EASAG-Net: A Residual-Decoder Network with Multi-Scale Edge-Aware Spatial Attention for Prostate Lesion Segmentation in T2W MRI

Anonymous WACV Algorithms Track submission

Paper ID 1532

Abstract

Accurate detection and segmentation of prostate lesions are essential for diagnosis and treatment planning. Prostate MRI is a standard imaging modality in detection, staging, and surveillance due to higher resolution and safety. However, challenges such as low contrast, class imbalance between foreground and background, small lesion size and ambiguous boundaries persist. Prior works often rely on multi-parametric MRI sequences (T2W, ADC, DWI), leads an extra burden of image misalignment, noise, size of dataset or clinical impracticalities. Due to these empirical challenges, most of the approaches lacks in generalization and reproducibility. We propose a lightweight integrated architecture using a multi-scale Edge-Aware Spatial Attention Gate (EASAG) within a residual decoder to enhance both boundary precision and spatial context modeling. The proposed method is evaluated on two-site patient cohort, while the model training was conducted on PI-CAI dataset, validation performed using an external cohort of 82 csPCa subjects (Gleason grade $\geq$ 2) in Prostate158 dataset. Our model outperformed state-of-the-art approaches based on cohort-wise prostate lesion segmentation and yield 3.57±0.7% gain in Dice score, 2.18±0.50 mm reduction in Hausdorff distance (HD), and 10.17±3.3% increase in boundary precision. These results highlight the effectiveness and generalization of our attention-enhanced design for clinically practical, single-modality prostate lesion segmentation.

1. Introduction

Prostate cancer (PCa) is the second most prevailing cancer and a leading cause of cancer-related mortality among men globally [23, 25]. Early detection significantly improves prognosis, especially for clinically significant PCa. Screening typically involves serum prostate-specific antigen (PSA) testing [16], followed by biopsy and imaging. Compared to ultrasound and biopsy alone, multi-parametric MRI (mpMRI) offers superior soft-tissue contrast and has become central to PCa diagnosis [3, 26, 28]. T2-weighted (T2W) MRI, in particular, is critical for localizing lesions and guiding interventions [2].

In prostate care and management, the standard diagnostic grading systems such as pathologic Gleason grade group (GGG) and clinical prostate imaging reporting and data system (PI-RADS) based assessments [2, 9] help stratify cancer risk. However, manual interpretation of MRI scans is prone to inter-observer variability, subjectivity, and annotation inconsistency [14]. In the past works, an automated, DL-based segmentation methods have shown promise in mitigating these limitations, enhancing reproducibility and reducing patient workload [1, 8].

While several deep learning approaches have shown promising results for prostate lesion segmentation, many rely on multi-channel inputs, stacking T2W, apparent diffusion coefficient (ADC), and diffusion-weighted imaging (DWI) MRIs [30] to extract richer information. However, this multi-modality requirement introduces practical challenges: images from different sequences may suffer from inconsistent resolution, motion artifacts, or misalignment, and in some clinical settings, certain modalities may be unavailable altogether. Moreover, many prior models have been trained and evaluated on limited or internal datasets, often lacking rigorous external validation, which raises concerns about generalizability and real-world deployment. From a modeling perspective, prostate lesion segmentation is inherently difficult due to the small size and ill-defined boundaries of many lesions. Conventional encoder-decoder architectures may fail to capture sufficient fine-grained edge information or struggle to integrate spatial context at multiple resolutions. As a result, segmentation masks may either under-segment the lesion or include excess background, particularly in cases where the lesion closely resembles surrounding tissue.

To address these challenges, we propose a novel deep learning framework that performs accurate lesion segmentation using only T2W MRI. Our architecture integrates a

Multi-scale Edge-Aware Spatial Attention Gate (EASAG) within a residual decoder design to enhance both boundary delineation and contextual understanding. The model avoids reliance on multi-modal inputs, reducing data acquisition overhead and improving clinical accessibility.

Our main contributions can be summarized as follows:

- In contrast to previously proposed models leveraging stacked multi-parametric MRI (T2W, ADC and DWI), we consider developing a lightweight, T2W-only lesion segmentation architecture tailored for real-world clinical deployment, thereby enhancing its clinical deployability across settings where ADC\DWI may be misaligned, unavailable, or low-quality.
- We introduce EASAG — a novel edge-aware spatial attention mechanism using edge details and multi-scale spatial context fusion to improve boundary sensitivity and localization.
- We incorporate a residual decoder that enables iterative feature refinement across decoding stages.
- We evaluate the proposed model using multi-site validation cohort, including ablation studies to establish the robustness and generalizability of our design.

2. Related Works

Prostate lesion segmentation from multi-parametric MRI has been extensively investigated using deep learning architectures, particularly encoder-decoder-based CNNs. Early methods adopted modified U-Net architectures to delineate lesion boundaries from T2-weighted (T2W) or ADC images [11, 20, 21]. For instance, Hambarde et al.[11] trained a radiomics-informed U-Net on 40 T2W cases, reserving 10 internal data for testing (415 slices), achieving 91.76% Dice similarity. Cao et al.[5] proposed a CNN-CRF-based meta-learning approach on 397 biparametric MRI cases with Gleason score $\geq$ 6 (with 20% held out for testing), achieving $\sim$40% Dice score on biparametric (T2W with ADC) MRI. However, many such pipelines relied on handcrafted features or cascaded stages, limiting generalizability. An end-to-end CNN-based strategies have been gaining attraction in previous years. Eidex et al.[7] proposed a cascaded ROI-scoring CNN on 77 T1-weighted MRI cases, achieving Dice of 85% through coarse-to-fine refinement. Gunashekar et al.[10] used Grad-CAM-based interpretability modules with U-Net trained on 122 multi-parametric MRI cases and evaluated on 15 whole-mount histopathology scans, reporting 31% of Dice score. Furthermore, these end-to-end deep segmentation models have also improved robustness, especially when trained on larger datasets such as the PI-CAI challenge or private hospital cohorts [6, 27] However, many of these methods report high Dice scores based on relatively small test sets, often limited to internal validation on a small external cohort which restricts conclusions about model generalization.

In addition to several deep learning based segmentation networks, Song et al.[24] used a deep multi-scale attention U-Net trained on the PROSTATEX dataset of 97 patients, reporting 70% dice similarity and 86.5% sensitivity. Meanwhile, Ren et al.[19] introduced a 3D attention-guided model with ASPP blocks, yielding Dice score of 93.9% on 180 DWI cases. Attention mechanisms have become a key component in medical image segmentation architectures. Attention U-Net [17] model introduced the spatial attention gates to selectively focus on relevant regions during decoding, improving performance on challenging organs like the pancreas. Channel and spatial attention modules, such as those found in CBAM [29] and SE blocks [12], have been applied to emphasize semantically rich feature maps and suppress irrelevant activations. However, many of these designs operate on single-scale features or late-stage decoder outputs, which limits their ability to model spatial detail across multiple resolutions. Moreover, attention modules are often sensitive to noise in low-contrast lesions, as seen in prostate cancer segmentation where boundary precision is critical.

ProLesNet [30] recently introduced a strong baseline for prostate lesion segmentation, using stacked T2W, ADC, and DWI images along with spatial and channel attention to improve lesion localization. In a past work, an edge-aware attention decoders were introduced to explicitly guide boundary refinement in segmentation tasks [18, 31]. These methods typically extract edge cues from early encoder features and fuse them with high-level context to enhance delineation. However, their application in prostate lesion segmentation remains limited.

3. Methodology

3.1. Overall Architecture

Our proposed model adopts a U-shaped encoder-decoder structure inspired by the classic U-Net [20], but significantly enhanced through two key innovations: (1) a residual decoder block (RDB) for improved feature reuse and boundary refinement, and (2) a novel Multi-scale Edge-Aware Spatial Attention Gate (EASAG) that selectively enhances spatial detail with edge guidance at each decoder level. The encoder extracts hierarchical features, while the decoder progressively upsamples and fuses them with attention-guided skip connections to reconstruct fine-grained lesion boundaries. Figure 1 illustrates the overall design.

Let $X \in \mathbb{R}^{H \times W \times C}$ denote the input T2W image. The encoder $\mathcal{E}$ extracts multi-resolution features as:

$$F_e = \mathcal{E}(X) = \{f_1, f_2, \dots, f_L\}, \quad f_l \in \mathbb{R}^{H_l \times W_l \times C_l} \quad (1)$$

where L is the number of encoder stages, and each f_l represents the output feature map at level l with progressively reduced spatial resolution.

During decoding, each level l performs three operations: (1) the decoder feature d_{l+1} is upsampled, (2) the encoder skip connection f_l is refined using EASAG with a gating signal g_l from the decoder, (3) the upsampled decoder feature and EASAG-enhanced skip feature are fused via a residual decoder block:

$$f'_l = \text{EASAG}(f_l, g_l) \tag{2}$$

$$d_l = \mathcal{D}_{\text{res}}(d^{\uparrow}_{l+1}, f'_l) \tag{3}$$

Here, EASAG enhances f_l using contextual and edge cues, and $\mathcal{D}_{\text{res}}$ is a residual decoder block that fuses attention-guided features with decoder context. The upsampling operation $d^{\uparrow}_{l+1}$ is implemented using transposed convolution.

Finally, the output segmentation map $\hat{Y} \in [0,1]^{H \times W}$ is generated via a 1×1 convolution followed by sigmoid or softmax activation:

$$\hat{Y} = \sigma(\text{Conv}_{1\times1}(d_1)) \tag{4}$$

where σ is the sigmoid function for binary segmentation, or softmax in the multi-class setting. The decoder thus combines multi-level upsampling with edge-aware spatial guidance to achieve high-precision lesion delineation in T2W MRI.

3.2. Edge-Aware Spatial Attention Gate (EASAG)

Prostate lesion boundaries in T2-weighted MRI are often subtle and fragmented, posing a challenge for traditional attention mechanisms that rely solely on semantic context. To address this, we propose the Multi-scale Edge-Aware Spatial Attention Gate (EASAG), implemented via a learnable fusion of contextual attention and explicit edge enhancement. EASAG is applied at each decoder stage to enhance skip-connected features before fusion with the upsampled decoder stream.

Our EASAG module consists of two main components: (1) A Grid Attention Block that learns to emphasize semantically important regions conditioned on a gating signal, and (2) A fixed 3D Laplacian edge detector that extracts boundary-sensitive features from skip inputs.

Let $x \in \mathbb{R}^{C \times D \times H \times W}$ be the encoder feature (skip connection), and $g \in \mathbb{R}^{C' \times D' \times H' \times W'}$ be the gating signal from a deeper decoder level.

Step 1: Contextual Attention via Grid Attention Block We use the gating signal g to modulate the encoder feature map x using a grid attention mechanism [22], which computes a soft attention map highlighting semantically:

$$\begin{aligned} G &= \text{GridAttn}(x, g) \\ &= \mathcal{W}_1\left(x \cdot \sigma\left(\psi^\top \cdot \text{ReLU}(\theta(x) + \phi(g))\right)\right) \end{aligned} \tag{5}$$

Here, θ and ϕ are learnable $1 \times 1 \times 1$ or $3 \times 3 \times 3$ convolutions applied to the encoder input x and the gating signal g, respectively. Their outputs are combined and passed through a ReLU non-linearity and a ψ projection to produce a coarse attention map. The sigmoid function σ transforms this into a spatial attention weight map.

To ensure alignment with the original input resolution, the attention weights are upsampled via trilinear interpolation before being applied element-wise to the input feature map x. The modulated result is passed through a $1 \times 1 \times 1$ convolution and instance normalization $\mathcal{W}_1$ to obtain the attention-weighted output G.

Step 2: Edge Feature Extraction via Laplacian Filtering In parallel, we apply a non-trainable 3D Laplacian kernel K_{lap} to extract edge information from the same encoder feature:

$$K_{\text{lap}} = \left[\begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -1 & 0 \\ -1 & 6 & -1 \\ 0 & -1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right]$$

This results in an edge-emphasized output $E \in \mathbb{R}^{1 \times D \times H \times W}$:

$$E = K_{\text{lap}} * x \tag{6}$$

Step 3: Feature Fusion and Attention Enhancement The attention-weighted context G and the edge map E are concatenated channel-wise and passed through a $1 \times 1 \times 1$ convolutional block with batch normalization and ReLU:

$$F' = \mathcal{F}_{\text{conv}}(\text{Concat}(G, E)) \tag{7}$$

where the fusion function is defined as:

$$\mathcal{F}_{\text{conv}}(z) = \text{ReLU}\left(\text{BN}\left(\text{Conv}_{1\times1\times1}(z)\right)\right) \tag{8}$$

The final output F' is passed to the next decoder stage for further refinement.

3.3. Encoder and Residual Decoder Blocks

Our architecture builds on a U-shaped encoder-decoder backbone, where the encoder abstracts spatial context and the decoder refines lesion localization. The decoder consists of three tightly integrated modules:

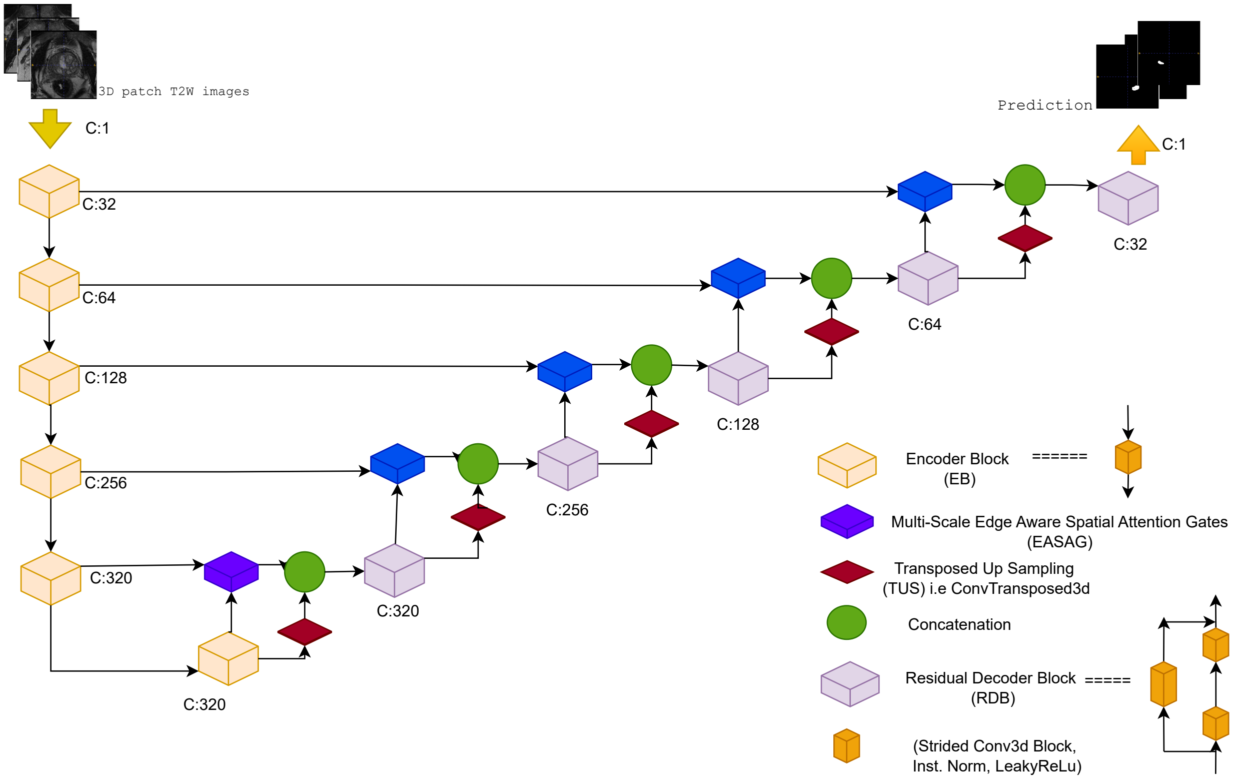
3D patch T2W images
C:1
C:32
C:64
C:128
C:256
C:320
C:320
C:320
C:256
C:128
C:64
C:32
Prediction
C:1
Encoder Block (EB)
======
Multi-Scale Edge Aware Spatial Attention Gates (EASAG)
Transposed Up Sampling (TUS) i.e ConvTransposed3d
Concatenation
Residual Decoder Block (RDB)
=====
(Strided Conv3d Block, Inst. Norm, LeakyReLu)

Figure 1. Overview of the proposed overall architecture

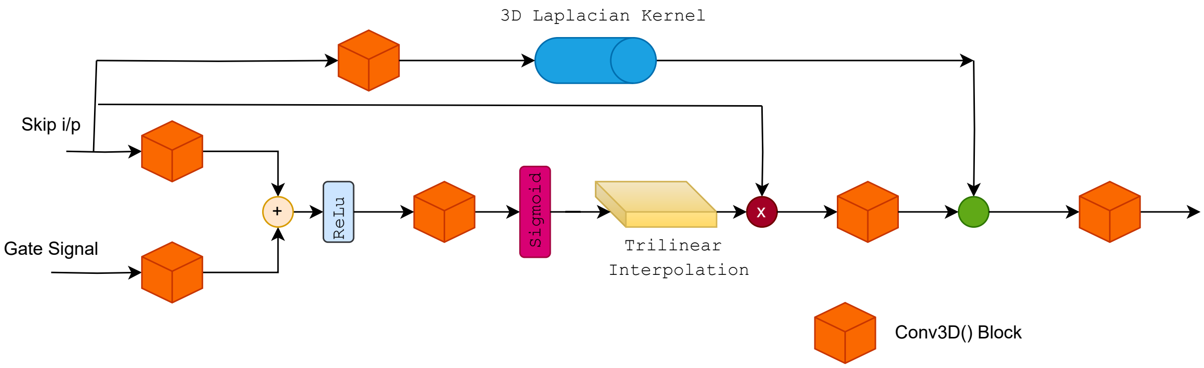

Figure 2. Internal structure of the proposed Multi-scale Edge-Aware Spatial Attention Gate (EASAG). The module combines grid attention conditioned on gating signals with an edge map generated by a fixed 3D Laplacian kernel. The two are concatenated and fused via a $1 \times 1 \times 1$ convolution to enhance feature refinement at decoder stages.

i. Encoder Blocks (EB) for deep hierarchical multi-stage feature extraction.
ii. Transposed Upsampling (TUS) for resolution recovery.
iii. Residual Decoder Blocks (RDB) for spatial refinement and gradient preservation

Encoder Block (EB) The encoder consists of a sequence of stages EB_l, each composed of a single 3D convolution block:

$$f_l = \phi\left(\text{IN}\left(\text{Conv}_{3\times3}(f_{l-1})\right)\right) \tag{9}$$

where ϕ denotes a non-linear activation function

(LeakyReLU), IN refers to Instance Normalization. Each encoder block reduces spatial resolution using a strided convolution, forming a hierarchical feature pyramid. Unlike the decoder, the encoder uses a **single convolution per stage** and omits 2 Convolution block or residual connections. This design choice is empirically motivated: as shown in our ablation study, increased encoder complexity (e.g., stacked or residual convolution) resulted in degraded performance, and found that simpler encoder generalize better in our task.

Transposed Upsampling (TUS) To recover spatial resolution in the decoder, we use a transposed convolution operation:

$$d_l^{\uparrow} = \mathrm{TUS}_l(d_{l+1}) = \mathrm{ConvTranspose}(d_{l+1}) \tag{10}$$

This produces an upsampled decoder feature $d_l^{\uparrow}$, which is then fused with the EASAG-enhanced skip connection.

Residual Decoder Block (RDB) After upsampling and edge-aware skip refinement, the residual decoder block fuses decoder context with attention-enhanced features to refine the output:

$$\begin{aligned} z_l &= \mathrm{Concat}(d_{l+1}^{\uparrow}, f_l') \\ d_l &= \mathrm{ResidualBlock}(z_l) + z_l \end{aligned} \tag{11}$$

The residual unit is defined as:

$$\mathrm{ResidualBlock}(x) = \phi\left(\mathrm{IN}\left(\mathrm{Conv}_2\left(\phi\left(\mathrm{IN}\left(\mathrm{Conv}_1(x)\right)\right)\right)\right)\right) \tag{12}$$

Here, ϕ denotes a non-linear activation (e.g., LeakyReLU), and IN refers to instance normalization. This residual formulation helps preserve low-level spatial features, improves gradient flow, and enhances lesion boundary delineation across decoder stages.

4. Experimental Setup

4.1. Dataset Description

We utilize two publicly available prostate MRI datasets:

i **PI-CAI (Prostate Imaging–Cancer AI):** A large-scale dataset of 1,500 mpMRI scans collected from over 100 institutions using Siemens and Philips 3T scanners. Among these, 425 cases are csPCa-positive, but only 219 scans include expert-annotated lesion masks, which we use as our training set. These 219 cases include diverse Gleason Grade Groups (GGG): GGG 1: 3, GGG 2: 128, GGG 3: 49, GGG 4: 18, GGG 5: 18. [pi-cai.grand-challenge.org]

ii **Prostate158:** A multi-center dataset with 158 prostate MRI cases acquired using Siemens VIDA and Skyra 3.0T scanners. Of the publicly available 139 cases, 82 were csPCa-confirmed subjects which is our external test set. Their GGG distribution is: GGG 1: 9, GGG 2: 29, GGG 3: 19, GGG 4: 18, GGG 5: 7. [zenodo.org/records/6481141]

4.2. Preprocessing

All experiments are performed on the axial axis, using 3D volumes oriented along the z-axis. Instead of processing full volumes, we adopt a patch-based approach, where fixed-size patches of (16, 224, 224) are extracted during training and inference. Here, 16 refers to the number of axial slices (z-depth), and 224×224 is the in-plane spatial resolution (y, x). These patches are sampled around the prostate region to balance local detail with contextual information.

As part of preprocessing, we apply non-zero cropping based on the prostate region using the nnUNet pipeline [13]. This removes irrelevant background and focuses computation on the anatomical region of interest, improving both memory efficiency and learning focus. All images are resampled to a voxel spacing of 3.0 mm × 0.5 mm × 0.5 mm (z, y, x) to ensure consistent resolution across subjects and datasets. Intensity normalization is performed using z-score normalization, and data augmentation follows nnUNet's default configuration, including random flipping, elastic deformation, gamma correction, scaling, and rotation. This preprocessing setup ensures spatial consistency, training robustness, and generalizability across diverse clinical data sources.

4.3. Model Development

We adopt the default optimizer, learning rate schedule, weight initialization, and training configuration from nnUNet [13], which provides a robust and generalizable framework for medical image segmentation. Specifically: Optimizer: Stochastic Gradient Descent (SGD), Learning rate: 0.01 with polynomial decay, Momentum: 0.99, Weight decay: 3e-5, Batch size: 1 (constrained by GPU memory), Training epochs: 1000, Loss weighting: nnUNet's foreground-background balancing

We experimented with several loss functions and found that a combination of Dice and Cross-Entropy loss delivered the most stable and accurate performance across validation metrics. While other losses such as Focal, Tversky, and boundary-aware variants were briefly explored, we observed unstable training (e.g., foreground loss collapse) in some configurations. Due to time constraints and lack of consistent improvement, we selected the Dice + Cross-Entropy formulation for final training and evaluation. This setup, paired with nnUNet-style dynamic data augmentation and learning rate scheduling, enabled stable conver-

Mean Values of Dice Score, IoU, Hausdorff Distance, Recall, and Precision (External Validation: Prostate158)					
Model	**Dice (%)**	**IoU (%)**	**HD (mm)**	**Recall**	**Precision**
U-Net[20]	17.26	12.01	24.73	16.54	25.03
V-Net[15]	22.32	12.80	21.56	18.25	39.14
nnUNet[13]	20.02	13.65	23.21	19.21	35.04
Attention U-Net[17]	31.23	22.37	19.83	31.27	41.07
Swin U-Net[4]	19.05	13.54	23.22	20.32	31.10
nnFormer[32]	23.81	16.01	21.03	22.29	34.16
ProLesNet[30]	30.22	21.34	19.89	29.25	42.17
Proposed Method	**34.13**	**23.99**	**17.43**	**32.93**	**43.25**

Table 1. Segmentation performance comparisons of our proposed method and SOTA baselines. All are trained on 219 expert-annotated cases from the T2W PI-CAI Challenge dataset and evaluated on external csPCa 82 patients from the T2W Prostate158 dataset.

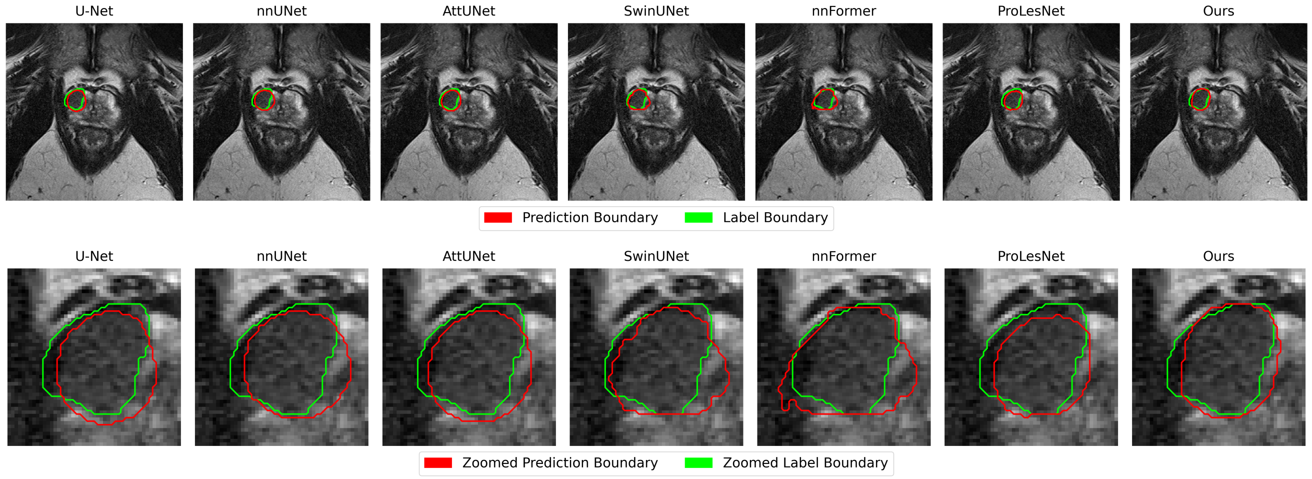

Figure 3. Qualitative comparison of predictions on the external csPCa Prostate158 dataset. The top row shows axial T2-weighted MRI slices overlaid with predicted (red) and ground truth (green) contours across baseline models and our proposed method. The bottom row presents zoomed-in views of the lesion region, emphasizing boundary fidelity. Our method shows noticeably tighter alignment with expert annotations, particularly around irregular and fine-grained lesion edges.

gence and strong generalization across internal and external datasets.

4.4. Qualitative Evaluations and Analysis

We evaluate the performance of our model on the Prostate158 external test set using five standard segmentation metrics: Dice Similarity Coefficient (Dice), Hausdorff Distance (HD), Intersection over Union (IoU), Recall, and Precision. As shown in Table 1, our method is compared against several state-of-the-art baselines, including U-Net [20], nnUNet [13], Attention U-Net [17], V-Net [15], Swin U-Net [4], nnFormer [32], and the recent ProLesNet [30]. To ensure a fair and consistent comparison, all SOTA baselines that have mentioned in Table were trained on the same T2-weighted MRI dataset (219 expert-annotated cases from PI-CAI) and evaluated on the same external test set (82 csPCa cases from Prostate158), using identical data splits and preprocessing.

Our proposed method achieves the highest Dice score (34.13%), surpassing all baselines, including ProLesNet (30.22%), which was previously among the strongest performers in prostate lesion segmentation. Notably, our model also outperforms all others in IoU, HD, and precision, while maintaining high recall. These improvements demonstrate the strength of our architectural contributions—specifically the Multi-scale Edge-Aware Spatial Attention Gate (EASAG) and residual decoder—in capturing fine-grained lesion boundaries and contextual information in accurately segmenting lesions under single-modality constraints. In terms of HD, our method records the lowest distance value, indicating reduced outlier error and better boundary alignment. We also achieve the highest IoU

Models	**Dice (%)**	**HD (mm)**	**IoU (%)**	**Recall (%)**	**Precision (%)**
Spatial Attention (SA) U-Net	31.23	19.83	22.37	31.27	41.07
REB + SA	30.10	19.86	21.37	29.04	**41.59**
RDB + SA + 2 Conv/Encoder Stage	30.58	19.88	21.78	30.22	39.30
RDB + SA + 1 Conv/Encoder Stage	**33.10**	**17.64**	**23.53**	**32.99**	40.35
Residual SA	30.59	18.56	21.92	31.1	37.37

Table 2. Ablation study evaluating the effect of residual connection and encoder depth (1 vs 2 Convolutions per stage) in a spatial attention (SA) U-Net. Evaluated on the External T2W csPCa Prostate158 dataset. REB (Residual Encoder Block), RDB (Residual Decoder Block), Residual SA (when both Encoder & Decoder have residual block). Best performance is achieved using a residual decoder with a shallow encoder (1 conv/ encoder stage)

Models	**Dice (%)**	**HD (mm)**	**IoU (%)**	**Recall (%)**	**Precision (%)**
Multi-Scale Spatial Attention (MSSA) U-Net	29.43	19.55	20.99	27.75	**42.29**
REB + MSSA	28.80	19.98	20.14	28.20	38.91
RDB + MSSA + 2 Conv/Encoder Stage	30.42	19.07	21.75	30.04	39.83
RDB + MSSA + 1 Conv/Encoder Stage	**33.71**	18.26	**24.00**	33.61	41.95
Residual MSSA	32.5	18.79	23.18	**34.26**	39.71

Table 3. Impact of Multi-Scale Spatial Attention (MSSA) design. Just replace Single scale spatial attention (SA) multi-scale spatial attention else everything same as Table 2.

and Precision, demonstrating both spatial accuracy and high confidence in positive lesion prediction. Recall is also competitive, further supporting the model's balanced detection capability.

The results confirm that even within a T2W-only setting, our design yields superior segmentation accuracy and boundary consistency compared to recent transformer-based, attention-based, and residual network baselines. This validates the effectiveness of our model under realistic clinical input constraints. These results demonstrate that our model generalizes well to unseen clinical data and offers a viable alternative to complex multi-modal setups — a critical advantage in practical deployment scenarios.

4.5. Qualitative Analysis

To qualitatively evaluate segmentation performance, we present both full-slice and zoomed-in contour visualizations of one of the best-case prediction from the external csPCa Prostate158 test set (Figure 3). The top row shows the complete axial T2-weighted slice overlaid with predicted (red) and ground truth (green) contours for U-Net, nnUNet, AttUNet, SwinUNet, nnFormer, ProLesNet, and our proposed method. The bottom row provides a zoomed-in view focusing on the lesion boundary to better highlight segmentation differences.

In this representative case, our model achieves tighter and more consistent alignment with the expert-annotated ground truth compared to competing methods. Notably, baseline models tend to either under-segment or deviate around lesion edges, while our edge-aware attention and residual decoding structure more faithfully preserve boundary shape and extent. This visualization reinforces the advantage of our architectural design in handling fine lesion details and improving spatial accuracy in prostate MRI segmentation.

4.6. Ablation study

Effect of Residual Decoder and Encoder Depth: We evaluate how residual connections and encoder convolutional depth affect performance in a spatial attention (SA) U-Net backbone. The baseline SA U-Net yields 31.23% Dice and 19.83 mm HD. Adding residual blocks in the encoder (REB + SA) improves precision (41.59%) but degrades Dice and recall, suggesting over-smoothing of low-level features.

Introducing residual decoding with two convolutions per encoder stage (RDB + SA + 2 Conv) gives marginal gains but still underperforms. Reducing encoder depth to one convolution (RDB + SA + 1 Conv) achieves the best Dice (33.10%), HD (17.64 mm), and recall (32.99%), highlighting the benefit of a shallower encoder for small lesion segmentation. A fully residual model (Residual SA) shows no additional gain, reaffirming that residual encoders may hinder fine-grained localization in boundary-sensitive tasks.

In summary, the combination of a residual decoder and a lightweight encoder with only one convolution per stage offers the best trade-off between contextual representation and boundary precision for T2W-based prostate lesion segmentation.

Effect of Multi-Scale Spatial Attention and Residual Decoding: We analyze the effect of replacing the single-scale spatial attention (SA) with a Multi-Scale Spatial Attention (MSSA) module that captures features at multiple

Models	Dice (%)	HD (mm)	IoU (%)	Recall (%)	Precision (%)
SE + SA U-Net	30.22	19.89	21.34	29.25	42.17
RDB + SE + SA + 2 Conv/Encoder Stage	30.92	19.29	21.40	29.76	**43.37**
RDB + SE + SA + 1 Conv/Encoder Stage	29.16	20.74	20.46	29.18	40.85

Table 4. Effect of combining channel attention i.e squeeze excitation (SE) with spatial attention (SA) in a residual decoder block (RDB) U-Net on Multi-scale basis. Evaluation on the External T2W csPCa Prostate158 dataset.

Models	Dice (%)	HD (mm)	IoU (%)	Recall (%)	Precision (%)
Residual Multi-Scale Spatial + EASAG (Ours)	**34.13**	**17.43**	**23.99**	**32.93**	43.25
Residual Single-Scale Spatial + EASAG	32.99	17.87	23.38	32.79	**44.10**

Table 5. Ablation of spatial attention design (Multi-Scale vs. Single-Scale) under a residual decoder block (RDB) and 1 Conv per Encoder stage with EASAG. Evaluation on the external T2W csPCa Prostate158 dataset.

resolutions. Similar to the SA setup, we vary encoder depth and residual configuration in MSSA. Results in Table 3 show that the best configuration remains a shallow encoder (1 Conv per stage) with residual decoding (RDB), achieving 33.71% Dice and 18.26 mm HD and 24% IoU among it's model variants in it's own table. Compared to the SA version, this setting improves Dice by +0.61%, IoU by +0.47%, and precision by +1.6%. Fully residual variants or deeper encoders again perform worse, supporting the conclusion that shallow encoders, residual decoder and multi-scale context yield more accurate and boundary-aware segmentation.

Effect of Channel Attention (SE) Combined with Spatial Attention: In Table 4, we explore the addition of channel attention (via Squeeze-and-Excitation) alongside spatial attention in a residual decoder framework. This combination aims to enrich feature representations by focusing both across spatial dimensions and feature channels.

Surprisingly, introducing SE attention does not yield performance improvements over spatial attention alone. The base SE-Spatial model with residual decoding and 2 Conv per encoder stage achieved modest Dice (30.92%) and IoU (21.40%), with a notable boost in precision (43.37%) — the best among all models. This suggests SE attention may help in suppressing false positives.

However, reducing encoder complexity to 1 Conv per stage resulted in a drop across most metrics, with Dice decreasing to 29.16% and HD worsening to 20.74 mm. These results indicate that SE-based channel attention did not synergize well with our residual decoding and shallow encoder pipeline, likely due to over-regularization or redundancy in feature reweighting.

Effect of Edge-Aware Spatial Attention Gate (EASAG): Table 5 compares the impact of multi-scale versus single-scale spatial attention in the presence of a residual decoder and the proposed Edge-Aware Spatial Attention Gate (EASAG). To ensure fair comparison, both configurations employ a single 3D convolution per encoder stage.

Our full model — integrating multi-scale spatial attention with EASAG — delivers the best overall performance, achieving a Dice score of 34.13%, HD of 17.43 mm, and an IoU of 23.99%. In contrast, its single-scale counterpart (also with EASAG) yields a Dice score of 32.99% and IoU of 23.38%, indicating reduced segmentation fidelity.

The only metric where the single-scale version slightly outperforms is precision (44.10% vs. 43.25%), likely due to its more conservative boundary estimates. However, the multi-scale model demonstrates better recall (32.93% vs. 32.79%) and significantly improved boundary alignment, highlighting its capacity to capture difficult or ambiguous lesions more effectively.

5. Conclusion

We presented a novel T2W-only segmentation architecture tailored for prostate lesion delineation under real-world clinical constraints. By integrating a lightweight residual decoder with our proposed Multi-scale Edge-Aware Spatial Attention Gate (EASAG), the model improves both contextual understanding and boundary localization without relying on multi-modal MRI inputs.

Extensive evaluations on the external Prostate158 dataset show that our method achieves state-of-the-art performance in the single-modality setting. It remains computationally efficient, generalizes well, and aligns with deployment-ready requirements.

Future directions include extending to multi-modal setups and incorporating uncertainty-aware or boundary-guided training strategies to further enhance segmentation quality.

References

[1] Asim Ahmad et al. Artificial intelligence in prostate cancer: Current state and future perspectives. *Cancers*, 15(5):1350, 2023.

[2] American College of Radiology. Pi-rads prostate imaging – reporting and data system, version 2.1. https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/PI-RADS, 2019. Accessed: July 11, 2025.

[3] Edward J. Bass, Aleksandra Pantovic, Matthew Connor, Rebecca Gabe, Anwar R. Padhani, Andrea Rockall, Harbir Sokhi, Hwei Tam, Mark Winkler, and Hashim U. Ahmed. A systematic review and meta-analysis of the diagnostic accuracy of biparametric prostate mri for prostate cancer in men at risk. *Prostate Cancer and Prostatic Diseases*, 24(3):596–611, 2021.

[4] Jingyun Cao et al. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

[5] Ruidong Cao, Xiaoyuan Zhong, Soroosh Shakeri, Ali M. Bajgiran, Soroush A. Mirak, Dieter Enzmann, Steven S. Raman, and Kevin Sung. Prostate cancer detection and segmentation in multi-parametric mri via cnn and conditional random field. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1900–1904. IEEE, 2019.

[6] DIAG Research Group, Radboud University. Prostate158: Multi-center 3d mri dataset for prostate lesion segmentation. https://github.com/DIAGNijmegen/prostate158, 2022. Accessed July 2025.

[7] Zachary A. Eidex, Tonghe Wang, Yao Lei, Mircea Axente, Olabimpe O. Akin-Akintayo, Oluwaseun A. A. Ojo, Ayooluwa A. Akintayo, Justin Roper, Jeffrey D. Bradley, Tian Liu, et al. Mri-based prostate and dominant lesion segmentation using cascaded scoring convolutional neural network. *Medical Physics*, 49(8):5216–5224, 2022.

[8] Sara B. Ginsburg, Baris Turkbey, Peter L. Choyke, et al. Radiomic features for prostate cancer detection on mri: A prospective study. *Radiology*, 289(1):110–118, 2017.

[9] Donald F Gleason and George T Mellinger. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. *The Journal of Urology*, 111(1):58–64, 1974.

[10] Darshan D. Gunashekar, Lukas Bielak, Andy L. Ha, Thomas Brox, Christos Zamboglou, and Matthias Bock. Explainable ai for cnn-based prostate tumor segmentation in multi-parametric mri correlated to whole mount histopathology. *Radiation Oncology*, 17(1):65, 2022.

[11] Prasad Hambarde, Sanjay Talbar, Abhijit Mahajan, Sunil Chavan, Mukund Thakur, and Nilesh Sable. Prostate lesion segmentation in mr images using radiomics based deeply supervised u-net. *Biocybernetics and Biomedical Engineering*, 40(4):1421–1435, 2020.

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[13] Fabian Isensee, Paul F. Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18:203–211, 2021.

[14] Geert Litjens et al. Computer-aided detection of prostate cancer in mri. *IEEE Transactions on Medical Imaging*, 33(5):1083–1092, 2014.

[15] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *3DV*, 2016.

[16] Virginia A Moyer. Screening for prostate cancer: Us preventive services task force recommendation statement. *Annals of Internal Medicine*, 157(2):120–134, 2012.

[17] Ozan Oktay et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[18] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R. Zaiane, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7479–7489, 2019.

[19] Xinyu Ren, Yifeng Zhang, Hongen Zhang, Xiaomeng Xu, Yinghuan Hu, Liang Sun, Shaoting Zhang, and Yefeng Zheng. Interleaved 3d convolutional neural networks for joint segmentation of small-volume structures in head and neck ct images. *Medical Image Analysis*, 68:101852, 2021.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *MICCAI*, 2015.

[21] Holger R. Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B. Turkbey, and Ronald M. Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 556–564, 2015.

[22] Jo Schlemper, Ozan Oktay, et al. Attention gated networks: Learning to leverage salient regions in medical imaging. *Medical Image Analysis*, 53:197–207, 2019.

[23] David A. Siegel, Mary E. O'Neil, Thomas B. Richards, Nancy F. Dowling, and Hannah K. Weir. Prostate cancer incidence and survival, by stage and race/ethnicity — united states, 2001–2017. *MMWR Morbidity and Mortality Weekly Report*, 69(39):1473–1480, 2020.

[24] Enliang Song, Jie Long, Guoqiang Ma, Hong Liu, Chih-Cheng Hung, Ruoqi Jin, Peng Wang, and Wei Wang. Prostate lesion segmentation based on a 3d end-to-end convolution neural network with deep multi-scale attention. *Magnetic Resonance Imaging*, 99:98–109, 2023.

[25] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71 (3):209–249, 2021.

[26] Baris Turkbey, Peter A. Pinto, et al. Prostate cancer: Value of multiparametric mr imaging at 3 t for detection–histopathologic correlation. *Radiology*, 255(1):89–99, 2010.

[27] H. B. Vos, Y. Sushentsev, S. Winkel, G. Litjens, and H. Huisman. Pi-cai: The prostate imaging—cancer ai grand challenge. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2023. https://picai.org.

[28] Jeffrey C Weinreb, Jelle O Barentsz, Peter L Choyke, François Cornud, Masoom A Haider, Katarzyna J Macura, Daniel Margolis, Mitchell D Schnall, Faina Shtern, Clare M Tempany, Harriet C Thoeny, and Sadhna Verma. Pi-rads prostate imaging—reporting and data system: 2015, version 2. *European Urology*, 69(1):16–40, 2016.

[29] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[30] Yulun Zhang et al. Prolesnet: Prostate lesion segmentation from multi-parametric mri via channel-spatial attention and localization refinement. *IEEE Transactions on Medical Imaging*, 41(10):2813–2826, 2022.

[31] Shuhan Zhao, Xi Li, Huchuan Lu, Xuan Wang, and Liang Wang. Egnet: Edge guidance network for salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8772–8781, 2019.

[32] Zongwei Zhou et al. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021.