ITC 3/55
Information Technology and Control
Vol. 55 / No. 3/ 2026
pp. 1206-1226
DOI 10.5755/j01.itc.55.3.44582

PUFormer: Prior Knowledge-Guided and Uncertainty-Driven Multi-Scale Fusion Network for Head and Neck OAR Segmentation

Received 2026/03/18 | Accepted after revision 2026/06/28

HOW TO CITE: Han, Q., Xie, Y., Deng, L., Ma, X., Yang, Q., Zheng, Y. (2026) PUFormer: Prior Knowledge-Guided and Uncertainty-Driven Multi-Scale Fusion Network for Head and Neck OAR Segmentation. *Information Technology and Control*, 55(3), 1205-1225. https://doi.org/10.5755/j01.itc.55.3.44582

PUFormer: Prior Knowledge-Guided and Uncertainty-Driven Multi-Scale Fusion Network for Head and Neck OAR Segmentation

Qiang Han
Intelligent Perception and Control Key Laboratory of Sichuan Province, Sichuan University of Science and Engineering, Yibin 644000, China

Yunxiang Xie, Lan Deng, Xingyuan Ma, Qianguo Yang
School of Automation and Information Engineering, Sichuan University of Science and Engineering, Yibin 644000, China

Yuebin Zheng
Department of Otolaryngology Head and Neck Surgery, Zigong First People's Hospital, Zigong 643000, China

Corresponding author: zgyyyzyb@163.com

Computer-aided diagnosis plays a pivotal role in clinical medicine. In recent years, models based on convolutional neural networks (CNNs) and Transformers have achieved remarkable performance in medical image segmentation. However, existing methods still face significant challenges in multi-organ segmentation tasks, particularly when dealing with anatomically complex structures and targets exhibiting large shape variations. To address these limitations, we propose PUFormer, a novel CNN-Transformer hybrid architecture. The proposed model consists of three key components: (1) a prior knowledge-guided learning strategy, implemented through Prior Shunted Transformer Blocks, each of which incorporates a Prior Shunted Self-Attention (PSSA) mechanism as its core operation, leveraging anatomical prior knowledge to guide the model in learning consistent spatial relationships among organs; (2) a feature aggregation enhancement module,

designed to extract effective multi-scale features from aggregated multi-level representations; and (3) an uncertainty-driven multi-scale gated fusion module, which employs prediction entropy-based uncertainty to dynamically allocate fusion weights, thereby improving the effectiveness of multi-scale feature integration. We conducted extensive experiments on three public datasets: HaN-Seg, SegRap2023, and ACDC. The results demonstrate that PUFormer achieves competitive performance compared to state-of-the-art methods. Specifically, on the HaN-Seg dataset, it attains an average Dice coefficient of 77.04% and HD95 of 3.56 mm. On SegRap2023, it reaches an average Dice of 87.02% and HD95 of 3.04 mm. On the ACDC dataset, the model achieves an average Dice coefficient of 92.65% and HD95 of 3.22 mm. These findings validate the effectiveness of PUFormer in head and neck organs-at-risk (OAR) segmentation, and further demonstrate its applicability across different medical imaging modalities.

KEYWORDS: Medical image segmentation; Transformer; Convolutional neural networks; Priori knowledge; Multi-scale features fusion

1. Introduction

3D medical image segmentation is a cornerstone technology for precise radiotherapy, playing a critical role in disease diagnosis, treatment planning, and prognosis evaluation [18]. In the treatment of head and neck tumors, radiotherapy remains the primary curative modality, often supplemented by chemotherapy [5]. Given the prolonged treatment cycles and stringent dose control requirements, it is essential to deliver a radical dose to the tumor target while maximally sparing surrounding normal organs and critical neural structures [9]. However, the head and neck region features highly complex and densely packed anatomy, encompassing numerous radiation-sensitive organs such as the optic nerves, brainstem, cochlea, parotid glands, and thyroid. Even minor segmentation errors can be significantly amplified through steep dose gradients, directly compromising dose calculation accuracy and treatment safety [15]. Therefore, developing high-precision automatic segmentation techniques for head and neck organs-at-risk (OARs) represents a pressing clinical need to achieve "precise tumor ablation with maximum normal tissue protection."

Segmentation inaccuracies of head and neck OARs can lead to dose exceedance and severe complications. For instance, deviations in optic nerve or optic chiasm segmentation may cause radiation-induced optic neuropathy, potentially resulting in irreversible vision impairment or blindness. Inaccurate parotid gland delineation can induce permanent xerostomia and dysphagia, substantially diminishing patients' quality of life [25]. Due to its low tolerance dose, cochlear segmentation errors readily trigger sensorineural hearing loss. Excessive irradiation of the thyroid may lead to hypothyroidism, necessitating lifelong hormone replacement therapy [26]. As the vital center of the nervous system, brainstem segmentation inaccuracies can provoke severe neurological dysfunction or even life-threatening complications [23]. Current methods still struggle to achieve clinically acceptable precision on small-volume, low-contrast organs. There is thus an urgent need to develop specialized high-accuracy automatic segmentation models tailored to the complex anatomical environment of the head and neck, thereby reducing radiotherapy toxicity and improving patients' long-term quality of life.

U-shaped encoder-decoder architectures have gained widespread popularity in medical image semantic segmentation [2]. Convolutional neural networks (CNNs) extract features through successive convolutional layers and downsampling operations, followed by progressive upsampling and skip connections to gradually restore image details [7]. Although CNNs have delivered promising results, their inherent limitations – restricted receptive fields and fixed downsampling designs – make it difficult to model long-range dependencies [10]. In contrast, Vision Transformers excel at capturing global contextual relationships. When applied to medical image segmentation, Transformers have achieved improved performance [12]. However, medical datasets typically contain far fewer samples than natural image datasets, and Transformers require substan-

tial inductive biases, which limits their full potential in this domain [31].

To overcome these shortcomings, many researchers have designed hybrid CNN-Transformer architectures that effectively integrate the strengths of both paradigms for enhanced extraction and fusion of high-level and low-level features [19]. These hybrid designs are generally categorized as serial or parallel. Nevertheless, existing CNN-Transformer hybrids still face a core challenge in head and neck OAR segmentation [39]. As shown in Figure 1, the head and neck region contains numerous organs with highly intricate anatomical relationships and significant scale differences, making models prone to severe feature ambiguity. This manifests as Transformers struggling to focus accurately on small-volume organs without sufficient anatomical constraints, difficulty in simultaneously recovering fine local boundary details and high-level semantics during decoding, and imbalanced contributions of multi-scale features in fusion, particularly degrading performance on small, low-contrast, or boundary-ambiguous regions.

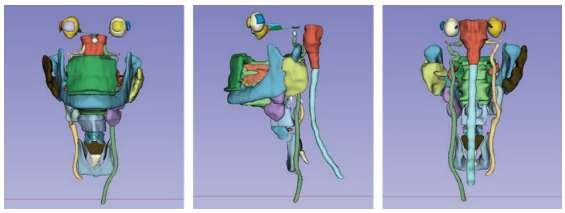
To address these issues, we propose PUFormer, a novel CNN-Transformer hybrid network. Unlike previous methods that focus primarily on local structural enhancement or straightforward feature fusion, PUFormer systematically alleviates feature ambiguity in complex anatomical scenarios through three synergistic levels: spatial localization, fine detail recovery, and difficult-region reinforcement. The main contributions of this work are as follows:

1 We propose a Prior-Shunted Self-Attention (PSSA) mechanism that injects structural prior knowledge as position-related biases into the Transformer attention logits. This effectively mitigates the spatial localization instability of small organs in complex anatomical scenes. The module employs distance transforms to generate dynamic structural prior maps, guiding the model to prioritize anatomically stable regions during early training and thereby enhancing sensitivity to low-contrast and small-volume organs.

2 We introduce a Feature Aggregation Enhancement (FAE) decoder. To overcome the loss of fine boundary details in traditional decoders, FAE aggregates high-level semantic features, low-level detail features, and edge boundary features, which are then refined through hierarchical feature enhancement modules. Compared with existing decoder designs, this module more effectively restores multi-scale fine-grained spatial information, making it particularly suitable for segmenting small organs with large morphological variations in the head and neck region.

3 We present an Uncertainty-driven Gated Multi-scale Feature Fusion (UGMFF) module. This module employs a dynamic gating mechanism based on prediction entropy to adaptively strengthen multi-scale feature responses in high-uncertainty regions. Unlike prior approaches that use uncertainty only for feature calibration or boundary refinement, UGMFF directly leverages uncertainty to dynamically adjust multi-scale fusion weights, thereby reducing semantic conflicts in challenging areas and improving overall segmentation robustness.

Figure 1
Three-dimensional visualization of head and neck organs-at-risk (OARs) from anterior, lateral, and posterior views, illustrating the complexity of anatomical structures, significant scale variations, and dense spatial relationships among organs in this region.



2. Related Works

2.1. Hybrid Models for Segmentation

Hybrid CNN-Transformer architectures effectively capture both local details and global contextual information. TransUNet [4] pioneered the integration of Transformers into U-shaped CNN architectures by incorporating Transformer layers to model long-range dependencies. MISSFormer [14] introduced a context bridging module to extract global correlations and local context from multi-scale features. UNETR [11] employed a pure Transformer as the encoder, reformulating 3D medical image segmentation as a sequence-to-sequence prediction task, with a CNN-based decoder. HiFormer [13] achieved

efficient multi-scale feature fusion through a parallel dual-branch encoder and cross-attention mechanism, improving both segmentation accuracy and computational efficiency. AgileFormer [28] proposed a deformable positional encoding that explicitly models irregular sampling patterns within self-attention, enabling the network to capture spatial dynamic information across different segmentation targets. UANet [44] introduced an uncertainty-guided feature recalibration module that combines adaptive convolution with pixel-wise uncertainty estimation, enhancing performance on small targets and deformable structures.

2.2. Network Based on Prior Knowledge

To address the challenges posed by complex medical images, incorporating prior knowledge has proven effective in improving the robustness and interpretability of segmentation models. For example, SPAN [22] leverages inherent spatial and semantic structures from unlabeled data to regularize Transformer attention mechanisms, generating more reliable attention maps and enhancing generalization in few-shot downstream tasks. DCNN [41] generates pancreas location prior maps from statistical analysis of training data and designs a prior propagation module to dynamically integrate them across different layers of the segmentation network, utilizing anatomical consistency to boost accuracy. SPC-Net [43] proposed a cyclic collaborative framework that iteratively optimizes a shape prior learning model alongside the segmentation network, effectively embedding shape knowledge into the segmentation process. MPGNet [8] introduced a correlation graph aggregation module that encodes brain anatomical priors as graph relations to enrich feature representations. MPF-Net [32] applies superpixel prior knowledge to both encoder and decoder stages, providing initial edge guidance during encoding and refining segmentation boundaries during decoding.

2.3. Multi-scale Feature Learning-based Methods for Medical Image Segmentation

Multi-scale feature fusion aims to integrate information from different resolutions, thereby recovering local and spatial details lost during downsampling. EMNet [42] combines edge guidance with multi-scale feature fusion to achieve synergistic edge-semantic multi-scale vessel segmentation. MCPA [35] employs cross-perception fusion to integrate multi-scale features, precisely capturing fine-grained local details and global context, which significantly improves segmentation accuracy. BCL-Former [33] proposed a multi-scale local Transformer fusion decoder that adopts a progressive feature fusion strategy to mitigate discrepancies between high- and low-level features. SMAFormer [45] introduced a collaborative multi-attention mechanism to enhance synergy among attention modules for effective multi-scale feature integration. MSPCNF-Net [38] incorporates bidirectional window-aware attention and bidirectional cross-attention to capture both local correlations and long-range dependencies between dual-encoder features, facilitating comprehensive fusion from fine-grained details to coarse semantic representations.

3. Methods

3.1. Overall Architecture

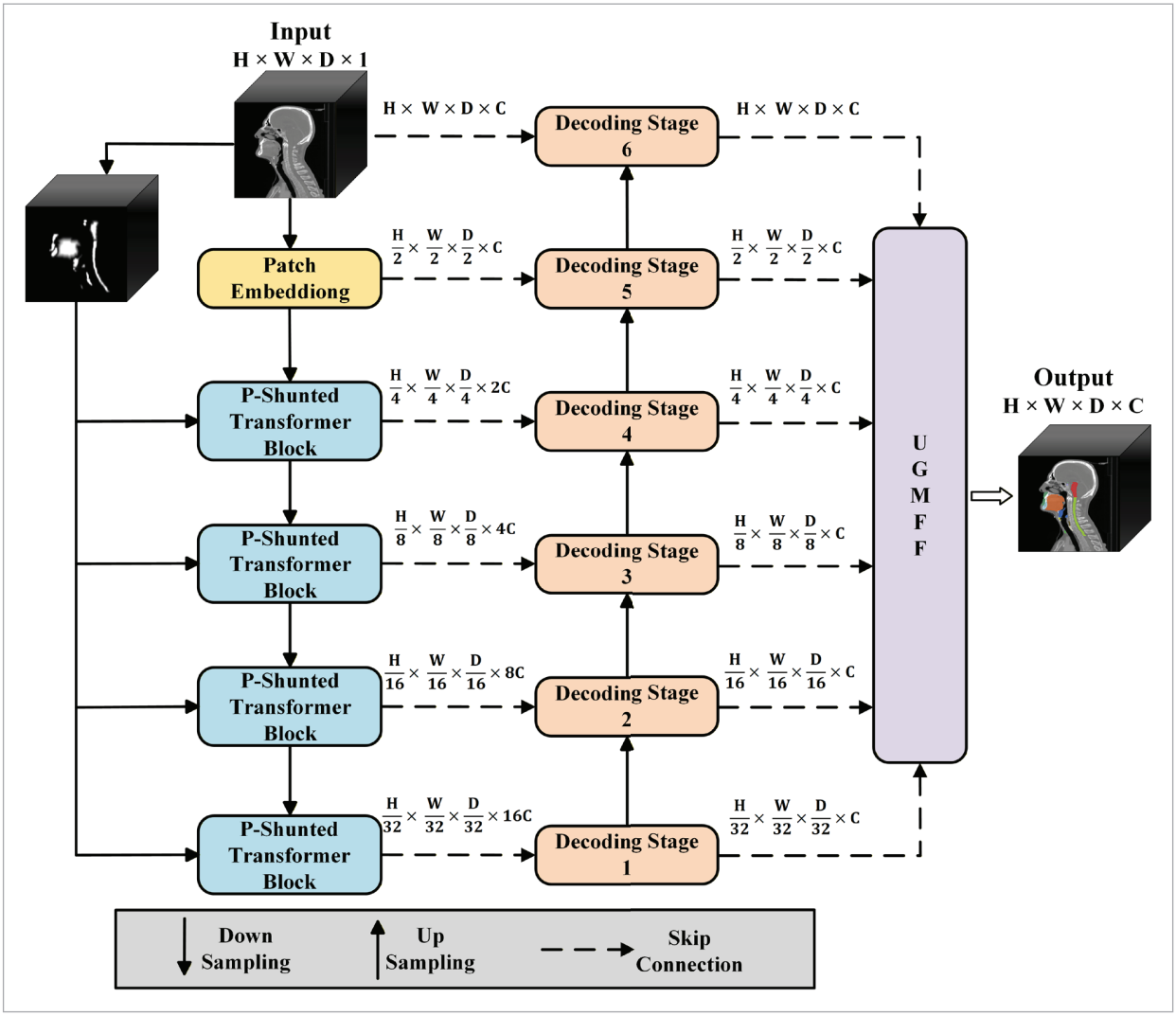
The overall architecture of PUFormer is illustrated in Figure 2. The network consists of three main components: an encoder, a decoder, and a multi-scale feature fusion module. The encoder is built upon P-Shunted Transformer blocks, which generate multi-level feature maps that are transmitted to the decoder via skip connections. The decoder employs the Feature Aggregation Enhancement (FAE) module to progressively recover fine-grained feature maps at six different scales. These multi-scale features are then fed into the multi-scale feature fusion module for adaptive integration. Finally, the network produces segmentation predictions at the same resolution as the input image.

3.2. Prior Knowledge-Guided Learning Strategy

When manually delineating head and neck OARs, clinicians typically exploit the relatively stable spatial and anatomical relationships between organs – such as superior-inferior, left-right, and adjacency cues – to make layer-by-layer decisions. This indicates that head and neck organs possess not only local textural features but also consistent spatial structural prior knowledge.

Figure 2

The overall architecture of PUFormer, consisting of three main components: a Prior-Shunted Transformer encoder, a Feature Aggregation Enhancement (FAE) decoder, and an Uncertainty-Guided Multi-scale Feature Fusion (UGMFF) module.



However, existing Transformer-based segmentation models primarily rely on data-driven self-attention mechanisms for global relationship modeling, lacking explicit anatomical constraints. In complex head and neck scenarios, characterized by numerous organs, large scale variations, and low-contrast boundaries, self-attention is easily distracted by background noise and ambiguous boundary regions. This leads to unstable attention distributions and frequent neglect of small-volume organs and fine-grained structures. To address this limitation, inspired by distance-transform-based spatial prior approaches such as DCNN [41], we propose a dynamic prior map based on structural center enhancement. The core idea is that, for most head and neck OARs, the central regions of organs typically exhibit more stable and consistent semantic representations, whereas boundary areas are more susceptible to partial volume effects, weak contrast, and inter-organ adjacency. By enhancing the central regions while moderately suppressing the boundaries, the prior map provides the Transformer with stable anatomical localization signals, thereby alleviating spatial attention ambiguity in complex anatomical environments. Compared with complex to-

pological constraints or statistical shape models, the distance-transform-based prior requires no additional anatomical modeling. It offers a computationally lightweight yet effective inductive bias for positional guidance and demonstrates strong cross-dataset generalization capability.

Specifically, given a 3D label map $L \in R^{H \times W \times D}$, where H, W, and D denote the height, width, and depth of the image, respectively. For the c-th organ class (c=1,2,... ,C, where C is the total number of segmentation classes), its binary mask is defined as:

$$M_c(x) = \begin{cases} 1, if\ L(x) = c \\ 0, otherwise \end{cases}, \quad (1)$$

where $x=(i,j,k)$ represents the 3D spatial coordinates. Next, the Euclidean distance transform is computed for each binary mask to obtain the distance map:

$$D_c(x) = \min_{y \in \partial M_c} \| x - y \|_2, \quad (2)$$

where ∂M_c denotes the boundary (surface voxel set) of the c-th organ mask.

In the distance map, voxels at the organ center take larger values, while voxels near the boundary take smaller values. To further enhance the structural response in the central region while suppressing unstable regions near the boundary, we normalize the distance map and apply Gaussian smoothing:

$$\widetilde{D}_c(x) = exp\left(-\frac{(D_c(x) - max D_c(x))^2}{2\sigma^2}\right) \cdot \frac{D_c(x)}{max(D_c)}, \quad (3)$$

where $maxD_c(x)$ represents the maximum distance value in the organ distance map, and σ is the Gaussian kernel parameter used to control the spatial decay range of the structural response.

Finally, the structural responses of all classes are fused at each spatial location to obtain the final structural prior map:

$$P(x) = \max_{c=1,\ldots,C} \widetilde{D}_c(x), \quad (4)$$

The generated structural prior map is shown in Figure 3, which presents the structural prior from three different perspectives.

Figure 3

Visualization of the structural prior map from three orthogonal views, where brighter regions correspond to organ central areas with higher structural confidence.

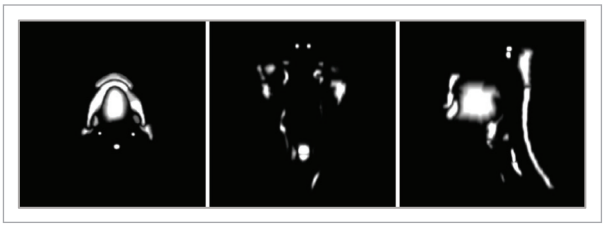
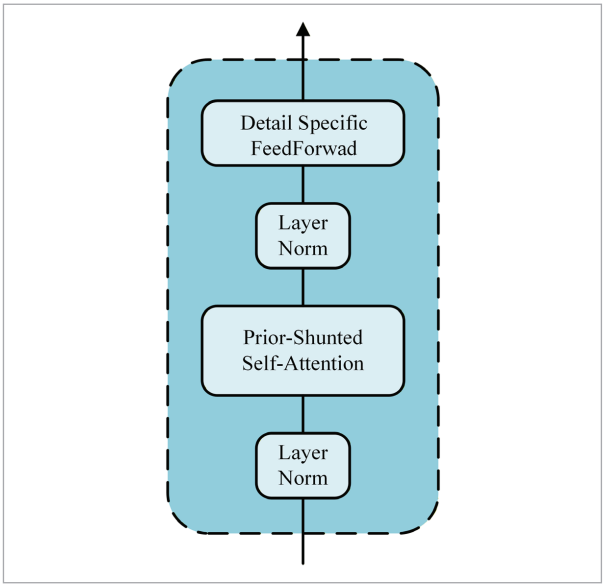


Figure 4

The P-Shunted Transformer block, comprising a Layer Normalization, a Prior-Shunted Self-Attention (PSSA) layer for anatomical prior-guided attention, a second Layer Normalization, and a Detail-Specific Feed-Forward (DSFF) layer for local detail enhancement.



As shown in Figure 4, the P-Shunted Transformer module consists of a Prior-Shunted Self-Attention (PSSA) layer, a Detail-Specific Feed-Forward (DSFF) layer, and two Layer Normalization (LN) layers.

Inspired by [6] and [40], we adopt an additive bias approach to integrate the structural prior map into the attention logits. Specifically, given the input feature $X \in R^{N \times D}$, where $N=h \times w \times d_s$ is the number of flattened tokens (h, w, and d_s denote the spatial height, width, and depth of the feature map, respectively) and D is the feature dimension, we first obtain the query, key, and value through linear projections:

$$Q = XW_Q, K = XW_K, V = XW_V, \quad (5)$$

where $Q,\ K,\ V \in R^{N\times D}$. The self-attention similarity matrix $QK^T \in R^{N\times N}$ is then obtained.

To inject anatomical prior knowledge, we require a bias matrix of the same dimensions as QK^T. The original structural prior map is first downsampled to the current feature map resolution via trilinear interpolation and then flattened into a one-dimensional vector $P \in R^N$, where each element P_i corresponds to the prior strength at the i-th token position. To apply this prior P to the similarity matrix $N\times N$, we use broadcasting addition by treating λP as a row bias added to every row of QK^T:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}} + \lambda \cdot P \cdot 1^T)V, \quad (6)$$

where $1^T \in R^N$ is an all-ones vector, and $P \cdot 1^T$ yields an $N\times N$ matrix. λ is a learnable parameter that controls the contribution of prior information to the self-attention weights.

From the perspective of the attention mechanism, this operation is equivalent to introducing a position-aware anatomical bias into the attention matrix before the softmax function. Unlike standard self-attention, which relies entirely on Query-Key similarity, this bias provides additional responses to positions with stable anatomical significance during the early stages of training. This mitigates attention instability caused by insufficient training data, blurry boundaries, and small organ sizes. Compared with multiplicative priors, the additive bias does not directly alter the relative structure of Query-Key similarities. It should be further noted that the prior bias $\lambda \cdot P \cdot 1^T$ acts as a position-dependent row bias rather than a uniform offset: the value added to the attention logit between query iand key jis determined by P_j, the prior strength at position j. Consequently, keys located in anatomically stable regions with high prior values receive selectively elevated logits before the softmax operation, while background positions with low prior values receive a comparatively smaller increment, resulting win position-selective enhancement rather than uniform amplification across all keys. Therefore, it guides the attention distribution in a smoother manner while preserving the Transformer's global modeling capability. Particularly in the early training phase, when Query-Key representations have not yet fully converged, the structural prior increases the response probability of organ center regions, enabling the model to more easily learn stable spatial localization relationships. As training progresses, the learnable parameter λ adaptively adjusts the contribution of the prior through backpropagation, preventing the model from over-relying on fixed anatomical templates.

The DSFF layer is designed to capture cross-token dependencies while enhancing local detail extraction via a dedicated detail-specific branch [29]. Its formulation is given by:

$$x' = FC(x;\ \theta_1) \quad (7)$$

$$x'' = FC(\sigma(x' + DS(x';\ \theta));\ \theta_2), \quad (8)$$

where $FC(\cdot)$ denotes a fully connected layer, denotes the GELU activation function, and $DS(\cdot)$ denotes the detail-specific layer, implemented via depth-wise convolution with learnable parameters.

3.3. Feature Aggregation Enhancement (FAE) Decoder

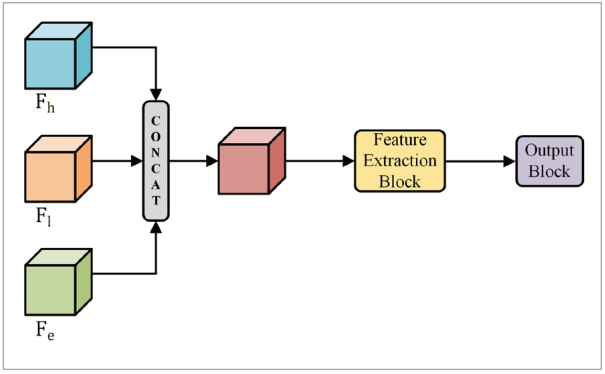
Although the encoder effectively extracts contextual information and multi-scale features, it does not sufficiently emphasize local feature extraction [36]. To enable the model to pay greater attention to local edge details, we introduce the Feature Aggregation Enhancement (FAE) module in the decoder of PUFormer.

As shown in Figure 5, the FAE module aggregates three types of feature maps: low-level feature maps (F_l) that contain rich edge and spatial details, high-level feature maps (F_h) that provide rich semantic information, and the mask boundary map (F_e) generated from the low-level feature map to supply explicit shape constraints.

Specifically, at the j-th decoding stage (j=1,2,...,6), we concatenate three inputs along the channel dimension: the low-level feature map $F_l(\frac{H}{2^{j-1}}\times\frac{W}{2^{j-1}}\times\frac{D}{2^{j-1}})$ from the corresponding skip connection of the encoder, the upsampled high-level feature map F_h $(\frac{H}{2^{j-1}}\times\frac{W}{2^{j-1}}\times\frac{D}{2^{j-1}})$ from the previous decoding stage, and the mask boundary map $F_e(\frac{H}{2^{j-1}}\times\frac{W}{2^{j-1}}\times\frac{D}{2^{j-1}})$. The boundary

Figure 5

The Feature Aggregation Enhancement (FAE) module, which aggregates high-level semantic features F_h, low-level spatial detail features F_l, and boundary map $F_e$via channel-wise concatenation, followed by a Feature Extraction Block and an Output Block for progressive fine-grained feature recovery.



map F_e is obtained from the low-level feature map F_l via a 3×3×3 convolution, followed by BatchNorm, ReLU, and then a 1×1×1 convolution. During concatenation, a 1×1×1 convolution is adopted to compress the channel dimensions. The compressed channel numbers corresponding to the 6 decoding stages are 1024, 512, 256, 128, 64, and 64, respectively.

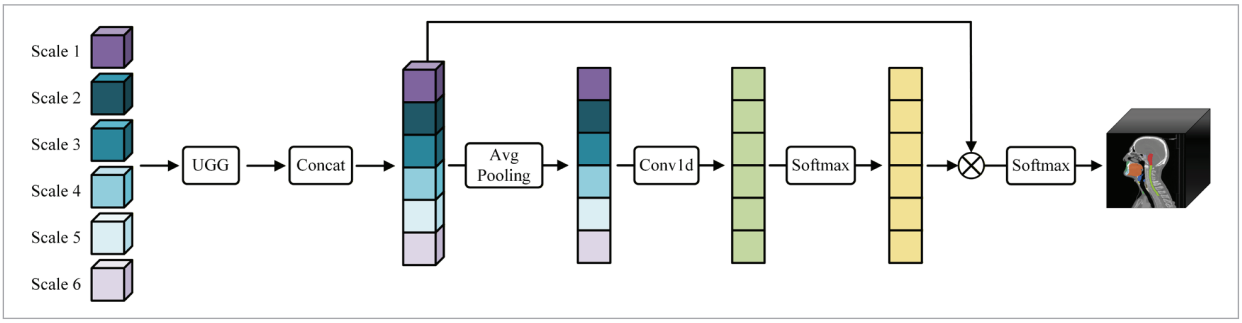
The concatenated feature map is then fed into the Feature Enhancement (FE) module [1]. Each FE Block consists of two convolutional layers (each followed by Instance Normalization and ReLU), a ResBlock, and an SCSE [30] attention module. The first four decoding stages contain 4, 3, 2, and 1 FE Blocks, respectively. The 5th and 6th decoding stages do not use FE Blocks; they only output the features after upsampling and skip-connection concatenation. The output module uses a 1×1×1 convolution to obtain the feature map F_s^j with a spatial resolution of $\frac{H}{2^{j-1}} \times \frac{W}{2^{j-1}} \times \frac{D}{2^{j-1}}$. Through the above feature aggregation mechanism, the FAE module effectively integrates multi-level features, mitigates semantic discrepancies across different scales, and significantly enhances the segmentation accuracy of small organs, their boundaries, and fine-grained structures.

3.4. Uncertainty-Driven Multi-scale Gated Fusion Module

In head and neck OAR segmentation tasks, organs exhibit significant differences in scale and morphological complexity [46]. Small-volume organs, low-contrast regions, and blurry boundary areas are prone to higher prediction uncertainty. Traditional multi-scale fusion methods typically adopt fixed or uniform fusion strategies, which fail to dynamically adjust feature contributions according to regional segmentation difficulty, resulting in insufficient feature representation in challenging areas. To address this issue, we propose the Uncertainty-Guided Multi-scale Feature Fusion (UGMFF) module, which explicitly incorporates model prediction uncertainty into the multi-scale feature fusion process. The core idea is that regions with high prediction entropy (difficult regions where the model is uncertain) should receive stronger multi-scale feature responses, thereby reinforcing the learning of complex structures such as boundaries and small organs [37]. As shown in Figure 6, the UGMFF module operates as follows. Let the decoder produce features at S dif-

Figure 6

The Uncertainty-Guided Multi-scale Feature Fusion (UGMFF) module. Features from six decoder scales are processed by the Uncertainty-Guided Gating (UGG) module and concatenated, followed by global average pooling and a Conv1d with softmax to generate adaptive scale-wise fusion weights, dynamically emphasizing complex regions such as organ boundaries and small-volume structures.



ferent scales. The output feature map of each scale is upsampled to the same spatial resolution as the input image and denoted as $F_s \in R^{C\times H\times W\times D}$, where s=1,2,... ,S, Cis the number of channels, and H, W, D represent the height, width, and depth, respectively. These features are then fed into the Uncertainty-Guided Gating (UGG) module.

For each scale s, the feature map F_s is passed through a 1×1×1convolution followed by a softmax operation to generate the corresponding pixel-wise class probability map.:

$$P_s = softmax(Conv_{1\times1\times1}(F_s)) \in R^{K\times H\times W\times D}, \quad (9)$$

where K is the total number of classes, and $P_s^{(k)}(i,j,l)$ denotes the predicted probability that the voxel at spatial position (i,j,l) belongs to class k at scale s.

To obtain a comprehensive cross-scale prediction, the probability maps from all S scales are aggregated by averaging:

$$\bar{P}(i,j,l) = \frac{1}{S}\sum_{s=1}^{S} P_s(i,j,l) \in R^K. \quad (10)$$

This aggregation unifies multi-scale semantic predictions into a single probability distribution, providing a reliable basis for subsequent uncertainty estimation.

To quantify the model's prediction confidence at each voxel position, we adopt prediction entropy as the uncertainty measure. Prediction entropy reflects the dispersion of the output probability distribution: a lower entropy value indicates high prediction confidence with a concentrated distribution, while a higher entropy value suggests greater ambiguity with a more uniform distribution. Based on the aggregated probability $\bar{P}$, the prediction entropy is computed at each voxel position (i,j,l) to construct the three-dimensional spatial uncertainty map $H \in R^{H\times W\times D}$:

$$H(i,j,l) = -\sum_{k=1}^{K} \bar{P}^{(k)}(i,j,l) \cdot log\left(\bar{P}^{(k)}(i,j,l) + \epsilon\right). \quad (11)$$

where ϵ is a small constant to avoid $log(0)$. When the model is highly confident about a voxel, the probability concentrates on a single class and the entropy approaches zero. When probabilities across classes are similar, the distribution becomes nearly uniform and the entropy approaches $log\ K$. This property allows H to effectively identify difficult voxels such as organ boundaries, small structures, and low-contrast regions.

To guide multi-scale feature fusion with uncertainty information, we normalize the prediction entropy map H by its theoretical maximum $log\ K$ to obtain the three-dimensional gating weight map:

$$G(i,j,l) = \frac{H(i,j,l)}{\log K}, \quad (12)$$

When $H\to0$ (low uncertainty), $G\to0$, and the gating weight approaches zero. When $H\to logK$ (high uncertainty), $G\to1$, fully activating the gate and assigning maximum multi-scale feature response to that voxel. The gating map G serves as a globally shared three-dimensional weight that modulates the entire feature fusion process without introducing additional learnable parameters.

Each upsampled feature map F_s is then spatially modulated using the gating weight G. The gated enhanced feature at scale sis defined as:

$$F_s' = F_s \odot (1+G), \quad (13)$$

where $\odot$ denotes element-wise multiplication. As described earlier, for low-uncertainty voxels, $G\to0$, so there is no significant amplification and the original feature representation remains stable. For high-uncertainty voxels, the gain approaches 2, providing strong enhancement for complex 3D structures such as boundaries and small organs. The residual connection effectively prevents vanishing gradients during early training.

After uncertainty-weighted re-calibration, features at different scales obtain differentiated responses in difficult regions. The gated enhanced features from all S scales are then concatenated along the channel dimension:

$$F_{\text{cat}} = \text{Concat}(F_1', F_2', \ldots, F_S'), \quad (14)$$

Global average pooling is subsequently applied to compress the spatial dimensions. A 1×1×1 convolution followed by a softmax operation generates the adaptive fusion weight vector for each scale:

$$w = SoftmaxConv_{1\times1\times1}\left(GAP(F_{cat})\right). \quad (15)$$

where the s-th element of w_s reflects the global importance of the s-th scale's features for the final segmentation under the current semantic context. Finally, the weighted sum of the gated enhanced features produces the final fused feature map:

$$F_{out} = \sum_{s=1}^{S} w_s \cdot \widetilde{F}_s \in R^{C\times H\times W\times D}. \quad (16)$$

The final semantic segmentation result is then obtained by passing F_{out} through a 3D segmentation head.

4. Experiments

4.1. Datasets

We evaluated the proposed PUFormer on three publicly available datasets: HaN-Seg [27], SegRap2023 [20], and ACDC [3].

4.1.1. HaN-Seg Dataset

The HaN-Seg challenge provides 42 accessible cases. Each case consists of contrast-enhanced head and neck CT scans from patients who received radiotherapy for previous cancer diagnoses, along with manual labels for 30 organs-at-risk (OARs). Due to the relatively small sample size, we employed 5-fold cross-validation for evaluation. The test sets in each fold contained 9, 9, 8, 8, and 8 cases, respectively. We report the average performance across all five folds.

4.1.2. SegRap2023 Dataset

The SegRap2023 challenge provides 120 CT scans from patients with histologically confirmed nasopharyngeal carcinoma who were receiving their first course of treatment. The dataset includes fine-grained annotations for 45 OARs. We randomly split the dataset into 84, 12, and 24 cases for training, validation, and testing, respectively.

4.1.3. ACDC Dataset

The ACDC dataset is a cardiac MRI dataset containing scans from 100 patients. Each sample includes annotations for the right ventricle (RV), left ventricular myocardium (Myo), and left ventricle (LV). For fair comparison, we followed the data split used in TransUNet [4], employing 70, 10, and 24 cases for training, validation, and testing, respectively

4.2. Implementation Details

The proposed method was implemented in PyTorch 2.7.0 and MONAI 1.5.1, and trained on an NVIDIA RTX 5070 Ti GPU. We used the AdamW optimizer with an initial learning rate of 1e-4 and a weight decay of 1e-5.

During training, data augmentation strategies including random rotation, flipping, and Gaussian smoothing were applied only to the ACDC dataset. For HaN-Seg and SegRap2023, such flipping was avoided because the segmentation targets exhibit strong anatomical symmetry, and flipping would adversely affect model training.

For HaN-Seg and SegRap2023, all images were resampled to a voxel spacing of 1×1×2 mm and cropped to a volume size of 512×512×196. A sliding window strategy was adopted, with random crops of size 96×96×64 used as 3D input patches. The batch size was set to 1, and the model was trained for 4000 iterations. For the ACDC dataset, images were resampled to a voxel spacing of 1.25×1.25×10 mm with a volume size of 256×256×10. The input patch size was 128×128×6, the batch size was 8, and training was performed for 4000 iterations.

4.3. Loss Function

We adopted a combination of the Dice loss [24] and the cross-entropy loss [21]:

$$L = \alpha L_{dice} + \beta L_{ce}, \quad (17)$$

where α and β are the weights for Dice loss and cross-entropy loss, respectively, both set to 1 by default in our experiments. The Dice loss directly optimizes region overlap and is naturally robust to class imbalance, while the cross-entropy loss provides voxel-wise supervision that helps stabilize training. This combination has been widely validated as a stable and effective scheme in medical image segmentation.

4.4. Comparison Methods

We compared PUFormer with several recent state-of-the-art (SOTA) methods, including HiFormer [13], CoTr [34], nnUNet [16], MS-TCNet [1], and E2MISeg [17].

4.5. Evaluation Metrics

To comprehensively evaluate segmentation performance, we employed two widely used metrics: the Dice Similarity Coefficient (DSC) and the 95% Hausdorff Distance (HD95). DSC measures the overlap between the predicted segmentation and the ground truth, while HD95 quantifies the maximum boundary distance between them. Their formulations are as follows:

$$Dice = \frac{2|A \cap B|}{|A| + |B|} \tag{18}$$

$$HD(A, B) = \max\left(\max_{a \in A} \min_{b \in B} d(a, b), \max_{b \in B} \min_{a \in A} d(a, b)\right), \tag{19}$$

where $|A|$ and $|B|$ denote the number of voxels in the prediction and ground truth, respectively, and $d(a,b)$ represents the Euclidean distance between points a and b. The 95% in HD95 indicates that the top 5% of extreme distances are ignored to enhance robustness to outliers.

5. Results

5.1. Comparison on Three Tasks

5.1.1. Results on HaN-Seg

Table 1 presents a quantitative comparison between PUFormer and several state-of-the-art methods on the HaN-Seg dataset. The last row reports the average Dice similarity coefficient (%) and HD95 (mm) across all 30 organs. Our proposed PUFormer achieves the best overall performance with an average Dice score of 77.04% and an HD95 of 3.55 mm. Notably, the method demonstrates superior segmentation accuracy on radiation-sensitive glandular structures, including the lacrimal gland, submandibular gland, thyroid gland, and parotid gland. Figure 7 shows qualitative visualization results, where the first column represents the ground truth. Compared with other competing methods, PUFormer produces more accurate segmentation masks, particularly for organs with irregular shapes and complex boundaries.

Visual comparison of segmentation results on the HaN-Seg dataset. The first column shows the ground truth, and the remaining columns present the predictions of different methods. PUFormer produces more accurate segmentation masks, particularly for organs with irregular shapes and complex boundaries.

Figure 7

Visual comparison of segmentation results on the HaN-Seg dataset. The first column shows the ground truth, and the remaining columns present the predictions of different methods. PUFormer produces more accurate segmentation masks, particularly for organs with irregular shapes and complex boundaries.

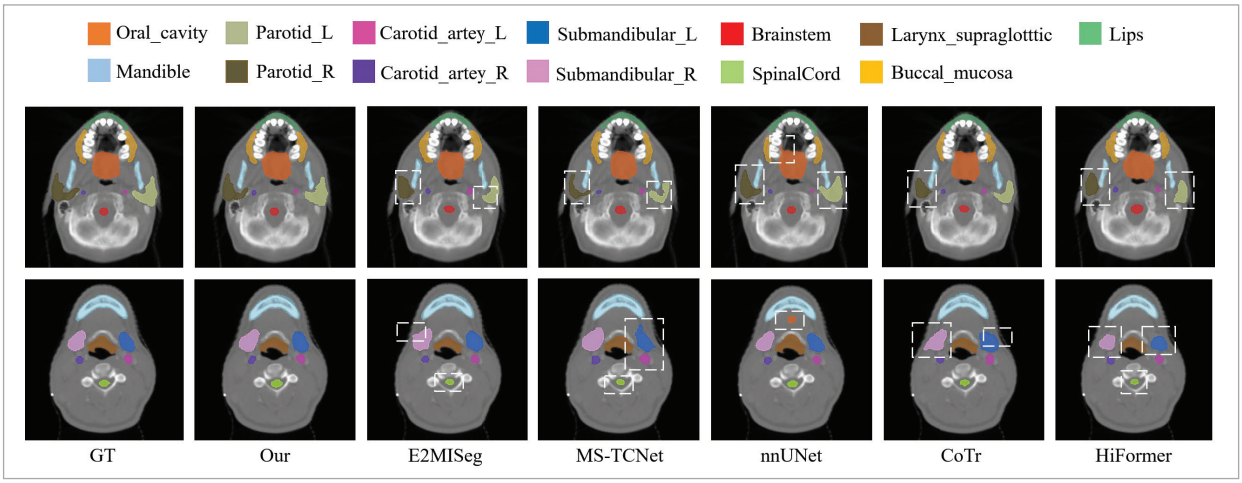


Table 1
Quantitative comparison of segmentation performance on the HaN-Seg dataset in terms of Dice similarity coefficient (%) and HD95 (mm). The best results are highlighted in bold.

Model	Our		E2MISeg		MS-TCNet		nnUNet		CoTr		HiFormer	
	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95
Carotid artery_L	81.76	3.41	**83.23**	**3.35**	80.83	3.94	78.98	4.20	77.44	4.86	77.18	4.92
Carotid artery_R	**84.22**	**3.24**	83.04	3.59	82.16	3.72	80.72	4.03	80.37	4.25	79.76	3.84
Arytenoids	**61.28**	**2.96**	58.34	3.82	57.34	4.18	57.09	5.64	60.18	3.62	55.23	5.91
Mandible	93.97	1.48	**94.52**	**1.26**	94.11	1.29	94.08	1.31	93.79	1.42	93.42	1.48
Brainstem	85.60	4.53	85.58	4.39	84.44	4.76	**86.42**	**4.26**	83.95	4.84	82.56	5.03
Buccal_mucosa	**70.43**	**4.87**	68.75	5.31	68.40	5.46	67.14	5.77	66.46	6.13	67.25	5.70
Oral_cavity	**89.76**	**5.34**	89.48	5.40	88.03	5.97	88.25	5.82	86.60	6.08	86.74	5.99
Cochlea_L	**74.22**	**1.27**	71.39	1.62	70.14	2.14	70.97	1.94	71.62	1.57	69.33	2.42
Cochlea_R_	**76.16**	**1.31**	72.25	1.78	71.57	2.09	72.49	1.54	73.96	1.68	71.08	2.24
Cricopharyngeal_inlet	**65.35**	**6.09**	63.02	6.74	63.76	6.58	64.13	6.29	63.53	6.93	62.97	7.11
Esophagus	62.83	8.14	**62.94**	7.93	61.97	**7.82**	58.38	9.06	60.32	8.33	58.69	8.85
Eyeball_A_L	77.48	**2.35**	76.78	2.49	76.22	2.76	**77.55**	2.40	76.91	2.84	75.45	3.01
Eyeball_A_R	**79.51**	**2.42**	76.46	2.75	77.16	2.59	76.77	2.82	76.08	2.78	76.14	2.76
Eyeball_P_L	91.96	1.62	**92.23**	**1.58**	92.04	1.60	92.08	1.77	90.15	1.95	91.26	1.83
Eyeball (P, R)	91.80	1.81	91.88	1.74	**91.95**	**1.65**	91.62	1.72	90.37	2.08	90.42	1.96
Lacrimal gland_L	**63.62**	**3.09**	59.94	3.42	60.36	3.27	58.14	4.03	58.52	3.67	59.23	3.52
Lacrimal gland_R	**61.10**	**3.57**	58.98	3.85	60.41	3.64	59.92	3.79	59.14	3.73	58.36	4.14
Submandibular_L	**82.49**	**3.96**	81.32	4.67	80.98	4.83	80.66	5.14	79.76	6.31	79.48	5.63
Submandibular_R	**82.35**	**4.15**	80.85	4.91	80.65	5.08	79.53	5.28	79.42	5.46	79.17	6.02
Thyroid	**88.54**	**2.82**	88.29	3.22	87.26	3.81	88.44	3.06	87.33	3.64	86.75	4.16
Larynx_Glottic	**73.48**	**3.20**	72.53	3.74	71.38	3.82	69.13	4.22	70.59	4.03	69.42	4.39
Larynx_Supraglot	**78.37**	3.93	77.95	**3.68**	75.16	4.84	75.44	4.45	74.87	4.67	76.63	4.36
Lips	73.75	7.33	72.92	8.17	**75.24**	**7.04**	70.38	8.83	72.46	7.96	70.14	9.21
Optic_chiasm	**52.44**	**3.68**	48.43	4.65	49.92	4.28	50.38	4.16	47.19	4.99	44.56	5.28
Optic_nerve_L	**70.18**	**2.26**	68.36	2.64	68.43	2.49	67.92	2.61	67.28	2.92	66.89	3.37
Optic_nerve_R	**73.82**	**2.02**	70.57	2.41	69.58	2.80	69.70	2.76	68.47	2.83	68.32	2.93
Parotid gland_L	84.13	**5.61**	**84.45**	5.97	82.93	6.42	81.72	6.12	81.88	7.56	81.54	8.14
Parotid gland_R	**83.26**	**5.94**	83.19	6.01	81.86	6.87	80.77	7.44	81.26	7.03	80.73	7.59
Pituitary gland	**74.63**	**2.07**	71.52	2.96	72.09	2.73	70.87	3.14	71.68	2.84	70.35	3.56
Spinal cord	**82.87**	**2.28**	82.04	2.53	81.70	2.61	80.87	2.78	80.82	2.86	80.46	3.27
Average	**77.04**	**3.56**	75.70	3.89	75.26	4.04	74.68	4.21	74.41	4.32	73.65	4.62

5.1.2. Results on SegRap2023

Table 2 provides a quantitative comparison between PUFormer and other state-of-the-art methods on the SegRap2023 dataset. The last row reports the average results across all organs. Our proposed PUFormer achieves the best performance with an average Dice coefficient of 87.02% and an HD95 of 3.04 mm, outperforming the second-best method by a margin of +1.55% in Dice and -0.59 mm in HD95, respectively. Figure 8 presents the qualitative visualization results. PUFormer produces segmentation masks with smaller errors across multiple organs. In particular, for the submandibular gland, which has highly irregular shapes, MS-TCNet, CoTr, and HiFormer exhibit noticeable deviations from the ground truth, whereas our method achieves more accurate delineation.

Table 2

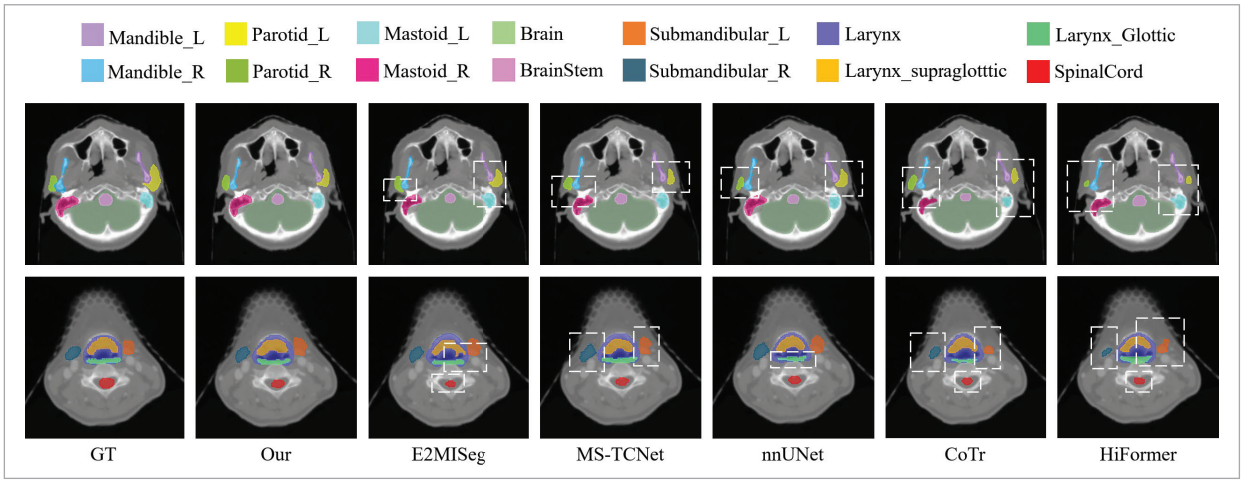
Quantitative comparison of segmentation performance on the SegRap2023 dataset in terms of Dice similarity coefficient (%) and HD95 (mm). The best results are highlighted in bold.

Model	Our		E2MISeg		MS-TCNet		nnUNet		CoTr		HiFormer	
	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95
Brain	98.64	1.28	**98.73**	**1.14**	98.42	1.25	98.24	1.32	98.31	1.30	98.46	1.22
BrainStem	91.97	3.44	**92.38**	**3.25**	91.60	3.60	91.51	3.63	91.43	3.79	91.65	3.67
Chiasm	**69.88**	**3.17**	67.26	3.64	66.32	3.98	67.44	3.51	66.04	4.22	63.72	4.73
Cochlea_L	**94.85**	**1.08**	94.37	1.49	93.49	1.96	93.18	2.16	92.96	2.47	92.77	2.44
Cochlea_R	94.39	1.15	**94.48**	**1.02**	93.57	1.89	93.66	1.84	93.02	1.99	92.43	2.50
Esophagus	**79.42**	**5.34**	76.19	6.25	76.35	6.04	77.12	5.88	75.87	6.74	74.26	7.26
ETbone_L	**82.76**	**1.91**	81.27	2.18	81.09	2.33	80.23	2.47	79.39	2.84	79.45	2.76
ETbone_R	**92.98**	**1.72**	92.70	1.95	92.26	2.49	91.84	2.54	91.43	2.61	91.64	2.85
Eye_L	91.86	2.04	**92.13**	**1.92**	91.33	2.21	91.07	2.68	90.88	2.56	90.63	2.97
Eye_R	**78.93**	**2.35**	78.29	2.56	77.62	2.74	77.05	2.93	76.41	2.99	74.75	3.24
Hippocampus_L	**75.42**	**3.62**	74.83	4.19	74.46	4.33	75.08	3.86	71.56	4.82	70.27	5.88
Hippocampus_R	**78.95**	3.57	76.54	4.80	77.29	4.12	78.36	**3.45**	75.34	4.43	74.82	5.62
IAC_L	**80.52**	**1.23**	78.29	1.58	79.14	1.45	77.93	1.98	74.86	2.47	75.35	2.02
IAC_R	**86.85**	**1.14**	85.73	1.39	84.88	1.67	83.98	1.73	82.53	2.19	83.46	1.98
Larynx	**90.54**	**2.96**	88.62	3.41	88.36	3.54	87.49	3.62	87.10	3.66	85.83	3.97
Larynx_Glottic	83.96	4.34	82.25	6.50	**84.13**	**4.05**	82.42	6.89	81.19	6.97	80.87	7.35
Larynx_Supraglot	**84.08**	**4.51**	83.34	5.98	81.94	5.69	81.82	5.73	79.25	6.64	79.69	6.49
Lens_L	**83.77**	**1.22**	80.13	1.76	80.38	1.65	79.24	1.94	78.47	2.18	77.82	2.51
Lens_R	**84.92**	**1.27**	81.99	1.84	82.26	1.80	81.79	2.09	81.21	2.31	80.34	2.65
Mandible_L	**84.72**	**6.36**	82.93	7.07	82.47	7.52	81.84	8.48	81.38	8.67	79.21	10.44
Mandible_R	**84.65**	**6.25**	82.89	7.13	80.91	8.64	81.03	8.27	80.20	9.43	78.65	11.21
Mastoid_L	**85.67**	**4.43**	83.31	5.28	82.64	5.56	82.43	5.79	79.29	6.53	79.09	6.75
Mastoid_R	**84.91**	**4.67**	82.74	5.43	81.12	5.91	80.85	6.50	80.72	6.22	80.15	6.81
MiddleEar_L	78.59	4.16	77.43	4.61	**79.40**	**3.94**	71.62	7.31	73.24	5.31	71.27	8.25
MiddleEar_R	**76.78**	**4.34**	74.84	4.89	74.93	4.80	73.23	6.06	73.61	5.84	72.19	6.33
OpticNerve_L	**83.43**	**1.64**	78.26	2.56	77.81	2.66	78.49	2.43	76.73	2.81	75.82	2.94

Model	Our		E2MISeg		MS-TCNet		nnUNet		CoTr		HiFormer	
	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95
OpticNerve_R	**96.79**	**1.42**	94.68	2.05	93.87	2.36	94.04	2.18	93.47	2.46	93.16	2.72
OralCavity	**96.13**	**4.89**	94.54	5.42	94.18	5.70	93.55	5.87	93.06	5.91	92.79	6.80
Parotid_L	**94.74**	**5.06**	92.91	6.64	93.23	6.22	94.02	5.54	92.43	6.78	90.57	7.63
Parotid_R	87.61	5.13	84.38	7.23	85.24	6.67	**88.47**	**4.95**	86.30	6.86	85.28	8.01
PharynxConst	**88.45**	**4.32**	87.59	4.86	86.08	5.24	83.26	7.08	82.72	7.73	84.41	6.16
Pituitary	**91.77**	**1.14**	89.65	1.72	88.94	2.50	87.93	2.93	86.64	3.22	86.43	3.35
SpinalCord	89.68	2.19	**90.06**	**2.08**	87.79	2.63	87.34	2.78	86.26	3.05	85.35	3.17
Submandibular_L	**93.36**	**2.82**	92.10	2.98	90.35	3.12	89.81	3.45	87.44	4.03	86.56	4.73
Submandibular_R	**92.75**	**2.68**	92.23	2.90	90.73	3.19	90.17	3.38	88.91	3.87	88.24	4.02
TemporalLobe_L	**90.04**	**3.25**	89.17	3.82	87.31	4.71	88.42	4.36	86.34	6.53	87.15	5.66
TemporalLobe_R	**91.22**	**3.04**	89.46	3.99	88.19	4.53	88.23	4.44	87.47	5.62	87.28	4.89
Thyroid	**91.13**	**2.42**	88.38	3.54	87.94	4.24	86.13	4.85	82.24	5.41	83.49	4.67
TMjoint_L	85.58	4.60	85.02	4.81	**85.95**	**4.45**	85.27	4.84	84.63	4.93	84.32	5.06
TMjoint_R	88.84	4.22	**89.21**	**4.18**	87.31	5.28	87.94	4.72	87.10	5.74	85.24	6.80
Trachea	**83.90**	**3.16**	82.67	3.52	82.28	3.68	80.05	4.83	81.41	4.25	79.47	5.19
TympanicCavity_L	**88.78**	**2.23**	87.92	2.98	86.42	3.03	87.28	3.15	85.63	3.24	82.56	3.50
TympanicCavity_R	**86.63**	**2.45**	84.85	3.27	81.70	3.76	81.57	3.92	80.34	4.02	79.06	4.41
VestibulSemi_L	**89.67**	**1.30**	86.34	1.83	85.83	2.07	82.42	3.21	83.55	2.98	81.64	3.43
VestibulSemi_R	**85.72**	**1.42**	83.41	1.79	83.29	1.96	83.03	2.74	82.86	3.15	80.38	3.87
Average	**87.02**	**3.04**	85.47	3.63	84.90	3.80	84.39	4.09	83.36	4.48	82.62	4.86

Figure 8

Visual comparison of segmentation results on the SegRap2023 dataset. The first column shows the ground truth, and the remaining columns present the predictions of different methods. PUFormer produces more accurate segmentation masks with smaller errors across multiple organs.



5.1.3. Results on ACDC

Table 3 summarizes the performance of our method compared with other approaches on the ACDC dataset. PUFormer ranks second in terms of overall average performance. Specifically, it achieves the highest Dice coefficient of 91.74% on the right ventricle (RV). For the left ventricular myocardium (Myo), E2MISeg obtains the best result with 96.97% Dice, while nnUNet achieves the highest Dice of 95.92% on the left ventricle (LV). The superior performance of E2MISeg on Myo can be attributed to its progressive feature aggregation strategy, which enhances edge awareness and improves boundary refinement for the segmentation target. Figure 9 shows the qualitative visualization results. PUFormer produces segmentation contours in the right ventricle region that are closer to the ground truth annotations compared with other methods.

5.2. Ablation Study

To thoroughly investigate the contribution of each core component of PUFormer and their synergistic effects, we conducted a comprehensive ablation study on the HaN-Seg dataset. Specifically, we evaluated the impact of the three key modules by selectively removing them: the Prior-Shunted Self-Attention (PSSA) module, the Feature Aggregation Enhancement (FAE) module, and the Uncertainty-Guided Multi-scale Feature Fusion (UGMFF) module.

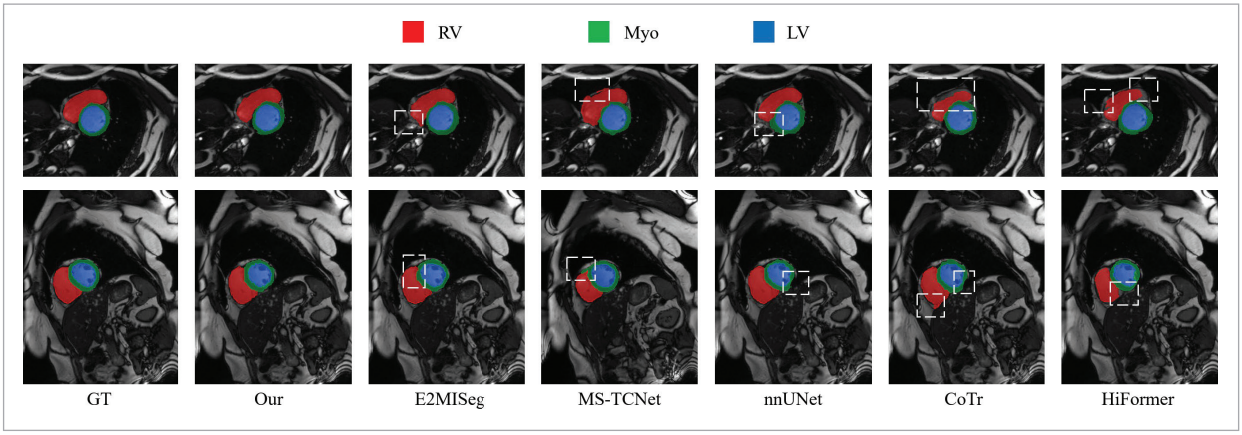
Table 3

Quantitative comparison of segmentation performance on the ACDC dataset in terms of Dice similarity coefficient (%) and HD95 (mm). The best results are highlighted in bold.

Model	Average		RV	Myo	LV	Params	Flops
	Dice	HD95	Dice	Dice	Dice		
HiFormer	90.40	3.83	88.24	87.63	95.30	22.36	54.31
CoTr	91.04	3.45	89.13	88.84	95.16	41.87	130.30
nnUNet	92.41	3.27	90.96	90.34	**95.92**	30.43	135.7
MS-TCNet	91.43	3.75	89.43	89.09	95.77	49.17	39.99
E2MISeg	**93.62**	**2.28**	91.15	**96.97**	93.62	85.13	61.03
Our	92.64	3.22	**91.74**	90.46	95.72	51.26	59.47

Figure 9

Visual comparison of segmentation results on the ACDC dataset. The first column shows the ground truth, and the remaining columns present the predictions of different methods. PUFormer produces segmentation contours in the right ventricle region that are closer to the ground truth annotations compared with other methods.



First, to validate the effectiveness of the prior knowledge-guided mechanism, we removed the structural prior map and replaced the Prior-Shunted Self-Attention in the P-Shunted Transformer with standard Shunted Self-Attention, denoted as Ablation-1. This experiment assesses the influence of explicit anatomical prior knowledge on the Transformer's ability to model spatial relationships.

Second, to evaluate the importance of the Feature Aggregation Enhancement module, we removed the FE Blocks from the decoder and retained only basic upsampling and skip-connection fusion, denoted as Ablation-2. This variant examines the contribution of jointly aggregating high-level semantic features, low-level textural details, and edge boundary information for fine-grained structure recovery.

Third, to verify the effectiveness of the uncertainty-driven fusion strategy, we removed the Uncertainty-Guided Gating (UGG) mechanism from the UGMFF module and adopted conventional multi-scale feature fusion, denoted as Ablation-3. This experiment analyzes the role of prediction uncertainty in enhancing feature representation and dynamic fusion in challenging regions.

Table 4 reports the ablation results. Removing any individual module leads to a noticeable performance degradation, demonstrating that PSSA, FAE, and UGMFF all contribute positively to head and neck OAR segmentation. When PSSA is removed (Ablation-1), the average Dice coefficient decreases by 1.07% and HD95 increases to 3.86 mm. This indicates that the structural prior knowledge effectively mitigates spatial localization ambiguity in complex anatomical environments, particularly benefiting small-volume and low-contrast organs by providing stable spatial relationship guidance for the Transformer. Removing the FAE module (Ablation-2) causes the most significant performance drop, with an average Dice reduction of 1.85%. This highlights that simple upsampling and skip connections are insufficient for recovering fine local boundaries and detailed structures in the highly complex head and neck region. The FAE module, by effectively integrating high-level semantics, low-level textures, and edge features, substantially alleviates semantic gaps across scales. Removing the UGMFF module (Ablation-3) results in an average Dice decrease of 0.89%, confirming that uncertainty-driven dynamic feature reweighting enhances multi-scale fusion in difficult regions. It enables the model to adaptively strengthen feature responses in high-uncertainty areas such as blurry boundaries and morphologically complex structures.

Table 4
Ablation analysis of the PUFormer module, examining the contributions of its configurations and submodules in HaN-Seg segmentation tasks.

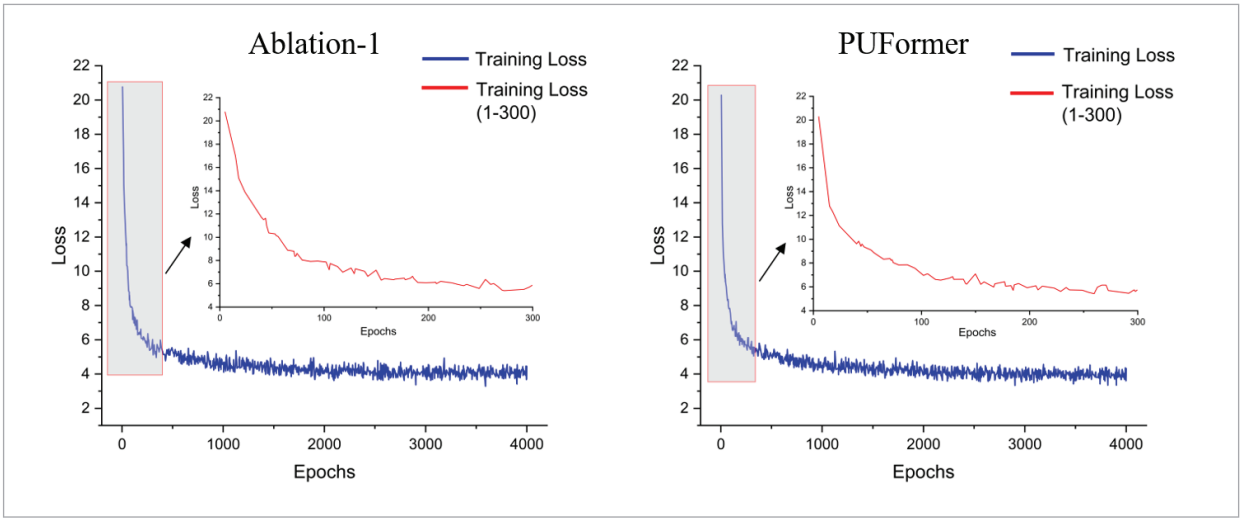
Model	Average		Params	FLOPs
	Dice	HD95		
Ablation-1	75.97	3.86	59.69	176.74
Ablation-2	75.19	4.18	50.87	150.35
Ablation-3	76.15	3.73	62.19	198.91
PUFormer	77.04	3.56	62.21	203.51

It is worth noting that although FAE yields the largest performance drop when removed individually, this does not imply that the other modules are less important. PSSA primarily provides stable spatial localization priors, while UGMFF focuses on dynamic feature reinforcement in challenging regions. These three modules operate at complementary levels – "spatial localization," "fine detail recovery," and "difficult-region fusion" – forming a synergistic framework. Retaining only a single module fails to simultaneously address global positioning, boundary refinement, and complex region modeling. The complete PUFormer achieves the best overall performance through their collaborative integration.

To further validate the effectiveness of the prior knowledge-guided learning strategy, we monitored the convergence behavior of the loss function during training on the HaN-Seg dataset. Figure 10 compares the training loss curves with and without the prior knowledge-guided strategy. The red box highlights a zoomed-in view of the loss curves from the 1st to 300th epoch (as indicated by the arrow). Compared with Ablation-1, the loss curve of the full PUFormer is noticeably smoother. Specifically, at the 100th epoch, Ablation-1 exhibits a loss value of 7.85, whereas PUFormer achieves a lower loss value of 6.93. These results demonstrate that the prior knowledge-guided learning strategy effectively directs the model to learn stable spatial relationships among organs during the early training stage, thereby improving training stability and convergence speed.

Figure 10

Comparison of training loss curves between the prior-guided and non-prior-guided learning strategies on the HaN-Seg dataset. The prior-guided strategy yields a smoother and faster converging loss curve, demonstrating its effectiveness in stabilizing model training, particularly during the early training stages.



6. Discussion

Over the past decade, deep learning-based medical image segmentation has witnessed remarkable progress. Nevertheless, complex tasks such as head and neck organs-at-risk (OAR) segmentation – characterized by densely packed anatomical structures, a large number of organs, and substantial scale variations – continue to pose significant challenges for existing CNN-Transformer hybrid architectures.

PUFormer differs fundamentally in design philosophy from representative models such as HiFormer and CoTr. HiFormer achieves strong global contextual modeling through a parallel dual-branch encoder and cross-attention mechanism. However, it underutilizes anatomical prior knowledge and lacks differentiated feature fusion based on regional prediction uncertainty. Consequently, PUFormer outperforms HiFormer by approximately 3–5 percentage points in Dice coefficient on small-volume, low-contrast, and anatomically stable organs such as the optic nerve, cochlea, lacrimal gland, and pituitary gland. These gains are primarily attributed to the structural prior guidance provided by PSSA and the uncertainty-aware fusion enabled by UGMFF. In contrast, for larger, high-contrast organs such as the brainstem and parotid gland, HiFormer's strong global modeling capability yields comparable or even marginally superior performance, particularly on organs with highly regular shapes.

Compared with CoTr, which enhances long-range dependencies by bridging CNN encoders and decoders via Transformer modules, PUFormer introduces dynamic structural prior guidance in the encoder and incorporates a Feature Aggregation Enhancement (FAE) module in the decoder. This design enables more effective integration of high-level semantics, low-level spatial details, and edge information. As a result, PUFormer achieves superior segmentation accuracy and boundary consistency (HD95) on organs with irregular morphology or ambiguous boundaries, such as the submandibular gland, laryngeal structures, and esophagus.

Overall, by incorporating anatomical prior guidance and uncertainty-driven multi-scale fusion, PUFormer effectively addresses the limitations of existing hybrid models in complex head and neck scenarios. It is particularly well-suited for OAR segmentation tasks characterized by numerous organs, tight anatomical relationships, and a high proportion of small structures. However, its advantage is relatively diminished in scenarios involving large dynamic deformations, such as cardiac imaging.

On the ACDC dataset, PUFormer achieves an average Dice coefficient of 92.64%, which is lower than that of E2MISeg (93.62%). We attribute this performance gap to two main factors. First, the distance-transform-based structural prior relies heavily on the relative spatial stability of organ positions. In cardiac MRI, however, the position and morphology of the heart vary considerably due to respiratory motion, cardiac contraction, inter-patient anatomical differences, and variations in scanning position, such that fixed priors may introduce spatial bias rather than effective guidance. Second, while the UGMFF module effectively enhances high-entropy regions corresponding to boundaries and small organs in CT images, high-entropy regions in cardiac MRI frequently coincide with motion artifacts rather than true anatomical boundaries, causing the module to inadvertently amplify noise in this setting.

In summary, PUFormer is best suited for CT-based segmentation tasks with anatomically stable structures and consistent image contrast. For tasks involving large morphological deformations or significant cross-modality differences, adaptive prior mechanisms tailored to deformable organs and modality-aware training strategies will be required in future work.

7. Conclusion

This paper proposes PUFormer, a novel CNN-Transformer hybrid network for high-precision automatic segmentation of head and neck organs-at-risk (OARs). By organically integrating anatomical prior knowledge, uncertainty estimation, and multi-scale feature fusion, the proposed method effectively improves both segmentation accuracy and robustness for small organs in complex anatomical environments.

Extensive experiments on two head and neck OAR datasets demonstrate the effectiveness of PUFormer. On the HaN-Seg dataset, PUFormer achieves an average Dice coefficient of 77.04%, outperforming the second-best method by 1.34% (77.04% vs. 75.70%). On the SegRap2023 dataset, it attains 87.02% Dice, surpassing the second-best method by 1.55% (87.02% vs. 85.47%). These results validate the effectiveness of our approach.

Compared with existing methods, the core advantage of PUFormer lies in the synergistic integration of structural prior guidance, feature aggregation enhancement, and uncertainty-driven multi-scale weighting. This complementary design effectively addresses the limitations of current approaches in exploiting anatomical priors and handling challenging multi-scale regions, providing more reliable technical support for accurate OAR segmentation in head and neck radiotherapy planning.

In future work, we will focus on further improving the modeling capability for extremely small organs and ambiguous boundaries, reducing computational complexity, and exploring multi-modal learning frameworks to better meet clinical requirements.

Funding

This work was supported in part by the Health Commission of Sichuan Province Medical Science and Technology Program (No. 24WSXT105), the Key Research and Development Project for High-Quality Development of Zigong First People's Hospital and Zigong Academy of Medical Sciences in 2024 (No. 2024GZL04), the Sichuan Provincial Key Research and Development Program (No. 2025YFHZ0006), the Sichuan University of Science and Engineering Talent Introduction Program (No. 2025RCZ064).

References

1. Ao, Y., Shi, W., Ji, B., Zhang, H. MS-TCNet: An Effective Transformer-CNN Combined Network Using Multi-Scale Feature Learning for 3D Medical Image Segmentation. Computers in Biology and Medicine, 2024, 170, 108057. https://doi.org/10.1016/j.compbiomed.2024.108057
2. Azad, R., Aghdam, E. K., Rauland, A., Kazerouni, A., Fayyaz, M., Merhof, D. Medical Image Segmentation Review: The Success of U-Net. arXiv Preprint, 2022. Available at: https://arxiv.org/abs/2211.14830
3. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Gon-

zalez Ballester, M. A., Sanroma, G., Napel, S., Petersen, S., Tziritas, G., Grinias, E., Khened, M., Kollerathu, V. A., Krishnamurthi, G., Rohe, M.-M., Pennec, X., Sermesant, M., Isensee, F., Jager, P., Maier-Hein, K. H., Full, P. M., Wolf, I., Engelhardt, S., Baumgartner, C. F., Koch, L. M., Wolterink, J. M., Isgum, I., Jang, Y., Hong, Y., Patravali, J., Jain, S., Humbert, O., Jodoin, P.-M. Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? IEEE Transactions on Medical Imaging, 2018, 37(11), 2514-2525. https://doi.org/10.1109/TMI.2018.2837502

4. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv Preprint, 2021. Available at: https://arxiv.org/abs/2102.04306
5. Chen, X., Sun, S., Bai, N., Han, K., Liu, Q., Yao, Y. A Deep Learning-Based Auto-Segmentation System for Organs-at-Risk on Whole-Body Computed Tomography Images for Radiation Therapy. Radiotherapy and Oncology, 2021, 160, 175-184. https://doi.org/10.1016/j.radonc.2021.04.019
6. Cheng, H., Yu, X., Wu, S., Fang, L., Cao, C., Zhang, J., Liu, T., Zhu, D., Zhong, W., Ma, P. DCMM-Transformer: Degree-Corrected Mixed-Membership Attention for Medical Imaging. Proceedings of the AAAI Conference on Artificial Intelligence, 2026, 40(5), 3228-3236. https://doi.org/10.1609/aaai.v40i5.37317
7. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2016, 424-432. https://doi.org/10.1007/978-3-319-46723-8_49
8. Feng, Y., Wang, W., Zhang, X., Liu, H. Medical Priors-Guided Feature Learning Network on Multimodal Imaging Raw Data for Brain Tumor Segmentation. Biomedical Signal Processing and Control, 2025, 107, 107855. https://doi.org/10.1016/j.bspc.2025.107855
9. Gao, Y., Huang, R., Yang, Y., Zhang, C., Chen, X. FocusNetv2: Imbalanced Large and Small Organ Segmentation with Adversarial Shape Constraint for Head and Neck CT Images. Medical Image Analysis, 2021, 67, 101831. https://doi.org/10.1016/j.media.2020.101831
10. Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S. CA-Net: Comprehensive Attention Convolutional Neural Networks for Explainable Medical Image Segmentation. IEEE Transactions on Medical Imaging, 2020, 40(2), 699-711. https://doi.org/10.1109/TMI.2020.3035253
11. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., Xu, D. UNETR: Transformers for 3D Medical Image Segmentation. WACV, 2022, 574-584. https://doi.org/10.1109/WACV51458.2022.00066
12. He, A., Wang, K., Li, T., Duan, Y., Fu, H. H2Former: An Efficient Hierarchical Hybrid Transformer for Medical Image Segmentation. IEEE Transactions on Medical Imaging, 2023, 42(9), 2763-2775. https://doi.org/10.1109/TMI.2023.3264513
13. Heidari, M., Kazerouni, A., Soltany, M., Azad, R., Merhof, D. HiFormer: Hierarchical Multi-Scale Representations Using Transformers for Medical Image Segmentation. WACV, 2023, 6202-6212. https://doi.org/10.1109/WACV56688.2023.00613
14. Huang, X., Deng, Z., Li, D., Yuan, X., Fu, Y. MISSFormer: An Effective Transformer for 2D Medical Image Segmentation. IEEE Transactions on Medical Imaging, 2022, 42(5), 1484-1494. https://doi.org/10.1109/TMI.2022.3230943
15. Ibragimov, B., Xing, L. Segmentation of Organs-at-Risk in Head and Neck CT Images Using Convolutional Neural Networks. Medical Physics, 2017, 44(2), 547-557. https://doi.org/10.1002/mp.12045
16. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., Maier-Hein, K. H. nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation. Nature Methods, 2021, 18(2), 203-211. https://doi.org/10.1038/s41592-020-01008-z
17. Jiang, L., Xu, W., Zheng, X., Zhang, Y., Zekun Jiang, Chong Jiang, Chen, Y., Ji, Y., Liu, S., Liu, J., Xu, J. E2MISeg: Enhancing Edge-Aware 3D Medical Image Segmentation via Feature Progressive Co-Aggregation. Expert Systems with Applications, 2026, 296, 128861. https://doi.org/10.1016/j.eswa.2025.128861
18. Kuang, H., Wang, Y., Tan, X., Liu, J., Zhang, Y. LW-CTrans: A Lightweight Hybrid Network of CNN and Transformer for 3D Medical Image Segmentation. Medical Image Analysis, 2025, 102, 103545. https://doi.org/10.1016/j.media.2025.103545
19. Lan, L., Cai, P., Jiang, L., Zhang, X. BRAU-Net++: U-Shaped Hybrid CNN-Transformer Network for Medical Image Segmentation. arXiv Preprint, 2024. Available at: https://arxiv.org/abs/2401.00722

20. Luo, X., Fu, J., Zhong, Y., Liu, S., Han, B., Astaraki, M., Bendazzoli, S., Toma-Dasu, I., Ye, Y., Chen, Z., Xia, Y., Su, Y., Ye, J., He, J., Xing, Z., Wang, H., Zhu, L., Yang, K., Fang, X., Wang, Z., Lee, C. W., Park, S. J., Chun, J., Ulrich, C., Maier-Hein, K. H., Ndipenoch, N., Miron, A., Li, Y., Zhang, Y., Chen, Y., Bai, L., Huang, J., An, C., Wang, L., Huang, K., Gu, Y., Zhou, T., Zhou, M., Zhang, S., Liao, W., Wang, G., Zhang, S. SegRap2023: A Benchmark of Organs-at-Risk and Gross Tumor Volume Segmentation for Radiotherapy Planning of Nasopharyngeal Carcinoma. Medical Image Analysis, 2025, 101, 103447. https://doi.org/10.1016/j.media.2024.103447

21. Mao, A., Mohri, M., Zhong, Y. Cross-Entropy Loss Functions: Theoretical Analysis and Applications. Proceedings of the 40th International Conference on Machine Learning (ICML), 2023, 23803-23828. Available: https://proceedings.mlr.press/v202/mao23a.html.

22. Miao, K., Gokul, A., Singh, R., Kim, J. Prior Knowledge-Guided Attention in Self-Supervised Vision Transformers. arXiv Preprint, 2022. https://arxiv.org/abs/2209.03745

23. Milano, M. T., Vargo, J. A., Yorke, E. D., Ronckers, C. M., Kremer, L. C., Chafe, S. M. J., van Santen, H. M., Marks, L. B., Bentzen, S. M., Constine, L. S., Vogelius, I. R. Primary Hypothyroidism in Childhood Cancer Survivors Treated with Radiation Therapy: A PENTEC Comprehensive Review. International Journal of Radiation Oncology, Biology, Physics, 2024, 119(2), 482-493. https://doi.org/10.1016/j.ijrobp.2021.02.001

24. Milletari, F., Navab, N., Ahmadi, S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. Proceedings of the Fourth International Conference on 3D Vision (3DV), 2016, 565-571. https://doi.org/10.1109/3DV.2016.79

25. Nutting, C. M., Morden, J. P., Harrington, K. J., Urbano, T. G., Bhide, S. A., Clark, C., Miles, E. A., Miah, A. B., Newbold, K., Tanay, M., Adab, F., Jefferies, S. J., Scrase, C., Yap, B. K., A'Hern, R. P., Sydenham, M. A., Emson, M., Hall, E., PARSPORT Trial Management Group. Parotid-Sparing Intensity Modulated Versus Conventional Radiotherapy in Head and Neck Cancer (PARSPORT): A Phase 3 Multicentre Randomised Controlled Trial. Lancet Oncology, 2011, 12(2), 127-136. https://doi.org/10.1016/S1470-2045(10)70290-4

26. Petras, K. G., Rademaker, A. W., Refaat, T., Choi, M., Thomas, T. O., Pauloski, B. R., Mittal, B. B. Dose-Volume Relationship for Laryngeal Substructures and Aspiration in Patients with Locally Advanced Head-and-Neck Cancer. Radiation Oncology, 2019, 14(1), 49. https://doi.org/10.1186/s13014-019-1247-7

27. Podobnik, G., Strojan, P., Peterlin, P., Ibragimov, B., Vrtovec, T. HaN-Seg: The Head and Neck Organ-at-Risk CT and MR Segmentation Dataset. Medical Physics, 2023, 50(3), 1917-1927. https://doi.org/10.1002/mp.16197

28. Qiu, P., Yang, J., Kumar, S., Zhang, L. AgileFormer: Spatially Agile and Scalable Transformer for Medical Image Segmentation. Biomedical Signal Processing and Control, 2026, 112, 108842. https://doi.org/10.1016/j.bspc.2025.108842

29. Ren, S., Zhou, D., He, S., Feng, J. Shunted Self-Attention via Multi-Scale Token Aggregation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 10853-10862. https://doi.org/10.1109/CVPR52688.2022.01058

30. Roy, A. G., Navab, N., Wachinger, C. Concurrent Spatial and Channel Squeeze & Excitation in Fully Convolutional Networks. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2018, 421-429. https://doi.org/10.1007/978-3-030-00928-1_48

31. Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., Yang, M. H. Transformers in Medical Imaging: A Survey. Medical Image Analysis, 2023, 88, 102802. https://doi.org/10.1016/j.media.2023.102802

32. Wang, Y., Liang, S., Xue, L., Zhang, H. MPF-Net: A Multi-Scale Feature Learning Network Enhanced by Prior Knowledge Integration for Medical Image Segmentation. Alexandria Engineering Journal, 2025, 128, 200-212. https://doi.org/10.1016/j.aej.2025.05.058

33. Wei, X., Sun, J., Su, P., Zhang, Y. BCL-Former: Localized Transformer Fusion with Balanced Constraint for Polyp Image Segmentation. Computers in Biology and Medicine, 2024, 182, 109182. https://doi.org/10.1016/j.compbiomed.2024.109182

34. Xie, Y., Zhang, J., Shen, C., Xia, Y. CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2021, 171-180. https://doi.org/10.1007/978-3-030-87193-2_17

35. Xu, L., Chen, M., Cheng, Y., Li, Z. MCPA: Multi-Scale Cross Perceptron Attention Network for 2D Medical Image Segmentation. Complex & Intelligent Systems, 2025, 11(1), 75. https://doi.org/10.1007/s40747-024-01671-1

36. Xue, B., Zheng, Q., Li, Z., et al. ISAR Weak Feature Enhancement with Perturbation Defense Using Hybrid Clustering Oversegmentation. IEEE Transactions on Aerospace and Electronic Systems, 2024, 60(5), 6256-6274. https://doi.org/10.1109/TAES.2024.3408139

37. Xue, B., Zheng, Q., Li, Z., Wang, J., Mu, C., Yang, J., Fan, H., Feng, X., Li, X. Perturbation Defense Ultra High-Speed Weak Target Recognition. Engineering Applications of Artificial Intelligence, 2024, 138, 109420. https://doi.org/10.1016/j.engappai.2024.109420

38. Yi, Y., Duan, Y., Wu, X., Zhang, H. MSPCNF-Net: Multi-Scale Parallel Cross-Neighborhood Fusion Network for Medical Image Segmentation. Knowledge-Based Systems, 2025, 114624. https://doi.org/10.1016/j.knosys.2025.114624

39. Yuan, F., Zhang, Z., Fang, Z. An Effective CNN and Transformer Complementary Network for Medical Image Segmentation. Pattern Recognition, 2023, 136, 109228. https://doi.org/10.1016/j.patcog.2022.109228

40. Zeng, Z., Xiao, J., Yi, S., Liu, Q., Zhu, Y. M3-TransUNet: Medical Image Segmentation Based on Spatial Prior Attention and Multi-Scale Gating. Journal of Imaging, 2026, 12(1), 15. https://doi.org/10.3390/jimaging12010015

41. Zhang, D., Zhang, J., Zhang, Q., Wang, Y. Automatic Pancreas Segmentation Based on Lightweight DCNN Modules and Spatial Prior Propagation. Pattern Recognition, 2021, 114, 107762. https://doi.org/10.1016/j.patcog.2020.107762

42. Zhang, J., Wang, Q., Tan, Y., Li, X. Retinal Vessel Image Segmentation Based on Edge Information Guidance and Multi-Scale Feature Fusion. Biomedical Signal Processing and Control, 2026, 113, 109055. https://doi.org/10.1016/j.bspc.2025.109055

43. Zhang, P., Cheng, Y., Tamura, S. Shape Prior-Constrained Deep Learning Network for Medical Image Segmentation. Computers in Biology and Medicine, 2024, 180, 108932. https://doi.org/10.1016/j.compbiomed.2024.108932

44. Zhang, Q., Wang, J., Wu, W., Li, H. UANeT: An Uncertainty-Aware Network for Feature Calibration and Boundary Refinement in Medical Image Segmentation. Biomedical Signal Processing and Control, 2026, 112, 108413. https://doi.org/10.1016/j.bspc.2025.108413

45. Zheng, F., Chen, X., Liu, W., Zhang, Y. SmaFormer: Synergistic Multi-Attention Transformer for Medical Image Segmentation. BIBM, 2024, 4048-4053. https://doi.org/10.1109/BIBM62325.2024.10822736

46. Zhong, J., Tian, W., Xie, Y., Liu, Z., Ou, J., Tian, T. PM-FSNet: Polarized Multi-Scale Feature Self-Attention Network for Lightweight Medical Image Segmentation. Computer Methods and Programs in Biomedicine, 2025, 261, 108611. https://doi.org/10.1016/j.cmpb.2025.108611



This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) License (http://creativecommons.org/licenses/by/4.0/).