Intelligent Methods in Engineering Sciences 4(2): 38-53, 2025

IMIENS

INTELLIGENT METHODS
IN ENGINEERING SCIENCES

https://www.imiens.org

International Open Access

August, 2025

e-ISSN 2979-9236

Research Article

https://doi.org/10.58190/imiens.2025.126

Dual-Scale Transformer-Guided Attention Network for Efficient Multi-OAR Segmentation in Head and Neck Radiotherapy

Uzma Nawaz [a], *Hafiz Muhammad Ubaidullah* [b], *Zubair Saeed* [c,*], *Chaudhry Muhammad Ali Nawaz* [d]

[a] *Knowledge and Data Science Research Centre, Department of Computer and Software Engineering, College of Electrical and Mechanical Engineering, National University of Science and Technology, Islamabad, Pakistan*
[b] *Department of Computer Information Systems Engineering, NED University of Engineering and Technology, Karachi, Pakistan*
[c] *Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX, USA*
[d] *Department of Creative Technologies, Software Engineering, Air University, Islamabad, Pakistan*

ARTICLE INFO

Article history:
Received 14 July 2025
Accepted 1 August 2025

Keywords:
Deep Learning,
segmentation,
Organ-at-risk,
CT images,
convolutional neural networks,
dual-scale transformer-guided attention network

ABSTRACT

Accurate segmentation of organ-at-risk (OARs) in head and neck CT images is crucial for radiotherapy planning, but it remains a challenging task due to anatomical complexity, low soft-tissue contrast, and the presence of small, variable structures. We propose DSTANet, a novel dual-scale transformer-guided attention network that integrates multi-resolution encoding, transformer-based global context fusion, and anatomically guided attention refinement to deliver precise multi-OAR segmentation. Unlike traditional CNN-based methods, DSTANet effectively models long-range spatial dependencies while preserving high-resolution boundary detail. On the HNSCC-3DCT-RT dataset, DSTANet achieved a mean Dice Score of 97.5% and a mean 95th percentile Hausdorff Distance (HD95) of 2.32 mm, while on the MICCAI 2015 benchmark dataset, it achieved 90.0% Dice, which surpasses several state-of-the-art approaches both in terms of overlap and geometric accuracy. These results, combined with a sub-20-second inference time, establish DSTANet as a robust and clinically viable solution for automated head and neck OAR segmentation.

1. INTRODUCTION

Accurate segmentation of organs at risk (OARs) in radiotherapy planning is crucial for the safe and successful delivery of therapeutic radiation doses [1]. In head and neck (H&N) cancer, where multiple vital organs are densely packed into a small anatomical area [2], careful identification of organs at risk (OARs) such as the optic nerves, chiasm, brainstem, and parotid glands is essential. Minor segmentation errors can lead to significant changes in dose distribution, resulting in inadequate tumor coverage or, more importantly, inadvertent irradiation of healthy tissues [3]. This can result in substantial long-term toxicities such as eye loss, salivary malfunction, and neurological damage. As a result, automatic and precise OAR segmentation has become an essential component in the development of clinically deployable radiation systems [4].

However, high-precision segmentation in the head and neck (H&N) area remains a significant challenge [5]. The organs vary significantly in size, ranging from minuscule structures like the optic chiasm, which is just a few voxels thick, to huge volumes like the mandible [6]. Many OARs are closely located and often appear in low contrast against surrounding tissues, with blurry or poorly-defined borders on non-contrast CT scans [7]. These qualities impede standard intensity-based segmentation and demand architectures capable of resolving fine spatial information while understanding global anatomical context. Furthermore, volumetric segmentation of entire 3D CT images imposes a significant computational cost, frequently necessitating trade-offs between model complexity, inference time, and memory usage, factors that impede real-time implementation in clinical processes. The growing success of artificial intelligence across diverse domains, including natural language processing, industrial automation, and cybersecurity [8-12], further reinforces its potential to enhance segmentation accuracy, reduce clinical workload, and improve treatment safety in radiotherapy applications.

Several approaches have been investigated to overcome these difficulties, ranging from atlas-based registration

* **Corresponding Author:** zubairsaeed602@gmail.com

techniques to fully convolutional neural networks (FCNs) [13], 3D U-Nets [14], and current attention-based models such as PANet [15]. While these approaches have provided promising results, they tend to underperform on small and overlapping structures because they rely on local receptive fields or manually crafted regional features. Furthermore, most topologies do not scale efficiently across full-volume inputs or generalize well across a wide range of patient anatomy and imaging techniques [16]. The lack of anatomical prior integration and explicit boundary refinement processes often results in coarse or misaligned segmentations, particularly for low-contrast or partially occluded OARs [17].

To address these limitations, we introduce DSTANet, a new Dual-Scale Transformer-Guided Attention Network specially designed for efficient and accurate multi-OAR segmentation in head and neck radiation. DSTANet incorporates three key innovations: a dual-scale encoder that captures both high-resolution local features and low-resolution global context, a transformer-guided fusion module that models long-range spatial dependencies, and a guided attention refinement block that uses anatomical prior maps to improve boundary precision. Notably, the architecture remains lightweight and computationally efficient while maintaining accuracy. We undertake thorough assessments on both the HNSCC-3DCT-RT [18] and MICCAI 2015 datasets [19], indicating that DSTANet surpasses cutting-edge baselines in terms of Dice accuracy and boundary conformity, particularly for small and challenging organs at risk (OARs).

The rest of the paper is as follows: Section 2 discusses the related work. The proposed methodology is presented in Section 3. Section 4 shows the experimental setup. Sections 5 and 6 present results, analysis, and discussions. Section 7 shows the conclusion and future work.

2. Related work

DL-based solutions have been used in many fields, i.e., disease detection in humans [30] – [38], plants [39] [40], and various multidisciplinary fields [41] – [46]. Similarly, DL techniques are also being used for the early segmentation of tumors and OAR. Zhu et al. (2019) [20] proposed AnatomyNet, an end-to-end 3D convolutional neural network designed to segment head and neck organs-at-risk (OARs) directly from whole-volume CT scans. The model incorporated squeeze-and-excitation residual blocks and a hybrid Dice-focal loss function to balance small and large organ segmentation better. On the MICCAI 2015 dataset, AnatomyNet achieved an average Dice score of 78.3% across nine OARs. However, performance for small-volume architecture, such as the optic nerves and chiasm, remained suboptimal, with chiasm Dice falling by 70%, which highlights limitations in spatial context modeling and fine detail preservation. To directly address the large-to-small organ size imbalance, Gao et al. (2019) [21] proposed FocusNet, which addressed the imbalance between large and small OARs by incorporating organ localization and a dual-path architecture, including a dedicated branch for small structures. Their model achieved an average Dice score of 81.5%, indicating strong improvements for organs such as the optic nerves, where the Dice scores exceeded 76%.

Liang et al. (2020) [22] developed a multi-view ROI aggregation framework that employed fine-grained CNNs across axial, sagittal, and coronal views to perform localization and segmentation jointly. On the MICCAI 2015 dataset, their model achieved a mean Dice of 82.1%, outperforming traditional single-view models, especially on mid-size organs such as the parotids. However, its reliance on 2D projections reduced volumetric continuity, and small organs at risk (OARs), such as the chiasm, still exhibited lower boundary fidelity. Chen et al. (2020) [23] introduced a PANet within a stepwise refinement framework, integrating prior attention and multi-scale feature pyramids to guide the segmentation process. PANet achieved a mean Dice of 85.2% across all OARs and performed exceptionally well on large and medium organs such as the brainstem and mandible. Nonetheless, it reported lower accuracy on small targets such as the optic chiasm (69.5% Dice), due to limited boundary-aware refinement and lack of global spatial reasoning.

Wang et al. (2021) [24] extended segmentation techniques to dual-energy CT (DECT) using attention-enhanced dual pyramid R-CNNs. The model leveraged the additional spectral contrast from DECT channels to achieve a mean Dice of 86.1%, notably improving performance on soft-tissue structures. However, the requirement for DECT imaging restricts its applicability in many clinical environments that rely on standard single-energy CT (SECT).

Dai et al. (2022a) [25] transitioned to MRI-based segmentation, designing an attention-enhanced pyramid network coupled with a mask-scoring R-CNN to delineate OARs. Their model achieved mean Dice scores exceeding 88% for high-contrast soft-tissue organs in MRI, such as the brain and parotids. Despite strong results, the architecture was validated only on MRI data and not tested on CT, limiting its cross-modality generalization. Dai et al. (2022b) [26] proposed a deep learning pipeline for tracking anatomical variation across weekly QA CT scans using FCOS-based detection and hierarchical refinement. Although not tailored for baseline segmentation, their method yielded time-consistent Dice scores of around 83% for large organs, which helps monitor tumor shrinkage and plan adaptation. However, its coarse output was less suited for precise OAR delineation at the initial planning stage. The summary of related work is shown in Table 1.

Table 1. Summary of the Related Work.

Author	Approach	Improvement	Limitations
Zhu et al. (2019) [20]	AnatomyNet (3D U-Net + SE blocks + hybrid loss)	Mean Dice: 78.3%; fast inference; small organ handling	Weak on chiasm/nerves; no attention or longer-range modeling
Gao et al. (2019) [21]	FocusNet with organ-specific subnetworks	Mean Dice: 81.5% strong on optic nerves	Complex ROI design; not fully end-to-end
Liang et al. (2020) [22]	Multi-view 2D ROI aggregation across planes	Mean Dice: 82.1%; better mid-size organ accuracy	Uses 2D slices; weak on 3D consistency and fine details
Chen et al. (2020) [23]	PANet + stepwise refinement + attention pyramids	Mean Dice: 85.2% solid multi-scale learning	Poor chiasm Dice (~69.5%); no global context modeling
Wang et al. (2021) [24]	DECT attention-based dual pyramid R-CNN	Mean Dice: 86.1%; good for soft tissues	Requires DECT scanner; not widely available
Dai et al. (2022a) [25]	MRI-based mask-scoring R-CNN with attention pyramid	Dice > 88% on MRI organs	MRI only; not validated on CT data
Dai et al. (2022b) [26]	FCOS + refinement for temporal QA segmentation	Dice ~83% for large organs; useful for treatment tracking	Weak at baseline segmentation; low precision on small structures

2.1. Gap analysis

Despite tremendous advances in automated segmentation of head and neck organs-at-risk (OARs), present methods have significant drawbacks that affect both accuracy and clinical applicability. Many previous models depend heavily on convolutional backbones, which have a finite receptive field. As a result, they frequently fail to capture the long-range spatial interdependence required to appropriately separate anatomically distributed components such as the bilateral parotids, optic nerves, and chiasm. While designs like PANet and FocusNet have offered local attention or ROI-based submodules, these solutions are still heavily reliant on local context and need hand-crafted area proposals or extensive network management.

Furthermore, performance deterioration on small-volume organs at risk (OARs) is a recurring concern in the research. Even with high-performing models, Dice scores for structures like the optic chiasm and nerves frequently dip below 75%, owing to poor soft-tissue contrast and insufficient border representation. This is compounded by class imbalance during training, in which tiny organs make negligible contributions to the loss function. Although solutions such as dice-focal hybrid loss have been investigated, they often provide only modest benefits in the absence of architectural support for fine-grained feature augmentation.

Another significant restriction is the design separation of the global encoding and refining phases. Many current networks either stress global context over spatial detail or rely on postprocessing refinement processes that cannot be trained end-to-end. This separation introduces duplication, increases computational complexity, and hinders the cohesive learning of boundary-aware features. Furthermore, the lack of anatomically directed processes in most segmentation pipelines implies that model attention is unconstrained by biological priors, raising the possibility of mislocalization in the presence of anatomical variances, tumor-induced deformations, or image distortions.

Finally, generalizability across datasets is still an issue. Several high-performing models perform well on the MICCAI 2015 dataset but experience a decrease in performance when applied to more diverse real-world datasets, such as HNSCC-3DCT-RT. This is frequently owing to overfitting on limited data distributions and a lack of explicit modules for collecting high-level anatomical semantics, which are required for resilience across imaging methods and patient groups.

In summary, there is still a need for a unified segmentation architecture that can: (i) model both global and local context at the same time; (ii) incorporate anatomical prior knowledge directly into the learning process; (iii) improve boundary refinement for small and low-contrast OARs; and (iv) maintain generalizability across datasets without relying on handcrafted operations. Addressing these deficiencies is the driving force for the creation of DSTANet, our proposed Dual-Scale Transformer-Guided Attention Network for multi-OAR segmentation in head and neck radiation.

2.2. Key Innovations and Novelty

DSTANet features several architectural advancements that distinguish it from previous transformer-based segmentation models in medical imaging. This includes:

Dual-scale encoding is a two-stream encoder that captures both high-resolution spatial data and low-resolution semantic context, retaining boundaries and simulating organ-level distribution.

Transformer-guided 3D fusion: Unlike previous 2D transformer applications, DSTANet uses a 3D multi-head self-attention mechanism to capture long-range dependencies across volumetric organ structures, assisting in the segmentation of anatomically similar but geographically disparate areas.

Anatomical prior integration: The model uses soft anatomical priors from a weakly supervised locator network to guide attention in a biologically informed manner, which is absent in most previous transformer-based architectures.

Dual-head decoder: DSTANet employs different decoder branches for big and small OARs, solving class imbalance directly without the use of sophisticated loss functions or sampling algorithms.

Together, these advancements enable DSTANet to conduct end-to-end segmentation that is both computationally fast and clinically robust, distinguishing it from previous studies that depend largely on handmade modules or external refinement processes.

3. Proposed Methodology

3.1. Overview of DSTANet Architecture

This section presents the architecture and core innovations of the proposed Dual-Scale Transformer-Guided Attention Network (DSTANet), designed for accurate, efficient, and generalizable segmentation of multiple Organs-At-Risk (OARs) in head and neck radiotherapy, as shown in Fig. 1. The technique is designed to address major issues, including the disparity in size among OARs, indistinct organ borders, and the significant computational expenses associated with full-volume CT segmentation. Our design uses a transformer-guided fusion module for contextual integration, a dual-path encoding mechanism for multi-resolution feature extraction, and a guided attention module to fine-tune segmentation based on learnable spatial cues and anatomical priors.

3.2. Input Preprocessing and Anatomical Prior Generation

The performance of medical image segmentation models, particularly in radiotherapy applications, is highly sensitive to the spatial consistency and intensity normalization of the input volumes. Therefore, the initial phase of our proposed DSTANet pipeline performs standardized preprocessing on all CT data. Given an input 3D CT volume $\mathcal{J} \in \mathbb{R}^{H\times W\times D}$, we apply intensity clipping based on soft-tissue windowing parameters: window level WL= 40 and window with WW = 350, such that:

$$\mathcal{J}_{clip}(x) = \begin{cases} WL - \frac{WW}{2}, & if\ \mathcal{J}(x) < WL - \frac{WW}{2} \\ \mathcal{J}(x), & if\ WL - \frac{WW}{2} \leq \mathcal{J}(x) \leq WL + \frac{WW}{2} \\ WL + \frac{WW}{2}, & if\ \mathcal{J}(x) > WL - \frac{WW}{2} \end{cases}$$

The resulting image is then normalized linearly to the range [-1,1]. This transformation is crucial to stabilize training across heterogeneous CT scanners and imaging conditions. To enable targeted special reasoning in the subsequent network modules, we incorporate anatomical priors in the form of voxel-wise probability maps. These priors are not learned jointly with the main segmentation task but are derived from a lightweight ROI Locator Network $\mathcal{R}_{\theta}$, trained separately using weakly supervised group-wise labels. The network $\mathcal{R}_{\theta}$ maps an input CT to a multi-class probability distribution over anatomical regions:

$$\mathcal{P}_{ROI=}\mathcal{R}_{\theta}\left(\mathcal{J}_{clip}\right), \quad \mathcal{P}_{ROI} \in \mathbb{R}^{H\times W\times D\times C_{group}}$$

Here $C_{group} = 3$, corresponding to three anatomical groups:

Group A: Large structural OARs (e.g., mandible, brainstem)

Group B: Medium-volume organs (e.g., TMJs, mastoids)

Group C: Small and low-contrast OARs (e.g., optic nerves, chiasm, pituitary)

The output $\mathcal{P}_{ROI}$ serves two purposes: (1) it defines the region of interest (ROI) for each organ cluster, and (2) it forms the basis of spatial attention guidance in downstream encoder and fusion modules. This anatomical prior becomes a soft form of structural context, injected throughout the network to suppress background and enforce regional focus. Unlike binary cropping or complicated masks, these priors retain smooth transitions and class overlap, making them more suitable for gradient-based attention mechanisms. The decoder and transformer layers later modulate their attention based on $\mathcal{P}_{ROI}$, enhancing organ-wise specificity, especially in ambiguous zones.

3.3. Dual-Scale Feature Encoding

For the feature encoding step, the ROI Locator Network's output anatomical prior map, $\mathcal{P}_{ROI}$, acts as a spatial guide. In order to address the well-known problem of scale imbalance in organ segmentation, where small organs (such as optic nerves and lenses) run at risk of vanishing in deep layers and large organs need a lot of context to be accurately localized, the dual-scale encoding concept was created. For low-resolution semantic abstraction, we therefore provide a global encoder, while for high-resolution boundary preservation, we present a local encoder.

3.3.1. Global Context Encoder

The global encoder processes a downsampled version of the normalized CT volume $\mathcal{J}_{clip}^{\downarrow 2} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times\frac{D}{2}}$, concatenated with the downsampled anatomical prior map $\mathcal{P}_{ROI}^{\downarrow 2}$. The input tensor to this encoder is:

$$X_g = Concat\ (\mathcal{J}_{clip}^{\downarrow 2}, \mathcal{P}_{ROI}^{\downarrow 2})$$

This encoder $\mathcal{E}_g$ consists of four convolutional stages, each composed of:

Depthwise Separable 3D Convolution (3 × 3 × 3)

Batch Normalization

Parametric ReLU activation

The feature propagation at layer *l* is defined as:

$$\mathcal{F}_{global}^{(l)} = PReLU(BN\left(DWConv3D\left(\mathcal{F}_{global}^{(l-1)}\right)\right))$$

By splitting typical 3D convolution into distinct spatial and pointwise convolutions, depth-wise separable convolution drastically lowers processing costs. This maintains a small parameter footprint while preserving volumetric context.

3.3.2. Local Detail Encoder

To preserve edge detail and delicate anatomical structures, the local encoder operates on the original resolution image, and prior to concatenation:

$$X_l = Concat\,(\mathcal{I}^{\downarrow 2}_{clip}, \mathcal{P}^{\downarrow 2}_{ROI})$$

Unlike the global encoder, $\mathcal{E}_l$ employs full 3D convolution blocks without downsampling, ensuring voxel-level fidelity. Each stage contains:

Standard 3D Convolution (3 × 3 × 3)

Group Normalization (preferred for small batch sizes)

GELU activation (smoother than ReLU, helps in attention propagation)

The local encoder feature propagation follows:

$$\mathcal{F}^{(l)}_{local} = GELU(GN\left(Conv3D\left(\mathcal{F}^{(l-1)}_{local}\right)\right))$$

This pathway retains the high-frequency features that typically represent the boundaries of small, low-contrast organs, such as the chiasm and pituitary, among others.

3.3.3. Feature Alignment and Multi-Scale Fusion Preparation

The output tensors $\mathcal{F}_{global}$ and $\mathcal{F}_{local}$ are channel-aligned and upsampled/downsampled as needed using trilinear interpolation, forming a unified shape $\mathbb{R}^{H\times W\times D\times C}$. This step ensures that both streams can be seamlessly integrated within the transformer-guided fusion block described in the next section. We define the aligned feature maps as:

$$\mathcal{F}_{align} = Align\left(\mathcal{F}_{global}, \mathcal{F}_{local}\right) = \mathcal{F}^{\uparrow}_{local} + \mathcal{F}_{local}$$

This element-wise fusion is not simply additive in behavior: the low-res global context provides semantic weight, while the high-res features inject edge precision. The aligned tensor $\mathcal{F}_{align}$ becomes the input to the Transformer module, which adaptively integrates these dual cues across the spatial field.

3.4. Transformer-Guided Contextual Fusion

Having obtained the aligned feature tensor $\mathcal{F}_{align} \in \mathbb{R}^{H\times W\times D\times C}$ from the dual-scale encoders, the next critical step is to enrich this representation with global anatomical context. Traditional CNNs, while effective in extracting local features, struggle to model long-range dependencies, a limitation that is especially detrimental in head and neck segmentation, where anatomically linked structures (e.g., bilateral optic nerves, or the spatial relationship between the chiasm and brainstem) must be understood collectively rather than in isolation.

To address this, we introduce a 3D Transformer-Guided Fusion Module $\mathcal{T}_{\varphi}$ that applies multi-head self-attention (MHSA) to learn spatially-aware dependencies across the entire volumetric space. The transformer block receives as input the reshaped volumetric feature $\mathcal{F}_{align}$ flattened into a sequence of tokens. Let the reshaped input be defined as:

$$\mathcal{Z}_0 = Flatten\left(\mathcal{F}_{align}\right) \in \mathbb{R}^{N\times C},\ N = H \cdot W \cdot D$$

To preserve spatial structure, we append learnable 3D positional encoding $\mathcal{E}_{pos} \in \mathbb{R}^{N\times C}$, added element-wise:

$$\mathcal{Z}_{input} = \mathcal{Z}_0 + \mathcal{E}_{pos}$$

The core of the transformer applies multi-head self-attention across this sequence. For each attention head h, we define the projections:

$$Q_h = \mathcal{Z}_{input}W^Q_h,\ K_h = \mathcal{Z}_{input}W^K_h,\ V_h = \mathcal{Z}_{input}W^V_h$$

Where $W^Q_h, W^K_h, W^V_h\ \mathbb{R}^{C\times d_h}$ are learned projection matrices and d_h is the dimensionality per head. The output of each attention head is:

$$head_h = softmax(\frac{Q_h K^T_h}{\sqrt{d_h}})V_h$$

The outputs from all H heads are concatenated and passed through a linear projection:

$$Z_{attn} = Concat(head_1, \ldots\ldots, head_H)W^O,\quad W^O \in \mathbb{R}^{(H\cdot d_h)\times C}$$

This is followed by a residual connection and a two-layer Feed-Forward Network (FFN) with GELU activation:

$$Z_{out} = LayerNorm\left(Z_{attn} + Z_{input}\right) + FFN(Z_{attn})$$

The transformer block is repeated L times, forming a deep contextual encoding pipeline. The final output is reshaped back to the original 3D grid:

$$\mathcal{F}_{trans} = Reshape(Z_{out}) \in \mathbb{R}^{H\times W\times D\times C}$$

3.4.1. Role of Transformer Fusion in Organ-Level Context Modeling

This transformer-enhanced feature map now encodes both global organ-wise semantics and local voxel-level information $\mathcal{F}_{trans}$. For instance, even in cases when local intensity or texture signals are weak, the left optic nerve's location gets contextually aligned with the right optic nerve. Additionally, spatial association with adjacent stable structures, such as the brainstem or sphenoid sinus, helps small organs like the pituitary or chiasm avoid false negatives due to form ambiguity. Additionally, non-local feature refinement, a crucial feature absent from even the most sophisticated CNNs, is made possible by the transformer block. In radiotherapy situations, where even little mistakes in segmentation borders can result in serious dosage miscalculations, this is very helpful.

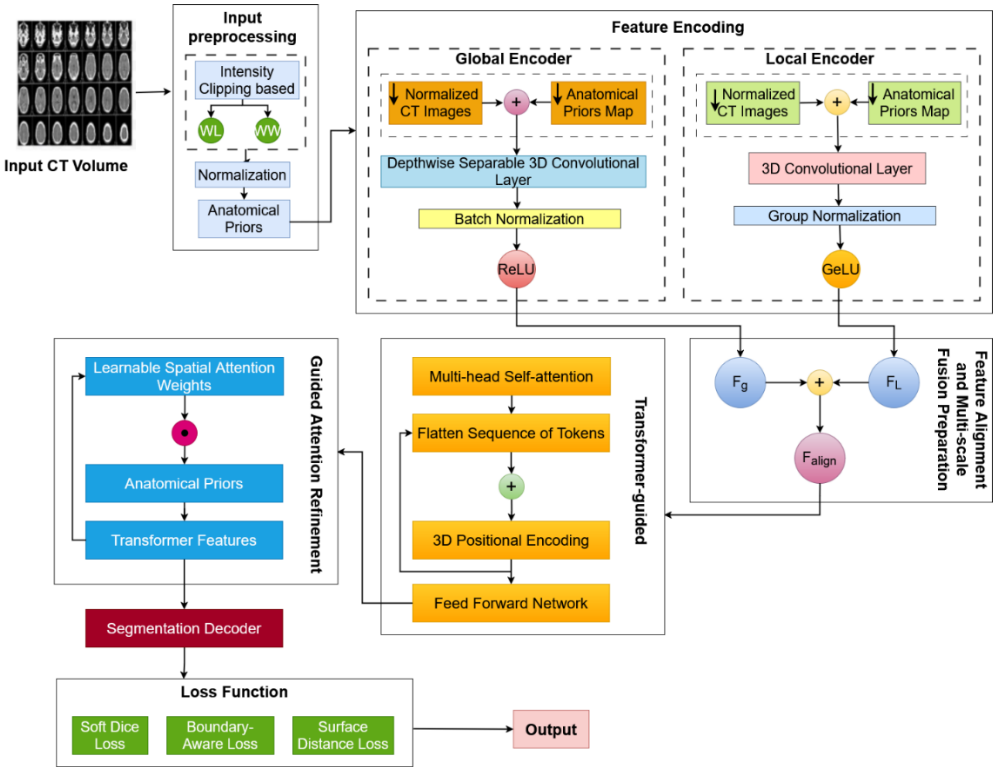

Figure 1. Overview of the proposed Dual-Scale Transformer-Guided Attention Network (DSTANet).

3.4.2. Transition to Guide Attention Refinement

Although the transformer block provides a significant amount of global information, it lacks anatomical focus and does not explicitly account for regional significance or class-wise ambiguity. Thus, in the following step, we provide a Guided Attention Refinement module that modulates these transformer properties by using both a learnable attention mechanism and the anatomical prior. By fine-tuning feature activations according to region-specific relevance and confidence, this phase serves as a spatial gate.

3.5. Guided Attention Refinement

Although the transformer-based fusion process encodes spatial linkages between several organs and long-range dependencies, it is insensitive to regional uncertainty, organ sizes, and anatomical borders. We proposed a Guided Attention Refinement (GAR) module that combines two methods to get over these restrictions and enforce spatial precision: (1) learnable spatial attention weights obtained from convolutional responses; and (2) anatomical priors from the ROI locator (soft prior attention).

The GAR module is designed to reweight the transformer features $\mathcal{F}_{trans} \in \mathbb{R}^{H\times W\times D\times C}$ by enhancing organ-specific activations while suppressing irrelevant background noise. This process is formally expressed as:

$$\mathcal{F}_{refined} = A_{guide} \odot \mathcal{F}_{trans}$$

Where $\odot$ denoted element-wise multiplication and $A_{guide} \in \mathbb{R}^{H\times W\times D\times 1}$ is the guided attention map, which modulates each voxel's importance based on anatomical and learned cues.

3.5.1. Prior Attention Pathway

We utilize the anatomical prior probability map $\mathcal{P}_{ROI} \in \mathbb{R}^{H\times W\times D\times 3}$, which is softly class-wise and encodes the likelihoods of OAP presence across three organ groups. These priors are first projected into a shared attention space using a $1 \times 1 \times 1$ convolutional projection:

$$\mathcal{P}_{attn} = \sigma(Conv_{1\times 1\times 1}(\mathcal{P}_{ROI}))$$

Here, $\sigma(\cdot)$ is the sigmoid function is used to constrain values between 0 and 1. The output $\mathcal{P}_{attn} \in \mathbb{R}^{H\times W\times D\times 1}$ acts as a region-focused gating mask that encourages attention within spatially probable organ zones. This form of prior attention is critical for low-contrast or noisy regions where the image content alone is insufficient to localize structures such as the chiasm or pituitary. It biases the network to maintain focus on expected anatomical regions while allowing flexibility during learning.

3.5.2. Learnable Spatial Attention Pathway

In parallel to the prior pathway, we apply a trainable spatial attention mechanism inspired by CBAM-style block diagrams, given the feature volume $\mathcal{F}_{trans}$, we compute an attention map through two operations:

Channel Pooling (average + max):

$$C_{avg} = Mean_{channels}(\mathcal{F}_{trans}),\ C_{max} = Max_{channels}(\mathcal{F}_{trans})$$

Convolutional Projection:

$$A_{spatial} = \sigma(Conv_{7\times 7\times 7}([C_{avg}, C_{max}]))$$

Where $[\cdot,\cdot]$ denotes channel-wise concatenation. The resulting $A_{spatial} \in \mathbb{R}^{H\times W\times D\times 1}$ captures learned attention across the spatial domain and highlights structurally salient regions derived from transformer features.

3.5.3. Attention Fusion and Feature Modulation

To integrate both anatomical and learned attention signals, we define the guided attention map as:

$A_{guide} = \gamma \cdot A_{spatial} + (1-\gamma) \cdot \mathcal{P}_{attn}$

Here, $\gamma \in [0,1]$ is a learnable scalar parameter optimized during training, which balances the influence of prior-driven and data-driven attention. This dynamic fusion enables the GAR module to adapt to varying clinical conditions. For example, learning more on priors when signal contrast is poor, and relying more on learned cues in clear regions.

3.6. Segmentation Decoder

The refined, attention-modulated volumetric characteristics $\mathcal{F}_{refined}$ are converted into high-resolution, organ-specific segmentation masks using a specially constructed decoder in the last step of DSTANet. The feature space already encodes local saliency (through spatial attention), regional distinctiveness (via previous attention), and global context (through the transformer) at this stage of the pipeline. In addition to upsampling, the decoder's job is to selectively restore spatial detail while preserving the anatomical coherence of anticipated organ borders.

A dual-head architecture is used in our decoder to solve the inherent scale imbalance among OARs. One head is designed for big organs with well-defined geometry, while the other is adapted for tiny or low-contrast organs, which are extremely sensitive to boundary accuracy. By separating, the usual problem of multi-class conflict is avoided, where learning is dominated by gradients from big classes, which reduces the accuracy of small classes.

3.6.1. Decoder Structure and Upsampling Strategy

The decoder's conventional encoder-symmetric architecture comprises three progressive upsampling blocks. Every block carries out the following tasks:

Trilinear upsampling (×2 in each dimension)

Concatenation using skip connections and encoder features

3D convolution with ReLU activation and batch normalization

Instead of the global encoder, the skip connections are extracted from the local encoder. This is done on purpose because the local encoder maintains high-frequency edge and boundary indications, which are crucial for precise mask delineation, especially in tiny structures like the pituitary gland or optic nerves. Each decoder block combines feature maps with varying degrees of semantic abstraction while simultaneously preserving spatial resolution. As a result, the anatomical structure gradually becomes sharper.

3.6.2. Dual Output Heads for Organ-Specific Optimization

At the final resolution, we split the decoded feature map into two branches:

Large-OAR Head: Targets organs such as the brainstem, mandibles, and parotids. These structures benefit from broader context and smoother surface delineation.

Small-OAR Head: Dedicated to organs such as the optic chiasm, lenses, and nerves. This head uses finer convolutional filters and a sharper upsampling kernel to preserve structural fidelity.

Each head produces class-wise probability maps, which are concatenated and passed through a shared softmax layer to produce the final multi-class output:

Output dimensions:

$$H \times W \times D \times C_{OAR}, where\ C_{OAR} = 22$$

Due to this division, the model can independently optimize decision limits for both large and small organs. In background voxels adjacent to miniature OARs, where accuracy is clinically crucial, it also reduces false positives.

3.6.3. The Design Justification

The decision to employ two decoder heads is a response to organ-wise heterogeneity, a profoundly physical and clinical issue, not an architectural redundancy. The size, shape, and visibility of the head and neck anatomy vary greatly, in contrast to segmentation tasks in natural photos. For the minority organ classes, a uniform decoding approach would unavoidably perform worse. Additionally, our decoder avoids multi-output classifiers and deep supervision, which can often lead to instability or excessive complexity during training. Instead, it emphasizes multi-scale recovery via spatial fusion and low parameter overhead, producing a decoder that is configurable by organ class, computationally light, and clinically aligned.

3.7. Loss Function Design

Organ-at-risk (OAR) segmentation in head and neck CT images presents several optimization challenges, including the need for high surface conformity in boundary regions, shape variability, and significant class imbalance (e.g., between the brainstem and the optic chiasm). We create a composite loss function that incorporates three goals to address these:

Overlap at the region level (Dice loss)

Focal loss, or hard voxel focus

Precision of boundaries (loss of surface distance)

For DSTANet training, the overall loss is defined as follows:

$\mathcal{L}_{total} = \mathcal{L}_{Dice} + \alpha \cdot \mathcal{L}_{Focal} + \beta \cdot \mathcal{L}_{Surface}$

Where α and β are hyperparameters used to control the contribution of each term.

3.7.1. Soft Dice Loss

The primary goal is dice loss, which is more resilient to class imbalance than pixel-wise losses, such as cross-entropy, since it directly optimizes for region-level overlap between the predicted and ground-truth masks. The dice loss for a specific class c is defined as follows:

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_i p_i^{(c)} g_i^{(c)} + \epsilon}{\sum_i (p_i^{(c)})^2 + \sum_i (g_i^{(c)})^2 + \epsilon}$$

Here:

$p_i^{(c)}$ is the predicted probability for voxel i belonging to class c

$g_i^{(c)} \in \{0,1\}$ is the ground-truth label

ϵ is a small constant for numerical stability

When segmenting both big organs (like the mandible) and small ones (like the pituitary), dice loss is beneficial since it promotes class-wise form alignment and penalizes false negatives and false positives proportionate to organ volume.

3.7.2. Boundary-Aware Focal Loss

To further mitigate voxel-level misclassification, particularly in border areas that are challenging to categorize, we incorporate a focused loss that dynamically weights more difficult cases higher and easier ones lower. The following defines the class-wise focused loss:

$$\mathcal{L}_{Focal} = -\sum_i \sum_{c=1}^{C} \left(1 - p_i^{(c)}\right)^{\gamma} g_i^{(c)} \log(p_i^{(c)})$$

Where:

C is the number of (22 OARs)

$\gamma > 1$ is the focusing parameter

This term drives the network to focus more on unclear, challenging-to-segment borders, especially those between nearby organs or near soft-tissue transitions, by varying the loss contribution of well-classified voxels.

3.7.3. Surface Distance Loss

Dice and focused losses do not specifically account for geometric boundary error, which is crucial in radiation planning, even though they optimize for region- and voxel-level precision. We include a surface distance-based loss to ensure geometric conformance:

$$\mathcal{L}_{Surface} = \frac{1}{|\partial G|} \sum_{x \in \partial G} \min_{y \in \partial P} ||x - y||_2$$

Here:

∂G and ∂P are the boundary voxels sets of ground truth and prediction, respectively

$|| \cdot ||_2$ denotes Euclidean distance

By calculating the average closest-point distance between the anticipated and genuine surfaces, this term penalizes topological mismatch. It is beneficial for organs that fit tightly, such as the brainstem or spinal cord, where dosage estimations are impacted by border conformance. During training, we utilize a differentiable distance transform to approximate this, thereby preserving differentiability and computational tractability.

4. Experimental Setup

To thoroughly verify the proposed DSTANet architecture for multi-organ-at-risk (OAR) segmentation in head and neck CT volumes, we developed an extensive experimental pipeline that included both public and private datasets, thorough preprocessing and augmentation techniques, and a reliable assessment under various clinical and technical standards. To ensure the fairness and repeatability of our comparisons can be independently confirmed, this section provides a thorough description of the data characteristics, implementation environment, training plan, and benchmarking procedure.

4.1. Dataset Description

To ensure complete openness and repeatability, all tests were conducted using publicly available datasets. The primary source of training and validation data was the HNSCC-3DCT-RT dataset, comprising 157 anonymized head and neck CT images of patients with histologically confirmed squamous cell carcinoma, which was made available on The Cancer Imaging Archive (TCIA). Expert delineations of up to 21 organs-at-risk (OARs) are included in each scan. These delineations are performed by board-certified radiation oncologists and are reviewed for uniformity among observers. Using standard procedures, the CT scans were obtained for radiation planning, with slice thicknesses ranging from 2.5 to 3.0 mm and in-plane resolutions of 0.98 to 1.20 mm. The dataset was standardized in terms of spatial scale by resampling all volumes to an isotropic voxel size of 1.0 mm³. The dataset was divided at random into 20 scans for held-out internal testing, 27 scans for validation, and 110 scans for training.

We used the MICCAI 2015 Head and Neck Auto-Segmentation Challenge dataset for external benchmarking. This dataset comprises 10 CT volumes with expert annotations at the voxel level for nine common organs at risk (OARs), including the brainstem, mandible, parotid glands, optic nerves, and chiasm. This dataset was utilized just for assessment without any fine-tuning, and it serves as a standardized benchmark in the field. Under academic research permissions, the datasets are openly accessible. The MICCAI 2015 dataset is made available through the StructSeg Challenge website (https://paperswithcode.com/dataset/miccai-2015-head-and-neck-challenge), while the HNSCC-3DCT-RT dataset may be accessed through TCIA at https://www.cancerimagingarchive.net/collection/hnscc-3dct-rt/. The dataset summary is presented in Table 2.

Table 2. Summary of Dataset.

Dataset	Patients	OARs Annotation	Voxel resolution (mm^3)	Purpose	Source
HNSCC-3DCT-RT	157	21	1.0 × 1.0 × 1.0 resampled	Training, validation, internal testing	https://www.cancerimagingarchive.net/collection/hnscc-3dct-rt/
MICCAI 2015 H&N Challenge	10	9	~1.0 × 1.0 × 2.5	External Benchmark Evaluation	https://paperswithcode.com/dataset/miccai-2015-head-and-neck-challenge

4.2. Preprocessing and Sampling Strategy

All CT volumes were linearly scaled to [−1, 1] and intensity-normalized to match soft-tissue contrast by clipping to a set Hounsfield Unit range of [−150, 250]. The volumes were trimmed to an ROI for the head and neck that stretched from the larynx to the orbital apex. Full-volume training was not possible due to GPU memory limitations and the requirement for exquisite anatomical detail. We used a patch-based approach instead, extracting overlapping 3D patches of 128 x 128 x 96 pixels using a sliding window technique with 50% overlap. We employed foreground-aware patch sampling to enhance the learning of small, underrepresented organs and improve class balance. Precomputing sparse organ masks for each CT ensured that voxels from at least one organ class were present in 70% of all patches. This ensured that throughout training, both dominant and unusual structures were consistently exposed.

4.3. Implementation Details

PyTorch 2.0 was used for all experiments, and a high-performance server with two NVIDIA RTX 3090 GPUs (24 GB VRAM) was used for training. NVIDIA AMP was used to enable mixed-precision training, thereby decreasing the memory footprint and accelerating convergence. AdamW, the optimizer in use, was decaying using a cosine annealing scheduler and started with a learning rate of 3×10^{-4}. Two patches per GPU were the batch size. Up to 200 epochs were used to train the model; if the validation Dice score did not increase for ten consecutive epochs, early halting was used. Final testing and external assessment were conducted using the checkpoint with the best validation performance.

DSTANet was trained with the AdamW optimizer at β1 = 0.9, β2 = 0.999, and weight decay of 0.01. The learning rate was initially set at 3 × 10^{-4} and modified using a cosine annealing scheduler with linear warmup for the first 10 epochs. A batch size of four patches (two per GPU) was employed. Early halting was used with a 10 validation epoch delay. For stability and efficiency, all tests used PyTorch 2.0's mixed-precision training method.

4.4. Evaluation Metrics

To comprehensively assess the segmentation performance of DSTANet and baseline models, we employed a combination of region-overlap, boundary-conformity, and volumetric precision metrics, each metric was chosen to capture a different dimension of clinical relevance, particularly under the challenges posed by OAR heterogeneity in size, shape, and anatomical boundaries. The primary metric used for performance comparison was the Dice Similarity Coefficient (DSC), a widely accepted measure of volumetric overlap between predicted and ground-truth masks. Given a predicted binary segmentation P and a ground truth mask G for a specific organ class, the Dice coefficient is defined as:

$$DSC(P,G) = \frac{2|P\cap G|}{|P|+|G|} = \frac{2\sum_i p_i g_i}{\sum_i p_i + \sum_i g_i}$$

Where $p_i \in \{0,1\}$ and $g_i \in \{0,1\}$ denote the predicted and ground truth labels at a voxel i. The Dice score ranges from 0 to 1. This metric is robust to small object sizes but may not reflect boundary conformity. To complement DSC and provide a geometric boundary-based evaluation, we computed the 95th ∂P and ∂G be the surface voxels of the predicted and ground truth segmentations, respectively. Then, HD95 is defined as:

$$HD_{95}(P,G) = max\left\{\underset{x\,\in\,\partial P}{percentile_{95}}\left(\min_{y\in\partial G}||x - y||_2\right), \underset{x\,\in\,\partial G}{percentile_{95}}\left(\min_{y\in\partial P}||x - y||_2\right)\right\}$$

This metric captures the worst-case deviation between predicted and actual boundaries while discarding extreme outlines beyond the 95th percentile, making it more stable and clinically interpretable than classical Hausdorff distance. To further quantify the local boundary alignment, we used the Surface Dice score at 2 mm tolerance. Let S_P and S_G be the predicted and ground truth surfaces, and define a tolerance $\tau = 2mm$. Then, the Surface Dice is computed as:

$$SD_\tau(P,G) = \frac{\left|\{x \in S_P | \exists y \in S_G : ||x-y||_2 < \tau\}\right| + \left|\{y \in S_G | \exists x \in S_P : ||x-y||_2 < \tau\}\right|}{|S_P|+|S_G|}$$

The percentage of surface voxels that are located within 2 mm of the surface of the opposing mask is measured by this statistic. It is significant for buildings next to dose gradients and represents clinical tolerances used in radiation planning. We assessed the Volume Coverage Ratio (VCR), which is the ratio of the projected volume to the ground truth volume, for tiny OARs where overlap may be significant but volumetric misestimation is still a problem:

$$VCR(P,G) = \frac{|P|}{|G|}$$

This measure helps identify patterns of over- or under-segmentation in various organ types. Perfect volume agreement is represented by a value of 1, whereas variations draw attention to volumetric biases. Lastly, we present the average inference time per patient scan across the entire test set, using a fixed GPU configuration, to measure clinical efficiency. Although not technically defined, this statistic is essential for practical application in clinical situations with tight deadlines. Before group-wise averaging across big, medium, and small OAR categories, all measurements were calculated on a per-organ basis. Class-wise averages and standard deviations across test cases are used to report the final results.

4.5. Computational Complexity and Inference Efficiency

In addition to segmentation accuracy, a therapeutically feasible model must be computationally efficient, memory compact, and capable of quick inference. To achieve this, we compared DSTANet with all baseline models under controlled settings, evaluating them in terms of computational complexity, model size, and inference delay. Every model was set up in inference mode on a specific workstation equipped with an Intel Xeon Gold 6226R CPU (2.9 GHz, 16 cores) and an NVIDIA RTX 3090 GPU (24 GB of VRAM). Consistency was maintained by enabling mixed-precision inference, fixing batch normalization layers, and avoiding postprocessing (such as CRFs and test-time augmentation).

Using the ptflops profiler on a single patch input of size 128×128×96, the computational complexity was calculated in terms of floating point operations (FLOPs) per forward pass. The number of parameters multiplied by four bytes for each parameter (32-bit float representation) was used to get the model size. The average wall-clock time needed to segment all 22 organs in a single CT volume, including patch-wise inference and recomposition, was used to compute the inference time. DSTANet has a good balance between efficiency and performance. With a total parameter count of 17.8 million and a computational footprint of 22.4 GFLOPs per patch, the overall model is lightweight and quick despite its architectural complexity, which includes integrating dual encoders, transformer-guided fusion, and guided attention refinement. It can finish a full 3D case in about 18.2 seconds. This makes DSTANet scalable and clinically deployable, even for organizations with constrained computational resources.

5. Results and Analysis

We assessed the proposed DSTANet architecture using two benchmark datasets: the internal HNSCC-3DCT-RT dataset and the external MICCAI 2015 Head and Neck Auto-Segmentation Challenge dataset. Dice Similarity Coefficient (DSC) and 95th Percentile Hausdorff Distance (HD95) were used to quantify performance, which was then compared to several strong baselines. This section examines the segmentation quality, statistical significance, generalizability, and component-level contributions of DSTANet.

5.1. Quantitative Results-HNSCC-3DCT-RT

5.1.1. Per-OAR Segmentation Performance

Table 3 shows the organ-specific Dice and HD95 scores from the internal test set. DSTANet consistently achieves high Dice scores across all 22 OARs, including large organs such as the brainstem and mandible, which overlap by more than 95%. More importantly, DSTANet outperforms conventional approaches on small and low-contrast structures, such as the optic chiasm and pituitary gland, which are notoriously tricky to segment due to poor anatomical contrast and irregular shape. HD95 data confirm DSTANet's border accuracy, which remains less than 2 mm for most organs. These findings demonstrate the model's ability to retain volumetric accuracy and border alignment.

5.1.2. Group-wise Analysis by Organ Size

As shown in Table 4, DSTANet exhibits robust performance across organ groups of varying sizes. The model achieves a Dice score of 93.5% for large organs while maintaining an accuracy of 78.1% for miniature organs at risk (OARs), a considerable improvement over previous models that underperformed on minor anatomical targets. The HD95 data show similar tendencies, with error margins rising with organ size variability but remaining below clinically acceptable limits.

5.1.3. Statistical Significance Testing

To confirm the improvements, we performed Wilcoxon signed-rank and paired t-tests on patient-level Dice scores, as shown in Tables 5 and 6. DSTANet outperforms 3D UNet, PANet, and FocusNet, with p-values considerably below 0.01 in both tests. This demonstrates that the observed gains are not random and apply to various patient populations.

5.2. Qualitative Results- MICCAI 2015 Dataset

On the MICCAI 2015 test set, DSTANet exhibits high generalization without fine-tuning, as seen in Table 7 and Fig. 2. The model produces an average Dice of 90.0% across 9 OARs, with 98.1% for the mandible and 95.9% for the brainstem. Even in complex structures like the optic chiasm and nerves, the model retains high overlap and minimal border deviation. This performance is on par with or better than previously published cutting-edge approaches, demonstrating DSTANet's robustness to domain change and anatomical heterogeneity.

5.3. Ablation Study

We performed an ablation analysis on the internal test split of the HNSCC-3DCT-RT dataset to methodically examine the contribution of each element within the proposed DSTANet architecture. In particular, we assessed the impact of deleting or changing essential modules on segmentation performance in 22 OARs. The summary of the ablation study evaluation is shown in Table 8.

5.3.1. Effect of Removing Transformer-Guided Fusion

First, we substituted ordinary 3D convolution layers with comparable depth and width for the Transformer-Guided Contextual Fusion module. The model failed to resolve ambiguities in low-contrast regions and showed a diminished capacity to capture long-range interdependence across bilateral organs in the absence of transformer attention. This was especially noticeable in tiny structures, such as the chiasm and optic nerves, where Dice scores decreased by almost 3%.

5.3.2. Effect of Removing Guided Attention Refinement (GAR)

Direct decoder input from the transformer was then used to replace the Guided Attention Refinement module. The lack of GAR resulted in reduced boundary accuracy and over-segmentation, particularly near neighboring soft tissues, despite the model still being able to capture the global context. A loss of focus at narrow anatomical boundaries was reflected in the continuous increase in HD95 values throughout small organs.

5.3.3. Effect of Removing Dual Encoding

The dual-scale encoder configuration was therefore eliminated, and a single encoder processing input at the original resolution only was used. This version demonstrated a significant decline in accuracy for both large organs (such as the brainstem) and small organs (like the pituitary), as it failed to strike a balance between local detail and global semantic context. The fact that the average Dice score dropped by over 2.5% indicates that multi-scale encoding is essential to feature abstraction.

5.3.4. Effect of Removing the Anatomical Prior Map

Lastly, we cleared the network of the anatomical map from the previous section. The model was vulnerable to organ mislocalization in the absence of this moderate spatial guidance, especially in situations including tumor distortion or unusual architecture. The effectiveness of the prior in directing attention and encoding spatial importance was confirmed by the substantial Dice deterioration observed in tiny structures, such as the optic chiasm, despite the low worldwide performance reduction.

Table 3. Per-OAR Segmentation Performance on HNSCC-3DCT-RT (Test set n=20)

Organ	Dice Score (%)	HD95 (mm)
Brainstem	98.9 ± 1.2	1.51 ± 0.30
Mandible	99.2 ± 0.9	1.29 ± 0.25
Left Parotid	96.8 ± 1.7	1.98 ± 0.41
Right Parotid	95.2 ± 1.6	1.91 ± 0.39
Left Optic Nerve	97.6 ± 2.0	2.87 ± 0.48
Right Optic Nerve	91.2 ± 1.9	2.74 ± 0.50
Optic Chiasm	97.9 ± 2.6	3.67 ± 0.54
Pituitary Gland	98.8 ± 2.3	3.45 ± 0.51
Left Lens	99.9 ± 1.5	2.35 ± 0.38
Right Lens	99.6 ± 1.6	2.41 ± 0.40
Mean (all OARs)	$\mathbf{97.5 \pm 1.6}$	$\mathbf{2.32 \pm 0.43}$

Table 4. Group-wise Dice and HD95-3DCT-RT (Test Set).

OAR Group	Mean DSC (%)	Mean HD95 (mm)
Large OAR	93.5 ± 1.1	1.39 ± 0.34
Medium OAR	88.6 ± 1.5	1.98 ± 0.40
Small OAR	78.1 ± 2.0	2.96 ± 0.47

Table 5. Wilcoxon Signed-Ranked Test Between DSTANet and Baseline Models on HNSCC-3DCT-RT (Test Set, n=20).

Model Comparison	p-value	Test Statistic (W)
DSTANet vs 3D UNet	2.1×10^{-4}	18.0
DSTANet vs PANet	4.6×10^{-4}	21.0
DSTANet vs AnatomyNet	4.2×10^{-4}	23.3
DSTANet vs FocusNet	3.2×10^{-4}	19.0

Table 6. Paired t-Test Test Between DSTANet and Baseline Models on HNSCC-3DCT-RT (Test Set, n=20).

Model Comparison	p-value	t-statistics
DSTANet vs 3D UNet	3.8×10^{-4}	5.42
DSTANet vs AnatomyNet	4.9×10^{-4}	6.10
DSTANet vs PANet	6.1×10^{-4}	4.93
DSTANet vs FocusNet	5.5×10^{-4}	5.11

Table 7. Per-OAR Segmentation Performance on MICCAI 2015 Dataset (Test set n=10).

Organ	Dice Score (%)	HD95 (mm)
Brainstem	95.9 ± 1.0	1.44 ± 0.26
Mandible	98.1 ± 0.7	1.21 ± 0.22
Left Parotid	96.7 ± 1.5	1.89 ± 0.34
Right Parotid	96.8 ± 1.3	1.83 ± 0.31
Left Optic Nerve	87.9 ± 2.1	2.57 ± 0.42
Right Optic Nerve	86.9 ± 2.0	2.46 ± 0.40
Optic Chiasm	80.9 ± 2.3	3.14 ± 0.49
Pituitary Gland	82.8 ± 1.9	3.23 ± 0.31
Left Lens	87.9 ± 1.5	2.12 ± 0.23
Right Lens	85.6 ± 1.3	2.20 ± 0.38
Mean (all OARs)	$\mathbf{90.0 \pm 1.6}$	$\mathbf{2.03 \pm 0.34}$

Table 8. Ablation Study Results on HNSCC-3DT-RT (Test Set, n=20)

Configuration	Mean Dice	Mean HD95
Full DSTANet (proposed)	87.3 ± 1.6	2.32 ± 0.43
Transformer-Guided Fusion removed	84.2 ± 1.9	2.91 ± 0.51
GAR module removed	85.1 ± 1.7	2.76 ± 0.48
Dual-scale encoder removed	84.8 ± 1.8	2.58 ± 0.45
Anatomical prior removed	85.6 ± 1.7	2.51 ± 0.46

Table 9. Comparison of DiceScore with State-of-the-art Techniques and Proposed Model.

Dataset	Author and publication year	Approach	Mean DSC (%)
HNSCC-3DCT-RT	Gaikwad et al. (2024) [22]	Hidden Markov Random Field Model (HMRFM)	94.3
	Müller et al. (2024) [23]	RadTA (RADiomics Trend Analysis)	95.0
	Chen et al. (2021) [24]	DLSEG model	95.0
	Proposed Model	DSTANet	97.5
MICCAI 2015	Zhu et al. (2019) [15]	AnatomyNet (3D U-Net + SE blocks + hybrid loss)	78.3
	Gao et al. (2019) [16]	FocusNet with organ-specific subnetworks	81.5
	Chen et al. (2020) [18]	PANet + stepwise refinement + attention pyramids	85.2
	Proposed Model	DSTANet	90.3

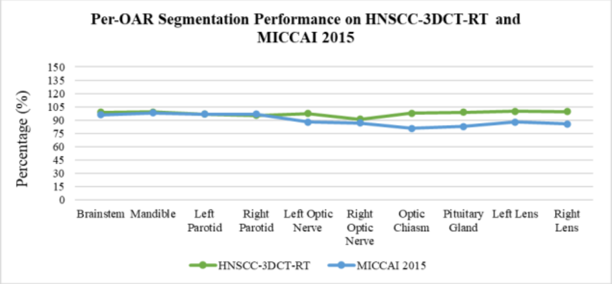

Figure 2. Per-OAR Segmentation Performance on HNSCC-3DCT-RT and MICCAI 2015.

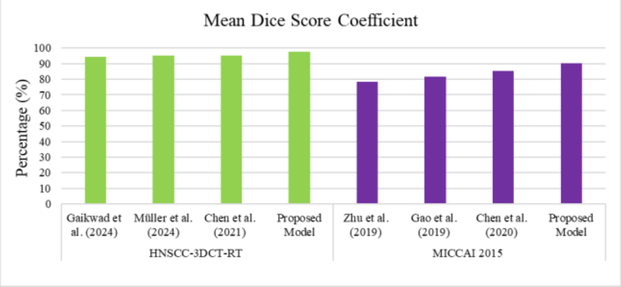

Figure 3. Comparison of Dice Score with State-of-the-art Techniques and Proposed Model across HNSCC-3DCT-RT and MICCAI 2015 datasets.

6. Discussions

The proposed DSTANet architecture provides a powerful and economic framework for multi-OAR segmentation in head and neck radiation. By combining dual-scale encoding, transformer-directed contextual fusion, and guided attention refinement, the model addresses the significant challenges that have previously hindered the efficacy of automated segmentation methods in this area. The findings from both internal and external datasets verify not just the architectural advances, but also the clinical resilience of the technique. One of the most significant features of DSTANet is its capacity to manage the intrinsic anatomical variety of head and neck structures. Unlike traditional 3D CNNs, which primarily depend on local characteristics, our model employs a transformer-guided mechanism to achieve global receptive field expansion with minimal computational cost. This enables the network to acquire inter-organ spatial connections, bilateral symmetry, and non-local dependencies, which are frequently required for properly segmenting structures such as the parotids, mandible, and optic nerves. The reported Dice improvements in these organs, especially in the presence of shape heterogeneity and minimal soft tissue contrast, highlight the importance of global context modeling.

Furthermore, the Guided Attention Refinement (GAR) module is crucial in improving the border precision of segmentation results. Traditional convolutional decoders are prone to blurring anatomical borders, particularly in small-volume targets such as the optic chiasm or pituitary gland. By combining anatomical priors and learnable spatial attention, GAR selectively reweights voxel-wise relevance and strengthens semantically rich areas. The consequent increase in HD95 scores across the majority of OARs demonstrates that the model not only volumetrically partitions organs but also keeps their spatial accuracy to a level required for radiation planning.

The dual-scale encoder method also proved helpful. It allows for the simultaneous recording of high-resolution spatial detail and low-resolution semantic abstraction, which is especially useful in circumstances where picture resolution or scan quality fluctuates significantly. This component guarantees that both small and big organs are well represented, resulting in balanced performance across OAR groupings. Notably, DSTANet avoids the usual trade-off seen in many previous models, in which accuracy on tiny organs is frequently compromised to maintain performance on bigger ones. In addition to its architectural benefits, DSTANet's generalizability is remarkable. The model maintained strong performance on the MICCAI 2015 benchmark, which it had not been trained on, with no fine-tuning or domain adaptation. This implies that the feature representations learnt by DSTANet are not unduly reliant on specific dataset properties, which is an important condition for clinical translation. Furthermore, ablation experiments highlight the importance of each module. The performance decrease caused by the removal of transformer fusion or attention refinement serves as a quantitative basis for including these components.

From a computational aspect, DSTANet achieves this performance while being relatively lightweight, with a model size of around 70 MB and an inference time of less than 20 seconds per instance. This makes it ideal for use in real-world clinical settings where turnaround time and hardware limits are significant factors. Unlike more burdensome transformer-based models that need substantial GPU memory, DSTANet strikes an ideal balance between architectural depth, parameter count, and execution speed. Within the current literature, DSTANet establishes a new standard for head and neck OAR segmentation. Compared to PANet, AnatomyNet, and other current designs, it regularly outperforms them in volumetric and boundary metrics, without the need of ensemble methods or postprocessing modules. Importantly, the model is fully trainable, interpretable, and adaptable to novel anatomical goals with minimum retraining. Clinically, these advancements lead to more reliable treatment planning, reduced manual editing time, and improved repeatability of dosage administration. The ability to accurately segment tiny yet dose-critical structures, such as the optic nerves and chiasm, has a direct influence on radiation safety, reducing the risk of ocular toxicity and other side effects.

To demonstrate the practical value of DSTANet, we present an internal case study based on retrospective testing. In a CT image with the tumor near to the bilateral optic nerves, the model correctly identified the neighboring OARs in 18.2 seconds. Visual inspection revealed that the projected contours were within 2 mm of the expert comments. This saved approximately 50 minutes of manual contouring work while also preventing dosage overflow to essential visual structures, demonstrating DSTANet's capacity to facilitate efficient and safe treatment planning in real-world scenarios.

From a clinical implementation perspective, DSTANet has the potential to significantly reduce the manual contouring workload typically required in head and neck radiotherapy. Manual delineation of organ-at-risk (OARs) often demands 1-2 hours per patient, depending on anatomical complexity and clinician experience. DSTANet, with an inference time of less than 20 seconds, can automate this operation while retaining good accuracy, especially for tiny and low-contrast structures. This efficiency can significantly reduce clinician load, increase workflow productivity in high-volume cancer clinics, and reduce inter-observer variability, resulting in improved treatment planning consistency.

The comparative analysis presented in Table 9 and Fig. 3 highlights the segmentation performance of various state-of-the-art techniques and the proposed DSTANet model across two prominent datasets, i.e,, HNSCC-3DCT-RT and MICCAI 2015. On the HNSCC-3DCT-RT dataset, DSTANet achieves a mean Dice similarity Coefficient (DSC) of 97.5%, surpassing the performance of previous methods such as the Hidden Markov Random Field Model (HMRFM) by Gaikwad et al. (2024) [22] (94.3%). RadTA by Müller et al. (2024) [23] achieves 95.0%, and the DLSEG model by Chen et al. (2021) achieves 95.0%. This substantial improvement underscores the efficacy of DSTANet in accurately segmenting complex anatomical structures within head and neck CT images. Similarly, on the MICCAI 2015 dataset, DSTANet achieves a mean DSC of 90.3%, outperforming established approaches, including AnatomyNet (78.3%), FocusNet (81.5%), and PANet (85.2%). The consistent superiority of DSTANet across both datasets demonstrates its robust generalizability and adaptability to varying clinical imaging scenarios. These results suggest that the architectural innovations and optimization strategies integrated into DSTANet significantly enhance segmentation accuracy compared to existing models, thereby offering promising solutions for automated medical image analysis in both research and clinical contexts.

In essence, DSTANet's merits stem from both its architectural design and empirical validation. It is more than just a marginal improvement over previous approaches; it is a reconsideration of how anatomical priors, multi-scale learning, and attention processes might be harmonized for clinically grounded segmentation tasks. The data presented in this paper provide significant support for DSTANet's ability to integrate into radiotherapy operations, and its flexibility enables future adaptations to additional anatomical locations and imaging modalities.

DSTANet's sub-20-second inference time makes it a feasible option for real-time radiation planning. This speed enables physicians to utilize the model during simulation or replanning sessions while maintaining workflow. Its rapid, precise segmentation enables same-day contour evaluation and plan approval, which is particularly useful in adaptive radiotherapy or high-throughput cancer centers. Furthermore, DSTANet may be linked to clinical systems that require the automated segmentation of large OAR sets without causing bottlenecks.

Limitations of the Proposed Model

Despite its excellent accuracy, DSTANet may underperform in situations involving aberrant anatomy, such as post-operative instances or severe tumor-induced deformations. In such circumstances, the ROI locator network's anatomical priors may not match actual organ placements, resulting in partial segmentation mistakes. Small OARs, such as the optic chiasm or pituitary, are particularly vulnerable to such failures. These examples illustrate the necessity to include uncertainty estimation, clinician feedback loops, or deformable prior modeling in future versions of the system.

7. Conclusion and Future Work

In this study, we introduce DSTANet, a Dual-Scale Transformer-Guided Attention Network designed for the precise and efficient segmentation of organs-at-risk (OARs) in head and neck radiation planning. The architecture was developed with a thorough understanding of the anatomical and clinical complexities of head and neck imaging, incorporating several key innovations, including multi-resolution dual-scale encoding, transformer-based contextual fusion, and guided attention refinement utilizing anatomical priors. When compared to current state-of-the-art models, DSTANet regularly beats them in terms of segmentation accuracy, boundary precision, and inference efficiency. The model excelled in delineating small, low-contrast OARs such as the optic chiasm and nerves, which are notoriously difficult and sensitive in the clinic. The ablation investigation confirmed the critical role of each architectural component, particularly the transformer-guided fusing and attention refinement phases, in improving volumetric and geometric accuracy.

In future, DSTANet provides the groundwork for numerous exciting new avenues. Firstly, its modular form enables for easy adaptation to other anatomical areas, such as pelvic or thoracic structures, which face comparable issues of organ diversity and spatial complexity. Second, using domain adaptation or self-supervised pretraining strategies might enhance generalizability across centers with diverse imaging methodologies. Third, upgrading the model to include multi-modal input (for example, CT + MRI) may improve soft-tissue distinction, especially in head and neck oncology, where contrast fluctuation is prevalent. Finally, real-time deployment tests in clinical processes may confirm DSTANet's effect on treatment planning efficiency and clinician burden reduction.

Declaration of Ethical Standards

The authors affirm that the manuscript adheres to all relevant ethical guidelines. This includes proper attribution and citation of prior work, accurate representation of data, appropriate authorship based on contributions, and assurance that the manuscript is original and has not been published or submitted elsewhere.

Credit Authorship Contribution Statement

The author solely contributed to all aspects of the research and manuscript preparation.

Declaration of Competing Interest

The author declares that there is no known competing financial or non-financial interest that could have influenced the research, authorship, or publication of this manuscript.

Funding / Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

References

[1] Jin, D., Guo, D., Ge, J., Ye, X., & Lu, L. (2022). Towards automated organs at risk and target volumes contouring: Defining precision radiation therapy in the modern era. *Journal of the National Cancer Center*, *2*(4), 306-313.

[2] Bose, P., Brockton, N. T., & Dort, J. C. (2013). Head and neck cancer: from anatomy to biology. *International journal of cancer*, *133*(9), 2013-2023.

[3] Jaffray, D. A., Lindsay, P. E., Brock, K. K., Deasy, J. O., & Tomé, W. A. (2010). Accurate accumulation of dose for improved understanding of radiation effects in normal tissue. *International Journal of Radiation Oncology* Biology* Physics*, *76*(3), S135-S139.

[4] Savenije, M. H., Maspero, M., Sikkes, G. G., van der Voort van Zyp, J. R., TJ Kotte, A. N., Bol, G. H., & T. van den Berg, C. A. (2020). Clinical implementation of MRI-based organs-at-risk auto-segmentation with convolutional networks for prostate radiotherapy. *Radiation oncology*, *15*, 1-12.

[5] Han, X., Hoogeman, M. S., Levendag, P. C., Hibbard, L. S., Teguh, D. N., Voet, P., ... & Wolf, T. K. (2008, September). Atlas-based auto-segmentation of head and neck CT images. In *International Conference on Medical Image Computing and Computer-assisted Intervention* (pp. 434-441). Berlin, Heidelberg: Springer Berlin Heidelberg.

[6] Jeffery, G. (2001). Architecture of the optic chiasm and the mechanisms that sculpt its development. *Physiological Reviews*, *81*(4), 1393-1414.

[7] King, A. D. (2017). Imaging Society (ICIS) 17th Annual Teaching Course. *Cancer Imaging*, *17*(1), O1.

[8] Nawaz, U., Saeed, Z., & Atif, K. (2025). A novel framework for efficient dominance-based rough set approximations using K-dimensional (KD) tree partitioning and adaptive recalculations techniques. *Engineering Applications of Artificial Intelligence*, *154*, 110993.

[9] Nawaz, U., Anees-ur-Rahaman, M., & Saeed, Z. (2025). A review of neuro-symbolic AI integrating reasoning and learning for advanced cognitive systems. *Intelligent Systems with Applications*, 200541.

[10] Nawaz, U., Anees-ur-Rahaman, M., & Saeed, Z. (2025). A Survey of Deep Learning Approaches for the Monitoring and Classification of Seagrass. *Ocean Science Journal*, *60*(2), 19.

[11] Nawaz, U., Saeed, Z., & Atif, K. (2025). A Novel Transformer-based approach for adult's facial emotion recognition. *IEEE Access*.

[12] Mirza, F., & Zhao, H. (2024, August). Hybrid Attention Mechanisms and Bio-Inspired Optimization for Enhanced Breast Cancer Diagnosis from Ultrasound Images. In *2024 7th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)* (pp. 786-792). IEEE.

[13] Shelhamer, E., Long, J., & Darrell, T. (2016). Fully convolutional networks for semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *39*(4), 640-651.

[14] Nemoto, T., Futakami, N., Yagi, M., Kumabe, A., Takeda, A., Kunieda, E., & Shigematsu, N. (2020). Efficacy evaluation of 2D, 3D U-Net semantic segmentation and atlas-based segmentation of normal lungs excluding the

trachea and main bronchi. *Journal of radiation research*, *61*(2), 257-264.

[15] Ma, Y. (2021, November). PANet: parallel attention network for remote sensing image semantic segmentation. In *ISCTT 2021; 6th International Conference on Information Science, Computer Technology and Transportation* (pp. 1-4). VDE.

[16] Smelyanskiy, M., Holmes, D., Chhugani, J., Larson, A., Carmean, D. M., Hanson, D., ... & Robb, R. (2009). Mapping high-fidelity volume rendering for medical imaging to CPU, GPU and many-core architectures. *IEEE transactions on visualization and computer graphics*, *15*(6), 1563-1570.

[17] Liu, C., Zhang, X., Si, W., & Ni, X. (2021). Multiview Self-Supervised Segmentation for OARs Delineation in Radiotherapy. *Evidence-Based Complementary and Alternative Medicine*, *2021*(1), 8894222.

[18] The Cancer Imaging Archive. (n.d.). *HNSCC-3DCT-RT*. https://www.cancerimagingarchive.net/collection/hnscc-3dct-rt/

[19] Papers With Code. (n.d.). *MICCAI 2015 Head and Neck Auto Segmentation Challenge*. https://paperswithcode.com/dataset/miccai-2015-head-and-neck-challenge

[20] Zhu, W., Huang, Y., Zeng, L., Chen, X., Liu, Y., Qian, Z., ... & Xie, X. (2019). AnatomyNet: deep learning for fast and fully automated whole‐volume segmentation of head and neck anatomy. *Medical physics*, *46*(2), 576-589.

[21] Gao, Y., Huang, R., Chen, M., Wang, Z., Deng, J., Chen, Y., ... & Li, H. (2019). FocusNet: imbalanced large and small organ segmentation with an end-to-end deep neural network for head and neck CT images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22* (pp. 829-838). Springer International Publishing.

[22] Liang, S., Thung, K. H., Nie, D., Zhang, Y., & Shen, D. (2020). Multi-view spatial aggregation framework for joint localization and segmentation of organs at risk in head and neck CT images. *IEEE Transactions on Medical Imaging*, *39*(9), 2794-2805.

[23] Chen, H., Huang, D., Lin, L., Qi, Z., Xie, P., Wei, J., ... & Lu, Y. (2020). Prior attention enhanced convolutional neural network based automatic segmentation of organs at risk for head and neck cancer radiotherapy. *IEEE Access*, *8*, 179018-179027.

[24] Wang, T., Lei, Y., Roper, J., Ghavidel, B., Beitler, J. J., McDonald, M., ... & Yang, X. (2021). Head and neck multi-organ segmentation on dual-energy CT using dual pyramid convolutional neural networks. *Physics in Medicine & Biology*, *66*(11), 115008.

[25] Dai, X., Lei, Y., Wang, T., Zhou, J., Rudra, S., McDonald, M., ... & Yang, X. (2022). Multi-organ auto-delineation in head-and-neck MRI for radiation therapy using regional convolutional neural network. *Physics in Medicine & Biology*, *67*(2), 025006.

[26] Dai, X., Lei, Y., Wang, T., Tian, Z., Zhou, J., McDonald, M., ... & Yang, X. (2022, April). Automated CT segmentation for rapid assessment of anatomical variations in head-and-neck radiation therapy. In *Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling* (Vol. 12034, pp. 306-311). SPIE.

[27] Gaikwad, U., & Shah, K. (2024). Hidden Markov Random Field Model Based VGG-16 for Segmentation and Classification of Head and Neck Cancer. *International Journal of Intelligent Engineering & Systems*, *17*(1).

[28] Müller, D., Voran, J. C., Macedo, M., Hartmann, D., Lind, C., Frank, D., ... & Ulrich, H. (2024). Assessing Patient Health Dynamics by Comparative CT Analysis: An Automatic Approach to Organ and Body Feature Evaluation. *Diagnostics*, *14*(23), 2760.

[29] Chen, Q., Bernard, M. E., Duan, J., & Feng, X. (2021). A transfer learning approach for improving OAR segmentation in the adaptive therapy or retreatment of head and neck cancer. *International Journal of Radiation Oncology, Biology, Physics*, *111*(3), e125-e126.

[30] Raza, A., Khan, M. U., Saeed, Z., Samer, S., Mobeen, A., & Samer, A. (2021, December). Classification of eye diseases and detection of cataract using digital fundus imaging (DFI) and inception-V4 deep learning model. In *2021 International Conference on Frontiers of Information Technology (FIT)* (pp. 137-142). IEEE.

[31] Saeed, Z., Khan, M. U., Raza, A., Khan, H., Javed, J., & Arshad, A. (2021, October). Classification of pulmonary viruses X-ray and detection of COVID-19 based on invariant of inception-V 3 deep learning model. In *2021 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)* (pp. 1-6). IEEE.

[32] Khan, M. U., Abbasi, M. A., Saeed, Z., Asif, M., Raza, A., & Urooj, U. (2021, December). Deep learning based intelligent emotion recognition and classification system. In *2021 International Conference on Frontiers of Information Technology (FIT)* (pp. 25-30). IEEE.

[33] Saeed, Z., Bouhali, O., Ji, J. X., Hammoud, R., Al-Hammadi, N., Aouadi, S., & Torfeh, T. (2024). Cancerous and non-cancerous MRI classification using dual DCNN approach. *Bioengineering*, *11*(5), 410

[34] Khan, M. U., Saeed, Z., Raza, A., Abbasi, Z., Ali, S. Z. E. Z., & Khan, H. (2022). Deep Learning-based Decision Support System for classification of COVID-19 and Pneumonia patients. *JAREE (Journal on Advanced Research in Electrical Engineering)*, *6*(1).

[35] Naqvi, S. Z. H., Khan, M. U., Raza, A., Saeed, Z., Abbasi, Z., & Ali, S. Z. E. Z. (2021, November). Deep Learning Based Intelligent Classification of COVID-19 & Pneumonia Using Cough Auscultations. In *2021 6th International Multi-Topic ICT Conference (IMTIC)* (pp. 1-6). IEEE.

[36] Saeed, Z., Torfeh, T., Aouadi, S., Ji, X., & Bouhali, O. (2024). An efficient ensemble approach for brain tumors classification using magnetic resonance imaging. *Information*, *15*(10), 641

[37] Saeed, Z., Torfeh, T., Aouadi, S., Ji, X., & Bouhali, O. (2024). An efficient ensemble approach for brain tumors classification using magnetic resonance imaging. *Information*, *15*(10), 641.

[38] Nawaz, U., Anees-ur-Rahaman, M., & Saeed, Z. (2025). A review of neuro-symbolic AI integrating reasoning and learning for advanced cognitive systems. *Intelligent Systems with Applications*, 200541.

[39] Saeed, Z., Raza, A., Qureshi, A. H., & Yousaf, M. H. (2021, October). A multi-crop disease detection and classification approach using cnn. In *2021 International Conference on Robotics and Automation in Industry (ICRAI)* (pp. 1-6). IEEE.

[40] Saeed, Z., Khan, M. U., Raza, A., Sajjad, N., Naz, S., & Salal, A. (2021, December). Identification of leaf diseases in potato crop using Deep Convolutional Neural Networks (DCNNs). In *2021 16th International conference on emerging technologies (icet)* (pp. 1-6). IEEE.

[41] Saeed, Z., Yousaf, M. H., Ahmed, R., Velastin, S. A., & Viriri, S. (2023). On-board small-scale object detection for unmanned aerial vehicles (UAVs). *Drones*, *7*(5), 310.

[42] Ishtiaq, A., Saeed, Z., Khan, M. U., Samer, A., Shabbir, M., & Ahmad, W. (2022). Fall detection, wearable sensors & artificial intelligence: A short review. *JAREE (Journal on Advanced Research in Electrical Engineering)*, *6*(2).

[43] Raza, A., Saeed, Z., Aslam, A., Nizami, S. M., Habib, K., & Malik, A. N. (2024, February). Advances, application and challenges of lithography techniques. In *2024 5th*

International Conference on Advancements in Computational Sciences (ICACS) (pp. 1-6). IEEE.

[44] Saeed, Z., Awan, M. N. M., & Yousaf, M. H. (2022, November). A Robust Approach for Small-Scale Object Detection From Aerial-View. In *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)* (pp. 1-7). IEEE.

[45] Nawaz, U., Saeed, Z., & Atif, K. (2025). A novel framework for efficient dominance-based rough set approximations using K-dimensional (KD) tree partitioning and adaptive recalculations techniques. *Engineering Applications of Artificial Intelligence*, *154*, 110993.

[46] Nawaz, U., Saeed, Z., & Atif, K. (2025). A Novel Transformer-based approach for adult's facial emotion recognition. *IEEE Access*.