each of these scales takes the CTV and OAR features as input, performing

©Journal of Central South University (Medical Science). All rights reserved.

windowed 3D cross-attention with learned relative positional biases. This yields OAR-contextualized CTV features $F'^{(s)}_{\mathrm{CTV}}$ at each scale s. The output of a MOCCA module at a coarser scale is upsampled and concatenated with skip-connection features from the encoder before potentially being processed by a MOCCA module at a finer scale.

3) Global axial refinement via TCA: The feature map resulting from the deepest (or final) MOCCA-enhanced decoder stage, denoted as $F_{\mathrm{MOCCA_out}}$, is then passed to the TCA module. As described in Section 1.4, the TCA module applies axis-wise feature encoding, generates tri-axial attention weights, and reweights $F_{\mathrm{MOCCA_out}}$ to produce globally refined features F_{TCA}. This step allows the network to emphasize features consistent with learned long-range, axis-aligned anatomical priors.

4) Final segmentation output: The refined features F_{TCA} from the TCA module are channeled into a final segmentation head, which comprises a $1 \times 1 \times 1$ convolution followed by a sigmoid activation function for binary CTV segmentation. This produces a probability map, subsequently thresholded at 0.5 during inference to yield the final binary CTV segmentation mask S_{CTV}.

1.6 Participants and experimental setup

1.6.1 Participants and datasets

The primary dataset used in this study was retrospectively collected from 132 patients diagnosed with cervical cancer undergoing radiotherapy at Xiangya Hospital of Central South University between May 2022 and May 2025.

Inclusion criteria: 1) Histologically confirmed cervical cancer (squamous cell carcinoma, adenocarcinoma, or adenosquamous carcinoma); 2) International Federation of Gynecology and Obstetrics (FIGO) 2018 stage IB2−IVA[36]; 3) planned for definitive or adjuvant external beam radiotherapy (EBRT); 4) available planning CT with standard acquisition protocol; 5) complete CTV and OAR delineations by senior radiation oncologists. Exclusion criteria: 1) Prior pelvic radiation therapy; 2) significant motion or metallic artifacts on CT; 3) incomplete or missing contour data; 4) concurrent non-cervical pelvic malignancy.

CT images were acquired using a SOMATOM Definition AS scanner (Siemens Healthineers, Germany) with a tube voltage of 120 kV and auto-modulated tube current (120 mAs). Slice thickness was 3 mm, and images were reconstructed with a voxel spacing of approximately 0.98 mm × 0.98 mm × 3.00 mm. Scans covered the region from the upper border of L2 vertebra to the inferior border of the ischial tuberosity.

CTVs and 5 relevant OARs (bladder, rectum, left femoral head, right femoral head, and small bowel) were manually delineated following RTOG consensus guideline[33] on planning CT scans. All contours were independently delineated by 2 radiation oncologists with more than 5 years of pelvic radiotherapy experience; discrepancies were resolved by a senior physician with over 10 years of experience.

The dataset was randomly split into training (92 cases), validation (15 cases), and test (25 cases) sets with no patient overlap, stratified by FIGO stage to ensure consistent stage distribution across subsets.

1.6.2 Evaluation metrics

Quantitative assessment of HAR-Net and the comparative methods was performed using 3 widely adopted segmentation metrics.

1) Dice similarity coefficient (DSC): Measures the overlap between the predicted segmentation S_P and the ground truth segmentation S_{GT}. It is defined in Equation (12) as:

$$\mathrm{DSC}\left(S_P, S_{GT}\right) = \frac{2 \times \left|S_P \cap S_{GT}\right|}{\left|S_P\right| + \left|S_{GT}\right|} \tag{12}$$

where $|\cdot|$ denotes the number of voxels in a segmentation. DSC ranges from 0 (no overlap) to 1 (perfect overlap).

2) 95th percentile Hausdorff distance (HD95): Measures the maximum distance between the surfaces of S_P and S_{GT}, but is robust to outliers by considering the 95th percentile of the distances. Let ∂S_P and ∂S_{GT} be the sets of surface points of the predicted and ground truth segmentations, respectively. The directed Hausdorff distance is $\mathrm{h}\left(A, B\right) = \max_{a \in A} \min_{b \in B} /\!/ a - b /\!/ _2$. HD95 is defined as the 95th percentile of the values in $\min_{b \in \partial S_{GT}} /\!/ a - b /\!/ _2 \forall\, a \in \partial S_P \cup \min_{a \in \partial S_P} /\!/ b - a /\!/ _2 \forall\, b \in \partial S_{GT}$. Lower HD95 values (typically in mm) indicate better agreement between the predicted and ground truth surfaces.

©Journal of Central South University (Medical Science). All rights reserved.

3) Average symmetric surface distance (ASSD): Calculates the average distance between the surfaces of S_P and S_{GT}. It is defined in Equation (13) as:

$$ASSD\left(S_P, S_{GT}\right) = \frac{1}{\left|\partial S_P\right| + \left|\partial S_{GT}\right|} \times \left(\sum_{p \in \partial S_P} d\left(p, \partial S_{GT}\right) + \sum_{q \in \partial S_{GT}} d\left(q, \partial S_P\right) \right) \tag{13}$$

where $d\left(p, \partial S\right) = \min_{s \in \partial S} \left\| p - s \right\|_2$. Lower ASSD values (typically in mm) indicate a closer match between the segmentation boundaries.

1.6.3 Implementation details

The HAR-Net framework was implemented using PyTorch (Version 2.6.0). Training was conducted on NVIDIA GeForce RTX 3090 Ti GPUs, each with 24 GB of VRAM. Network parameters were optimized using the AdamW optimizer, initialized with a learning rate of 2×10^{-4} and a weight decay of 5×10^{-4}. A multi-step learning rate schedule was employed to adjust the learning rate during training.

To enhance model robustness and mitigate overfitting, extensive online data augmentation was applied, encompassing random horizontal and vertical flipping, random rotation, random contrast adjustments, and random translations and scaling. Prior to network input, CT volumes underwent pre-processing, which involved clipping Hounsfield unit (HU) values to the relevant range for soft-tissue visualization, followed by normalization. The network was trained using randomly extracted 3D patches, with patch dimensions (64 × 128 × 128) determined by nnU-Net heuristics based on dataset characteristics, and a batch size of 2 was used. The training typically extended for 1 000 epochs for the initial stage. The optimization process was guided by the composite loss function L_{total} [Equation (11)], which combines the primary CTV segmentation loss L_{CTV} [Equation (10)] with an auxiliary loss $L_{\text{OAR_aux}}$ derived from OAR segmentations. The loss weights were set to $\lambda_1 = 1.0$ and $\lambda_2 = 0.5$. The learning rate followed a multi-step schedule with decay at epochs 500 and 800 ($\gamma = 0.1$). HU values were clipped to [−175, 250] for soft tissue visualization.

1.6.4 Compared methods (baselines)

To contextualize the performance of HAR-Net, it was benchmarked against several state-of-the-art deep learning architectures prevalent in medical image segmentation. These included:

1) MultiResUNet[13]: A multi-resolution U-Net architecture for biomedical image segmentation (adapted to the present 3D CTV task).

2) NnU-Net[12]: A self-configuring framework that automatically adapts U-Net-based architectures to various medical segmentation tasks, often serving as a strong baseline. This forms the backbone of our HAR-Net.

3) AMTA-Net[29]: An asymmetrical multi-task framework leveraging OAR context for challenging target volume segmentation.

4) Swin UNETR[15]: A Transformer-based model that uses a Transformer encoder with a U-Net-like decoder structure for 3D medical image segmentation.

5) ResUNet-a[16]: A 3D CNN architecture for enhanced feature learning.

For all baseline methods, we utilized their publicly available implementations where possible and trained them on our dataset using similar data pre-processing and augmentation strategies as HAR-Net to ensure a fair comparison. Hyper-parameters were tuned according to their respective papers or default configurations if specific tuning guidelines were unavailable for our dataset characteristics. The nnU-Net baseline corresponds to the output of the standard nnU-Net framework without our proposed MOCCA and TCA modules.

2 Results

2.1 Comparison with state-of-the-art methods

HAR-Net was benchmarked against several leading deep learning models for medical image segmentation: MultiResUNet, nnU-Net, AMTA-Net, Swin UNETR, and ResUNet-a. Quantitative comparison results are presented in Table 1. HAR-Net achieved the highest performance across all 3 metrics. Specifically, HAR-Net yielded a DSC of (82.420±5.420)%, an HD95 of (8.390±6.120) mm, and an ASSD of (1.920±0.980) mm. Compared to the strong nnU-Net baseline [DSC: (79.160±6.100)%; HD95: (12.470±7.380) mm; ASSD: (2.570±1.210) mm], HAR-Net demonstrated notable

©Journal of Central South University (Medical Science). All rights reserved.

improvements, including an increase in DSC by 3.26 percentage points and reductions in HD95 and ASSD by 4.08 mm and 0.65 mm, respectively.

Figure 4 further supplements the state-of-the-art comparison with visualizations illustrating the distribution of performance metrics. For DSC (Figure 4A), HAR-Net not only exhibits the highest median but also a competitive interquartile range (IQR), indicating consistent high performance. While some outliers are present, the bulk of HAR-Net's DSC results are concentrated at the higher end compared to other methods. In terms of surface distance metrics (Figure 4B for HD95 and Figure 4C for ASSD), HAR-Net consistently demonstrates lower medians and more compact IQRs, suggesting more accurate and reliable boundary delineation. Specifically, for HD95, HAR-Net's distribution is visibly shifted towards lower values, indicating fewer instances of large boundary errors relative to the other state-of-the-art approaches.

The quantitative improvements translate to clinically relevant enhancements in segmentation quality, with HAR-Net showing superior boundary delineation and reduced inter-patient variability compared to existing methods. Figure 5 provides visual comparisons between HAR-Net and the state-of-the-art methods across different anatomical views (axial, sagittal, coronal) and a 3D rendering. The figure demonstrates HAR-Net's superior ability to capture anatomically consistent CTV boundaries compared to other methods, with noticeably improved delineation quality and reduced boundary irregularities.

Table 1 Performance comparison of various models

Method	DSC/% ↑	HD95/mm ↓	ASSD/mm ↓	DSC *P*
MultiResUNet	68.650±9.730	13.750±8.790	3.260±1.630	<0.001
AMTA-Net	77.270±5.350	13.130±8.150	2.890±1.220	<0.001
Swin UNETR	76.860±6.340	12.490±7.360	2.400±1.270	<0.001
ResUNet-a	75.620±8.280	12.310±7.130	2.730±1.550	<0.001
nnU-Net	79.160±6.100	12.470±7.380	2.570±1.210	0.003
HAR-Net	82.420±5.420	8.390±6.120	1.920±0.980	

Data are presented as mean±standard deviation. Wilcoxon signed-rank tests were used on per-case metrics, comparing each baseline against HAR-Net. Arrows indicate the direction of better performance. DSC: Dice similarity coefficient; HD95: 95th Percentile Hausdorff distance; ASSD: Average symmetric surface distance.

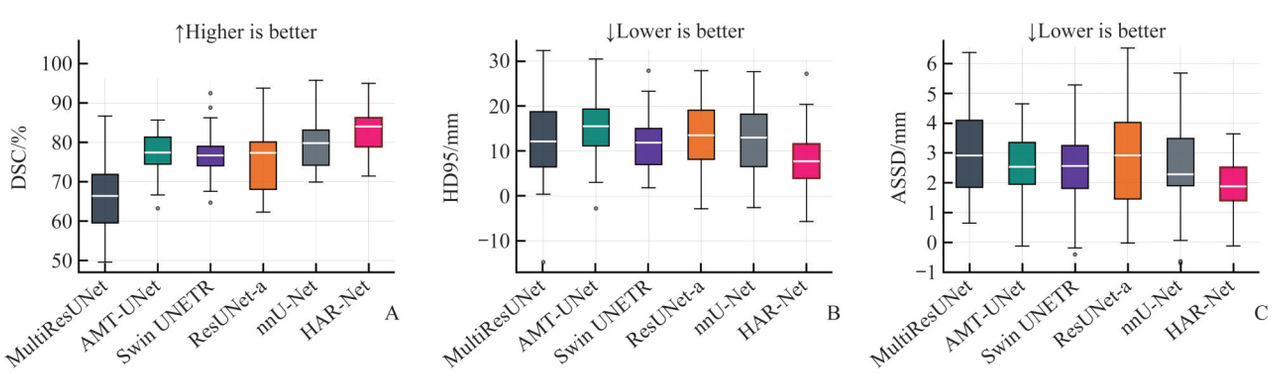


Figure 4 Comparison with state-of-the-art methods

A−C: Box plots comparing the distribution of DSC (A), HD95 (B), and ASSD (C) for HAR-Net and state-of-the-art methods. The central mark indicates the median, the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. Whiskers extend to 1.5× interquartile range; outliers are shown as individual points. Statistical comparisons of the differences between each method and HAR-Net were performed using the Wilcoxon signed-rank test. **P*<0.05. DSC: Dice similarity coefficient; HD95: 95th Percentile Hausdorff distance; ASSD: Average symmetric surface distance.

©Journal of Central South University (Medical Science). All rights reserved.

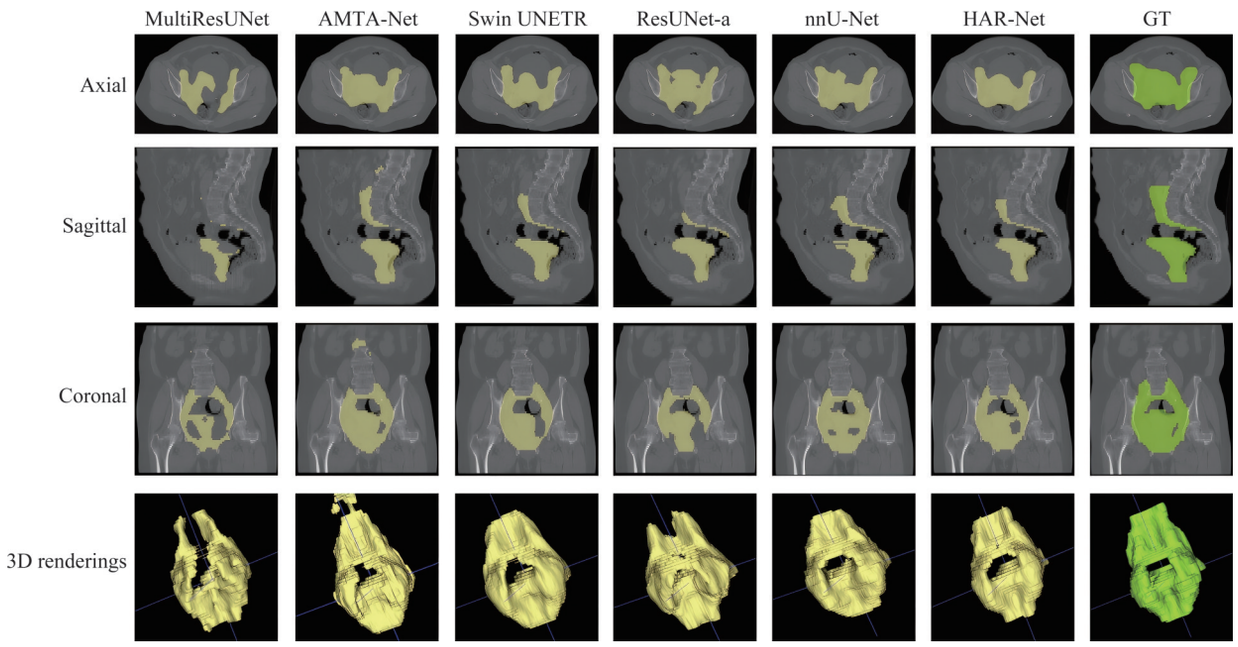


Figure 5 Visual comparison of CTV segmentation results across different methods

GT: Ground truth.

2.2 Ablation experiment

To assess the individual contributions of the MOCCA and TCA modules, ablation experiments were performed. The configurations evaluated were: 1) Baseline (nnU-Net backbone), 2) Baseline+MOCCA, 3) Baseline+TCA, and 4) HAR-Net (Baseline+MOCCA+TCA).

The results (Table 2 and Figure 6) illustrated the impact of each component. The baseline nnU-Net established the reference performance. Integrating the MOCCA module improved the DSC and ASSD, though HD95 slightly increased, highlighting the general benefit of OAR-CTV contextualization. The addition of the TCA module alone to the baseline yielded improved surface distance metrics. The complete HAR-Net model, incorporating both MOCCA and TCA modules, achieved the best overall performance. This demonstrated a synergistic effect, where the combination of MOCCA and tri-axial coordinate attention led to substantial improvements in segmentation accuracy. We further examined the window size of the MOCCA module (Table 3) and the activation function in TCA (Table 4). The intermediate window size configuration yielded the highest DSC and the lowest surface distances, and was therefore adopted. Among the tested activation functions, Hard-Swish consistently outperformed ReLU and GELU across all metrics (Table 4), justifying its selection in the final model.

Table 2 Ablation experiment results for HAR-Net components ($\bar{x}\pm s$)

Configuration	DSC/% ↑	HD95/mm ↓	ASSD/mm ↓
Baseline	77.920±6.280	13.290±6.920	2.470±1.030
+MOCCA	79.820±6.170	14.140±7.920	2.530±1.160
+TCA	78.470±6.070	13.950±6.230	2.420±1.170
+MOCCA+TCA*	82.420±5.420	8.390±6.120	1.920±0.980

*Best model. Progressive improvements shown from baseline to full model. Arrows indicate the direction of better performance.

2.3 OAR branch performance analysis

To validate the quality of OAR features utilized by the MOCCA module, the OAR decoding branch was evaluated on the test set (Table 5).

©Journal of Central South University (Medical Science). All rights reserved.

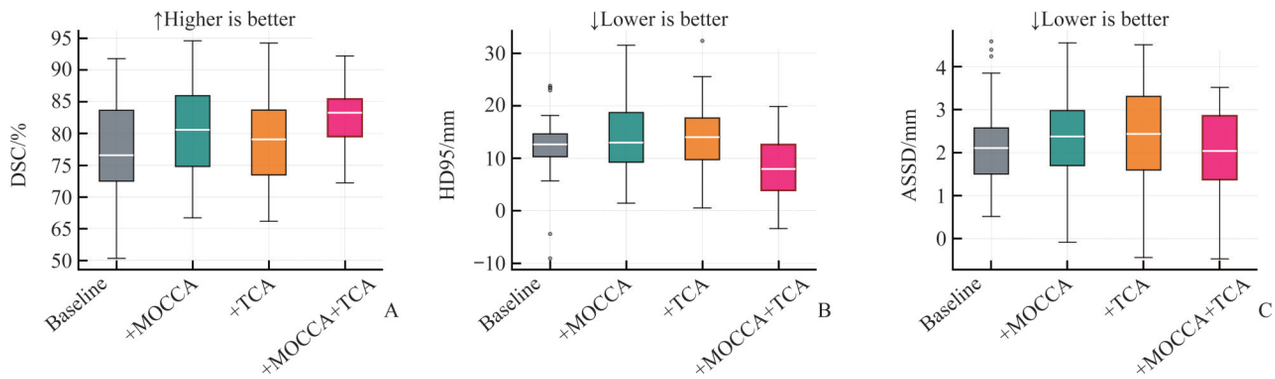


Figure 6 Box plots for ablation experiment

Distribution of DSC (A), HD95 (B), and ASSD (C) for different HAR-Net configurations. This illustrates the progressive performance gains from the MOCCA and TCA modules.

Table 3 Ablation experiment on MOCCA window size ($\bar{x}\pm s$)

Window Size	DSC/% ↑	HD95/mm ↓	ASSD/mm ↓
4 × 4 × 4	80.870±5.780	9.850±6.540	2.140±1.050
8 × 8 × 8*	82.420±5.420	8.390±6.120	1.920±0.980
16 × 16 × 16	81.930±5.610	8.720±6.280	2.030±1.020

*Chosen window size for HAR-Net. Arrows indicate the direction of better performance.

Table 4 Ablation experiment on TCA activation function ($\bar{x}\pm s$)

Activation	DSC/% ↑	HD95/mm ↓	ASSD/mm ↓
ReLU	81.560±5.630	9.070±6.380	2.080±1.030
GELU	82.030±5.550	8.850±6.240	2.010±1.000
Hard-Swish*	82.420±5.420	8.390±6.120	1.920±0.980

*Activation function used in HAR-Net. Arrows indicate the direction of better performance.

The OAR branch achieved an overall mean DSC of 90.410%, with performance consistent with organ-specific anatomical characteristics: Femoral heads, with clear boundaries, achieved the highest DSC (>94%), while small bowel, with complex morphology and positional variability, was the most challenging (DSC 81.360%). To quantify the relationship between OAR segmentation quality and CTV accuracy, we computed the Pearson correlation between per-case mean OAR DSC and CTV DSC: r=0.470 (P=0.018), indicating a moderate positive correlation that supports the premise that accurate OAR features contribute to better CTV delineation (Figure 7).

Table 5 OAR decoding branch segmentation performance on test set ($\bar{x}\pm s$, n=25)

OAR	DSC/%	HD95/mm
Bladder	93.470±2.150	3.820±1.430
Rectum	87.230±4.610	6.150±2.870
Left femoral head	95.120±1.380	2.310±0.920
Right femoral head	94.890±1.520	2.470±1.050
Small bowel	81.360±6.730	9.420±4.210
Mean	90.410±3.280	4.830±2.100

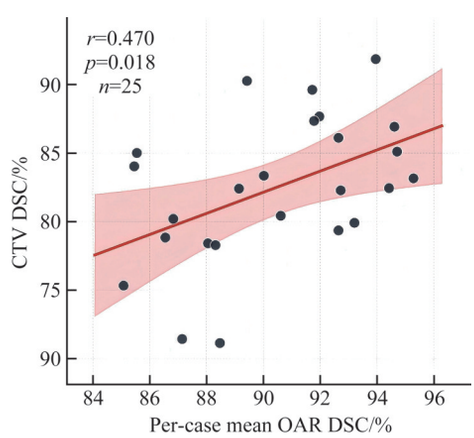


Figure 7 Scatter plot of per-case mean OAR DSC and CTV DSC on the test set (n=25)

The linear regression line and 95% confidence interval are shown. Pearson r=0.470 (P=0.018), indicating a moderate positive correlation between OAR segmentation quality and CTV accuracy.

2.4 Stage-wise performance and multi-task analysis

Table 6 presents HAR-Net performance stratified

©Journal of Central South University (Medical Science). All rights reserved.

by FIGO stage. HAR-Net performs better on early-to-intermediate stage (IB2－IIB) cases (DSC 84.870%), where CTV boundaries are relatively well-defined, compared to locally advanced (IIIA－IVA) cases (DSC 79.300%), where CTV morphology is more complex and boundaries are less distinct.

Table 6 HAR-Net performance stratified by FIGO stage ($\bar{x}\pm s$)

FIGO Stage	*n*	DSC/%	HD95/mm	ASSD/mm
IB2−IIB	14	84.870±4.310	6.820±4.450	1.630±0.710
IIIA−IVA	11	79.300±5.890	10.390±6.870	2.290±1.180

FIGO: International Federation of Gynecology and Obstetrics.

To quantify the contribution of multi-task learning, single-task (CTV-only decoder) and multi-task (shared encoder with OAR decoder but no MOCCA/TCA) configurations were compared against the full HAR-Net (Table 7). Multi-task learning alone improves DSC by 1.340 percentage points over the single-task baseline, indicating that joint OAR segmentation encourages more anatomically discriminative feature representations. Adding MOCCA and TCA further increases DSC by 2.550 points, confirming that explicit cross-attention and axis-aware refinement are essential for fully leveraging OAR context.

2.5 Interpretability analysis

To examine whether MOCCA learns clinically meaningful spatial relationships, cross-attention weights at the 3 decoder stages were visualized by extracting query-key attention maps for CTV boundary voxels and overlaying them on CT slices (Figure 8). At the 1/8 resolution (Stage 3), attention is diffusely distributed across OAR regions, reflecting coarse global position awareness. At the 1/4 resolution (Stage 4), attention concentrates at OAR-CTV interfaces, particularly near the bladder dome and anterior rectal wall. At the 1/2 resolution (Stage 5), attention is sharply focused on fine boundary regions, consistent with clinical practice of first establishing overall anatomical context before refining boundaries.

Table 7 Performance comparison of single-task vs multi-task comparison ($\bar{x}\pm s$)

Configuration	DSC/% ↑	HD95/mm ↓	ASSD/mm ↓
Single-task (CTV only)	78.530±6.180	13.050±7.240	2.510±1.150
Multi-task (+OAR decoder)	79.870±5.950	12.160±7.080	2.350±1.090
HAR-Net (full)	82.420±5.420	8.390±6.120	1.920±0.980

Comparisons were performed on the same train/validation/test split.

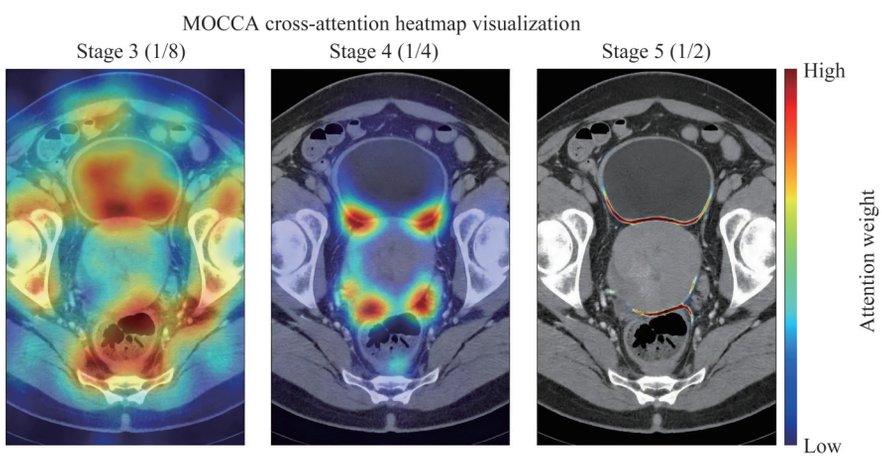


Figure 8 MOCCA cross-attention weight visualization at 3 decoder stages

Progressive focusing is observed from diffuse OAR-region attention at the 1/8 scale to sharp, boundary-specific attention at the 1/2 scale.

©Journal of Central South University (Medical Science). All rights reserved.

3 Discussion

This study introduced HAR-Net, a deep learning framework designed for precise CTV delineation in cervical cancer. HAR-Net addresses the inherent challenge of segmenting CTVs, which are often defined by complex anatomical relationships and consensus guidelines rather than distinct image features, by hierarchically integrating anatomical knowledge. This is primarily achieved through its novel MOCCA module and the TCA module, built upon a robust nnU-Net backbone.

Experimental results indicated that HAR-Net substantially outperformed several state-of-the-art methods in DSC, HD95, and ASSD. HAR-Net achieved a DSC of (82.420±5.420)%, an HD95 of (8.390±6.120) mm, and an ASSD of (1.920±0.980) mm. This represents a notable improvement over the strong nnU-Net baseline, with an increase in DSC by 3.26 percentage points and significant reductions in HD95 by 4.080 mm and ASSD by 0.650 mm. When compared to AMTA-Net, another method that leverages OAR context, HAR-Net demonstrated superior performance across all metrics. This suggests that HAR-Net's explicit multi-scale OAR feature integration via MOCCA, combined with the global refinement from TCA, provides a more effective approach to anatomical prior incorporation than the cascaded attention mechanism used in AMTA-Net. The consistent improvements by HAR-Net are particularly noteworthy given the inherent variability observed in surface distance metrics (HD95, ASSD) across all benchmarked methods. This underlying variability, reflected in the standard deviations (e.g., nnU-Net HD95 standard deviation: 7.380 mm; HAR-Net HD95 standard deviation: 6.120 mm), underscores the complexity of achieving uniformly precise CTV boundary delineation, likely stemming from factors such as subtle inter-patient anatomical differences, indistinct tumor margins on CT, and the guideline-driven nature of CTVs. While HAR-Net achieved the most favorable mean HD95 (8.390 mm), indicating superior average boundary accuracy, its associated standard deviation (6.120 mm) warrants consideration. This suggests that while HAR-Net is, on average, highly accurate, a subset of cases might still exhibit larger boundary deviations. This is a common challenge in automated segmentation, highlighting the ongoing need for refinement in worst-case scenario robustness, even with advanced architectures. Nevertheless, the average improvements in DSC, HD95, and ASSD are promising for clinical translation, potentially leading to more reliable target coverage and OAR sparing.

The ablation experiments provide critical insights into the interplay of HAR-Net's components and their individual contributions. The baseline nnU-Net achieved a DSC of (77.920±6.280)%, HD95 of (13.290±6.920) mm, and ASSD of (2.470±1.030) mm. The MOCCA module, when integrated alone (Baseline+MOCCA), yielded an improved DSC and a slightly higher ASSD over this baseline. This underscores the value of explicitly modeling OAR-CTV spatial relationships at multiple scales for better volumetric overlap. However, a notable observation was the degradation in HD95 [(14.140±7.920) mm vs (13.290±6.920) mm for baseline] with a concurrent increase in its standard deviation when MOCCA was used in isolation. This "DSC improves but HD95 degrades" phenomenon arises from the fundamentally different measurement dimensions of these metrics. DSC quantifies global volumetric overlap and is sensitive to overall segmentation accuracy; HD95 captures the 95th percentile worst-case surface distance and is highly sensitive to localized boundary errors. MOCCA effectively increases correctly predicted voxels (improving DSC) by incorporating OAR contextual features. However, in anatomically complex OAR-CTV interface regions (e. g., areas with variable bladder filling), uncertainties in OAR feature representations can propagate through cross-attention weights to CTV boundary predictions, inducing localized over- or under-segmentation. These sporadic large deviations, while negligible for global DSC, are acutely captured by HD95. When TCA is subsequently added (full HAR-Net), its global axis-wise structural priors effectively regularize these local boundary perturbations, reducing HD95 dramatically to 8.390 mm. This "local enhancement followed by global constraint" hierarchy produces the observed synergy: MOCCA enhances volumetric accuracy, while TCA ensures boundary

©Journal of Central South University (Medical Science). All rights reserved.

consistency.

This suggests that while local OAR context is powerful, it might, without subsequent global regularization, occasionally lead to larger localized boundary errors. This could occur if OAR features become overly influential in regions of ambiguity or if upstream OAR segmentations themselves possessed inaccuracies. Conversely, the TCA module alone (Baseline+TCA) demonstrated its specific strength in improving surface distance metrics relative to the baseline+MOCCA, consistent with its design to enforce long-range, axis-aligned structural priors relevant to anatomical features like the psoas muscles. The DSC with TCA alone (78.470±6.070) was slightly better than baseline, but did not reach the level of MOCCA alone, possibly reflecting a trade-off where global smoothing for anatomical consistency could slightly impact fine voxel-level agreement if not complemented by strong local context.

Crucially, the complete HAR-Net model (Baseline+MOCCA+TCA) not only achieved the best overall performance but also demonstrated a strong synergistic effect. The inclusion of TCA alongside MOCCA rectified the HD95 degradation seen with MOCCA alone, reducing it significantly to 8.390 mm, and simultaneously uplifted the DSC beyond what either module achieved individually. This robustly supports the hypothesis that a hierarchical approach—integrating local OAR-derived context (MOCCA) followed by global, axis-aware structural refinement (TCA) —is essential for achieving both high volumetric accuracy and precise, anatomically consistent CTV boundaries. This explicit, two-stage refinement strategy distinguishes HAR-Net from approaches that rely on more implicit learning of anatomical context or employ general attention mechanisms without specific adaptations for OAR-CTV interplay and tri-axial volumetric priors.

The primary strengths of HAR-Net lie in its novel architectural components designed to intelligently embed anatomical intelligence and its hierarchical refinement strategy. Leveraging the robust nnU-Net framework as a backbone further ensures effective multi-scale feature extraction. The enhanced accuracy and consistency demonstrated by HAR-Net could potentially translate to more reliable CTV delineations in clinical practice, possibly reducing inter-observer variability and streamlining the radiotherapy planning workflow.

The OAR branch achieved a mean DSC of 90.410% across 5 pelvic organs, with a moderate positive correlation (r=0.470, P=0.018) between per-case OAR accuracy and CTV segmentation quality. This supports the design premise that high-quality OAR features serve as reliable anatomical anchors for MOCCA, while also suggesting that further improvement in OAR segmentation—particularly for challenging organs like rectum (DSC of 87.230%) and small bowel (DSC of 81.360%)—could yield additional CTV accuracy gains. The stage-wise analysis revealed expected performance differences: early-stage cases (IB2－IIB, DSC of 84.870%) exhibited more favorable results than locally advanced cases (IIIA－IVA, DSC 79.300%), reflecting the increased CTV morphological complexity and boundary ambiguity in advanced disease. The multi-task comparison demonstrated that while joint OAR-CTV learning contributes meaningfully (a 1.340 percentage point improvement in DSC), the explicit spatial reasoning provided by MOCCA and TCA accounts for the majority of the performance gain (a 2.550 percentage point improvement in DSC), confirming that implicit feature sharing alone is insufficient for optimal CTV delineation. The interpretability analysis further demonstrates that MOCCA learns attention patterns consistent with clinical knowledge, with a progressive coarse-to-fine focusing behavior across decoder stages.

Despite promising results, this study has limitations. The findings are based on a single-institution dataset; multi-center validation is crucial for assessing generalizability. The observed variability in HD95 for HAR-Net, even with its superior mean, indicates that ensuring robustness across all patient cases, particularly those with highly challenging anatomical presentations or atypical OAR configurations, remains an area for improvement. Furthermore, the specific observation that the MOCCA module, in isolation, can sometimes adversely affect HD95 suggests that the interplay between OAR features

©Journal of Central South University (Medical Science). All rights reserved.

and CTV prediction needs careful management, and its performance is likely dependent on the accuracy of the OAR features it receives. The impact of OAR segmentation accuracy on MOCCA's effectiveness, and the optimal selection of OARs for contextualization, therefore warrant further dedicated study. Additionally, HAR-Net's computational cost is inherently higher than the baseline nnU-Net due to the additional modules.

Future research should prioritize several avenues: 1) External validation on diverse, multi-center datasets to confirm robustness and generalizability. 2) Targeted research to understand and mitigate the variance in HAR-Net's HD95 performance, potentially by exploring adaptive mechanisms that modulate the influence of MOCCA and TCA based on image-derived uncertainty or anatomical complexity. 3) In-depth investigation into optimizing OAR feature integration within MOCCA, including its sensitivity to upstream OAR segmentation quality, the impact of different OAR combinations, and methods to make it more robust to OAR segmentation errors. 4) Exploration of methods to enhance clinical guideline integration more directly, potentially through advanced loss functions or rule-based post-processing guided by the network's outputs. 5) Efforts to reduce computational demands, perhaps through model pruning or knowledge distillation, without sacrificing performance. 6) Qualitative clinical evaluation by radiation oncologists to assess the clinical acceptability and utility of HAR-Net's segmentations, particularly in ambiguous or challenging cases.

In conclusion, HAR-Net shows significant promise for advancing automated CTV delineation in cervical cancer. By explicitly modeling hierarchical anatomical relationships through its MOCCA and TCA modules, it addresses key limitations of existing approaches. The framework has the potential to improve the efficiency, accuracy, and consistency of radiotherapy planning, which may ultimately contribute to better patient outcomes.

Acknowledgements: We gratefully acknowledge the experimental assistance and computational resources provided by the Jiaxing Yaohe Tech research platform (PaperMind).

Contributions: CHENG Tingting Experimental data collection and statistical analysis; DENG Liang Data acquisition and preprocessing; LIANG Zhan and TANG Du Experiment oversight, computational analysis. ZHANG Zijian Research design, paper writing and revision. The final version of the manuscript has been approved and read by all authors.

Conflict of interest: The authors declare that they have no conflicts of interest to disclose.

Data availability statement: All data supporting the findings of this study are available from Xiangya Hospital, Central South University, but restrictions apply to the availability of these data, which were used under licence for the current study and so are not publicly available. The data are available upon request and with the permission of Xiangya Hospital, Central South University.

References

[1] Chino J, Annunziata C, Beriwal S, et al. Radiation therapy for cervical cancer: executive summary of an ASTRO clinical practice guideline[J]. Pract Radiat Oncol, 2020, 10(4): 220-234. https://doi.org/10.1016/j.prro.2020.04.002.

[2] Oh D, Huh SJ, Park W, et al. Clinical outcomes in cervical cancer patients treated by FDG-PET/CT-based 3-dimensional planning for the first brachytherapy session[J/OL]. Medicine, 2016, 95(25): e3895[2026-02-01]. https://doi.org/10.1097/md.0000000000003895.

[3] Zabihollahy F, Viswanathan AN, Schmidt EJ, et al. Fully automated segmentation of clinical target volume in cervical cancer from magnetic resonance imaging with convolutional neural network[J/OL]. J Applied Clin Med Phys, 2022, 23(9): e13725[2026-02-01]. https://doi.org/10.1002/acm2.13725.

[4] Vinod SK, Jameson MG, Min M, et al. Uncertainties in volume delineation in radiation oncology: a systematic review and recommendations for future studies[J]. Radiother Oncol, 2016, 121(2): 169-179. https://doi.org/10.1016/j.radonc.2016.09.009.

[5] Tai P, Van Dyk J, Yu E, et al. Variability of target volume delineation in cervical esophageal cancer[J]. Int J Radiat Oncol Biol Phys, 1998, 42(2): 277-288. https://doi.org/10.1016/s0360-3016(98)00216-8.

[6] Song Y, Hu J, Wu Q, et al. Automatic delineation of the clinical target volume and organs at risk by deep learning for rectal cancer postoperative radiotherapy[J]. Radiother Oncol, 2020, 145: 186-192. https://doi.org/10.1016/j.radonc.2020.01.020.

[7] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification

©Journal of Central South University (Medical Science). All rights reserved.

with deep convolutional neural networks[J]. Commun ACM, 2017, 60(6): 84-90. https://doi.org/10.1145/3065386.

[8] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. arXiv, (2014-09-04) [2026-02-01]. https://arxiv.org/abs/1409.1556.

[9] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 7-12, 2015, Boston, MA, USA. IEEE, 2015: 3431-3440. https://doi.org/10.1109/CVPR.2015.7298965.

[10] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[M]//Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. Cham: Springer International Publishing, 2015: 234-241. https://doi.org/10.1007/978-3-319-24574-4_28.

[11] Çiçek Ö, Abdulkadir A, Lienkamp SS, et al. 3D U-Net: learning dense volumetric segmentation from sparse annotation [M]//Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016. Cham: Springer International Publishing, 2016: 424-432. https://doi.org/10.1007/978-3-319-46723-8_49.

[12] Isensee F, Jaeger PF, Kohl S, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation [J]. Nat Meth, 2021, 18(2): 203-211. https://doi.org/10.1038/s41592-020-01008-z.

[13] Ibtehaz N, Rahman MS. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation[J]. Neural Netw, 2020, 121: 74-87. https://doi.org/10.1016/j.neunet.2019.08.025.

[14] Zhou HY, Guo J, Zhang Y, et al. nnFormer: interleaved transformer for volumetric segmentation[EB/OL]. arXiv, (2021-09-07) [2026-02-01]. https://arxiv.org/abs/2109.03201

[15] Hatamizadeh A, Nath V, Tang Y, et al. Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images[M]//Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Cham: Springer International Publishing, 2022: 272-284. https://doi.org/10.1007/978-3-031-08999-2_22.

[16] Diakogiannis FI, Waldner F, Caccetta P, et al. ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data[J]. ISPRS J Photogramm Remote Sens, 2020, 162: 94-114. https://doi.org/10.1016/j.isprsjprs.2020.01.013.

[17] Shi F, Hu W, Wu J, et al. Deep learning empowered volume delineation of whole-body organs-at-risk for accelerated radiotherapy[J]. Nat Commun, 2022, 13: 6566. https://doi.org/10.1038/s41467-022-34257-x.

[18] Rahman MM, Marculescu R. Medical image segmentation via cascaded attention decoding[C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). January 2-7, 2023, Waikoloa, HI, USA. IEEE, 2023: 6211-6220. https://doi.org/10.1109/WACV56688.2023.00616.

[19] Milletari F, Navab N, Ahmadi SA. V-net: fully convolutional neural networks for volumetric medical image segmentation [C]//2016 Fourth International Conference on 3D Vision (3DV). October 25-28, 2016, Stanford, CA, USA. IEEE, 2016: 565-571. https://doi.org/10.1109/3DV.2016.79.

[20] Ma CY, Zhou JY, Xu XT, et al. Deep learning-based auto-segmentation of clinical target volumes for radiotherapy treatment of cervical cancer[J/OL]. J Applied Clin Med Phys, 2022, 23(2): e13470 [2026-02-01]. https://doi.org/10.1002/acm2.13470.

[21] Peng H, Liu T, Li P, et al. Automatic delineation of cervical cancer target volumes in small samples based on multi-decoder and semi-supervised learning and clinical application[J]. Sci Rep, 2024, 14: 26937. https://doi.org/10.1038/s41598-024-78424-0.

[22] Zang L, Liu J, Zhang H, et al. A deep learning model based on Mamba for automatic segmentation in cervical cancer brachytherapy[J]. Sci Rep, 2025, 15: 10152. https://doi.org/10.1038/s41598-025-94431-1.

[23] Khor HG, Ning G, Sun Y, et al. Anatomically constrained and attention-guided deep feature fusion for joint segmentation and deformable medical image registration[J]. Med Image Anal, 2023, 88: 102811. https://doi.org/10.1016/j.media.2023.102811.

[24] Deng XW, Zhao HM, Jia LC, et al. Prior knowledge-guided U-Net for automatic clinical target volume segmentation in postmastectomy radiation therapy of breast cancer[J]. Int J Radiat Oncol, 2025, 121(5): 1361-1371. https://doi.org/10.1016/j.ijrobp.2024.11.104.

[25] Liu L, Wolterink JM, Brune C, et al. Anatomy-aided deep learning for medical image segmentation: a review[J/OL]. Phys Med Biol, 2021, 66(11): 11TR01[2026-02-01]. https://doi.org/10.1088/1361-6560/abfbf4.

[26] Zhang Y, Liang Y, Ding J, et al. A prior knowledge-guided, deep learning-based semiautomatic segmentation for complex anatomy on magnetic resonance imaging[J]. Int J Radiat Oncol, 2022, 114(2): 349-359. https://doi.org/10.1016/j.ijrobp.2022.05.039.

[27] Dalca AV, Guttag J, Sabuncu MR. Anatomical priors in convolutional networks for unsupervised biomedical segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. June 18-23, 2018, Salt Lake City, UT, USA. IEEE, 2018: 9290-9299. https://doi.org/10.1109/CVPR.2018.00968.

[28] Boutillon A, Conze PH, Pons C, et al. Generalizable multi-task, multi-domain deep segmentation of sparse pediatric imaging datasets via multi-scale contrastive regularization and multi-joint anatomical priors[J]. Med Image Anal, 2022, 81: 102556. https://doi.org/10.1016/j.media.2022.102556.

[29] Xu X, Lian C, Wang S, et al. Asymmetrical multi-task attention U-Net for the segmentation of prostate bed in CT image [M]//Medical Image Computing and Computer Assisted Intervention-MICCAI 2020. Cham: Springer International Publishing, 2020: 470-479. https://doi.org/10.1007/978-3-030-59719-1_46.

[30] Bohlender S, Oksuz I, Mukhopadhyay A. A survey on shape-constraint deep learning for medical image segmentation[J].

©Journal of Central South University (Medical Science). All rights reserved.

IEEE Rev Biomed Eng, 2023, 16: 225-240. https://doi.org/10.1109/RBME.2021.3136343.

[31] Mlynarski P, Delingette H, Alghamdi H, et al. Anatomically consistent CNN-based segmentation of organs-at-risk in cranial radiotherapy[J]. J Med Imag, 2020, 7(1): 1. https://doi.org/10.1117/1.jmi.7.1.014502.

[32] Sadikine A, Badic B, Tasu JP, et al. Improving abdominal image segmentation with overcomplete shape priors[J]. Comput Med Imag Graph, 2024, 113: 102356. https://doi.org/10.1016/j.compmedimag.2024.102356.

[33] Small W, Bosch WR, Harkenrider MM, et al. NRG oncology/RTOG consensus guidelines for delineation of clinical target volume for intensity modulated pelvic radiation therapy in postoperative treatment of endometrial and cervical cancer: an update[J]. Int J Radiat Oncol, 2021, 109(2): 413-424. https://doi.org/10.1016/j.ijrobp.2020.08.061.

[34] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[EB/OL]. arXiv, (2017-06-12) [2026-02-01]. https://arxiv.org/abs/1706.03762.

[35] Hou QB, Zhou DQ, Feng JS. Coordinate attention for efficient mobile network design[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 20-25, 2021. Nashville, TN, USA. IEEE, 2021: 13708-13717. https://doi.org/10.1109/cvpr46437.2021.01350.

[36] Bhatla N, Berek JS, Cuello Fredes M, et al. Revised FIGO staging for carcinoma of the cervix uteri[J]. Int J Gynecol Obstet, 2019, 145(1): 129-135. https://doi.org/10.1002/ijgo.12749.

(Edited by WU Xufang)

©Journal of Central South University (Medical Science). All rights reserved.