big data and cognitive computing

MDPI

Article

MDA-Net: A Segmentation Network for Kidney Tumor Based on Enhanced Multi-Scale Feature Extraction and Attention Refinement

Shaofu Lin [1], Yumiao Chang [1], Jianhui Chen [2,3,4,5,*] and Lianfang Ma [6]

[1] College of Computer Science, Beijing University of Technology, Beijing 100124, China; linshaofu@bjut.edu.cn (S.L.); changyumiao@emails.bjut.edu.cn (Y.C.)
[2] School of Artificial Intelligence, Beijing University of Technology, Beijing 100124, China
[3] Engineering Research Center of Digital Community, Ministry of Education, Beijing 100124, China
[4] Beijing Key Laboratory of MRI and Brain Informatics, Beijing University of Technology, Beijing 100124, China
[5] Engineering Research Center of Intelligent Perception and Autonomous Control, Beijing 100124, China
[6] School of Computer Science and Technology, Shandong Technology and Business University, Yantai 264005, China; malianfang@sdtbu.edu.cn
* Correspondence: chenjianhui@bjut.edu.cn

Abstract

Accurate kidney tumor segmentation from abdominal CT is essential for quantitative assessment and treatment planning. However, indistinct tumor boundaries and substantial inter-patient shape variability render traditional hand-crafted feature-based methods unreliable for precise delineation. Although deep learning has advanced this task, these methods still struggle with multi-scale tumor characteristics, complex morphological variations, and background noise in medical images. To address these challenges, we propose MDA-Net, an end-to-end segmentation method based on enhanced multi-scale feature extraction and attention refinement. Specifically, we introduce a Multi-Scale Feature Extraction (MSFE) module into encoder–decoder skip connections to aggregate dilated features across multiple receptive fields and learn branch-wise weights for adaptive refinement and fusion, thereby enhancing boundary details and semantic cues to reduce tumor-tissue ambiguity. At the bottleneck, a Deformable Pyramid Feature Refinement (DPFR) module combines deformable sampling with pyramid contextual modeling, thereby improving adaptability to variations in tumor shape and scale while preserving feature resolution. Moreover, a Channel and Spatial Attention (CASA) module is embedded in the decoder to suppress background interference and enhance boundary-sensitive structures during upsampling via coordinated channel and spatial reweighting, thereby improving the reconstruction of fine-grained tumor morphology and contours. Experiments on both KiTS19 and KiTS21 show that MDA-Net consistently improves tumor boundary delineation, lesion localization, and mask reconstruction, demonstrating stronger robustness and cross-dataset generalizability than representative baseline methods. Ablation studies further confirm the complementary effects of MSFE, DPFR, and CASA. In addition, Grad-CAM visualizations improve the clinical transparency and interpretability of the model. Overall, this method advances deep learning for medical image analysis and supports precise diagnosis and treatment of renal tumors.

Keywords: kidney tumor; multi-scale feature enhancement; TransUNet; medical image segmentation; attention mechanism

Check for updates

Academic Editor: Guangyao Li

Received: 22 March 2026
Revised: 4 May 2026
Accepted: 6 May 2026
Published: 8 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The kidney is a vital organ that eliminates metabolic waste, regulates electrolyte balance, and maintains blood pressure homeostasis [1]. Kidney cancer is one of the most common malignancies worldwide, accounting for approximately 2.4% of all cancer cases, with more than 330,000 new cases and over 170,000 deaths reported annually [2]. It is a major cause of renal dysfunction and imposes a substantial burden through increased morbidity, long-term treatment costs, and excess mortality [3]. Early-stage kidney cancer often presents without specific symptoms or reliable serological biomarkers. As a result, many lesions are detected incidentally during imaging examinations performed for unrelated conditions, which frequently delays diagnosis [4]. Therefore, rapid and accurate segmentation of kidney tumors, together with precise delineation of tumor boundaries, is essential for early detection and individualized treatment planning, thereby helping to limit disease progression and reduce unnecessary radical nephrectomy.

In recent years, AI-driven medical imaging has increasingly supported key clinical tasks, including disease diagnosis, prognosis assessment, and clinical decision-making. In particular, deep learning-based AI methods have made considerable progress in medical image segmentation and have been widely applied to the automated delineation and quantitative analysis of kidney and renal tumor regions [5,6]. Through end-to-end optimization, modern neural networks are able to learn discriminative representations directly from large-scale annotated datasets, integrate semantic and structural information across multiple scales, and progressively refine model fitting through data-driven parameter updates [7]. Consequently, these methods not only improve segmentation accuracy and reduce manual workload, but also simplify the clinical imaging workflow, thereby enhancing both efficiency and inter-observer consistency.

Despite these advances, the automatic segmentation of kidney tumors remains challenging. First, as shown in Figure 1, substantial inter-patient variability leads to pronounced heterogeneity in tumor size, location, and imaging appearance, making it difficult to learn discriminative features that are both transferable and generalizable. Second, the interface between renal tumors and normal parenchyma often exhibits low contrast or a gradual transition, which hinders precise boundary delineation [8]. Finally, abdominal CT images present complex, scale-dependent variations in morphology and intensity, further increasing the uncertainty of lesion localization.

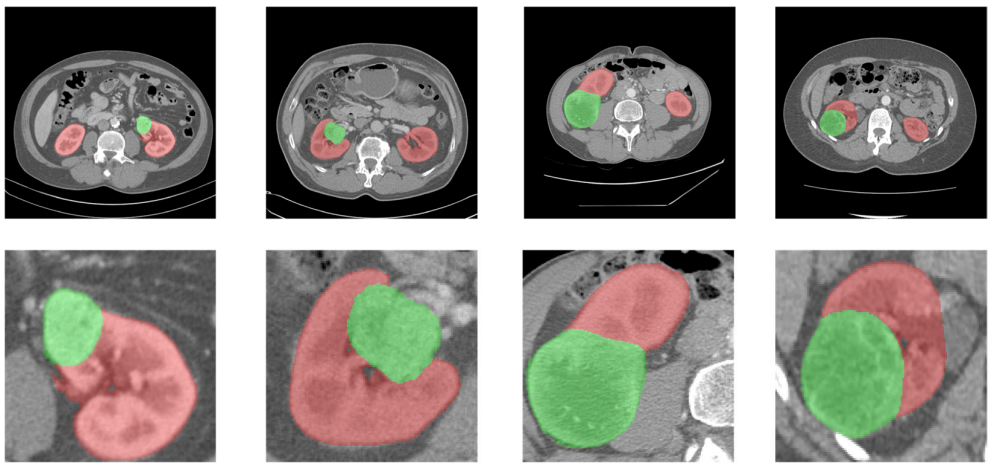


Figure 1. CT examples of renal lesions across different patients with blurred tumor boundaries (red denotes the kidney region, and green denotes the tumor region).

Given the challenges discussed above, there is a pressing need for more accurate algorithms to achieve fine-grained segmentation of renal tumors. With the continued devel-

opment of deep learning in medical image analysis, CNN-based models have become the dominant framework for automatic medical image segmentation, and early representative architectures, such as fully convolutional networks (FCNs) [9], have substantially advanced the field. However, when fine spatial details are progressively weakened during the encoding of features into high-level semantic representations, the subsequent upsampling process in FCN-style architectures often lacks sufficient localization information, resulting in blurred predictions and poorly defined tumor boundaries.

U-Net [10] mitigates the loss of spatial details in FCN-style architectures by introducing skip connections that fuse high-level semantic information with low-level spatial cues, and it has become a foundational architecture for medical image segmentation. Building on this framework, subsequent variants of U-Net have mainly improved performance through multi-scale context aggregation [11], residual feature learning, attention-based and boundary-aware refinement [12], and stronger supervision, such as hybrid loss functions or multi-task objectives [13]. These strategies enhance the integration of semantic and structural information and generally produce more accurate predictions in complex scenarios. However, they also suggest that accurate medical image segmentation increasingly relies on the coordinated enhancement of context modeling, feature refinement, and boundary recovery, rather than on simple architectural scaling.

However, CNN-based models are inherently constrained in their ability to capture global contextual information because convolutional receptive fields are local by design and are largely determined by factors such as kernel size. In complex kidney tumor segmentation tasks, these models often fail to fully exploit global semantic and anatomical context, which limits the stability of lesion localization and the precision of boundary delineation.

To address this limitation, researchers have introduced Transformer-based methods [14], Vision Transformers [15], and hybrid CNN-Transformer paradigms [16] that aim to combine the strengths of CNNs in local feature extraction and multi-scale detail modeling with the ability of Vision Transformers to capture global dependencies, model long-range contextual interactions, and learn dynamic feature relationships. A representative example is TransUNet [17], which adopts a hybrid CNN-Transformer encoder and uses self-attention-based global context modeling to compensate for the limited global representation capacity of CNNs. Swin-Unet [18] further incorporates hierarchical Transformer modules into a U-Net-like encoder–decoder architecture, thereby enabling effective medical image segmentation. Shen et al. [19] developed COTRNet, a hybrid segmentation framework in which convolutional blocks and Transformer blocks are alternately arranged to form a fused architecture; in addition, class imbalance is addressed through a composite objective that combines class-weighted cross-entropy loss with weighted Dice loss. Hou et al. [20] proposed an end-to-end network termed DCCTNet, which employs ResNet34 as the primary branch and Swin Transformer as the auxiliary branch for hybrid feature extraction; moreover, a multi-scale feature pyramid is incorporated into the decoder to strengthen contextual modeling, thereby improving the recognition of kidney and renal tumors. Hu et al. [21] introduced STC-UNet, which integrates selective kernel mechanisms with Vision Transformers and adopts a hierarchical strategy to promote the coordinated enhancement of local and global information, ultimately improving semantic feature extraction for the segmentation of the kidney and renal tumors. Yuan et al. [22] constructed a hybrid encoder based on parallel CNN and Swin Transformer streams, resulting in a complementary CNN-Transformer network; furthermore, channel attention is incorporated into the self-attention component of the Transformer to capture dual attention cues.

Recent studies on related tumor segmentation tasks further indicate that robust tumor delineation often benefits from multi-modal information fusion, attention-guided feature

refinement, and ensemble-based decision strategies. Mazher et al. [23] combine deep learning with radiomics by extracting features from multimodal MRI-derived segmentation results and integrating them with clinical information for survival prediction in patients with brain tumors. Akter et al. [24] develop a CNN- and U-Net-based framework in which segmentation information is used to improve brain tumor classification, highlighting the complementary relationship between segmentation and diagnostic prediction. More recently, Rasool et al. [25] propose ResMHA-Net, a glioma segmentation framework that combines ResNet and U-Net with multi-head attention and uses multiple MRI modalities, including T1, T1ce, T2, and FLAIR, for tumor delineation. They further integrate radiomic features extracted from predicted masks with clinical variables and build a voting-based ensemble model for survival prediction. These studies suggest that, although their imaging modalities and anatomical targets differ from those of renal tumor segmentation, multi-modal fusion, attention-enhanced feature representation, and ensemble-based prediction are important methodological directions for tumor segmentation and clinically oriented medical image analysis.

Although existing studies have achieved notable progress in the segmentation of the kidney and renal tumors, substantial challenges remain. Inter-patient variability in renal anatomy and tumor presentation often leads to ambiguous boundary delineation and complicates the reliable separation of the kidney from adjacent organs and tissues. Moreover, the complex multi-scale characteristics of tumors in medical images, together with image noise and interference from surrounding tissues, often prevent current methods from extracting truly discriminative lesion features. Consequently, accurate segmentation of renal tumors remains a critical challenge in medical image segmentation, which continues to motivate the development of more precise and robust approaches.

This study aims to develop an end-to-end segmentation method for more accurate renal tumor segmentation based on enhanced multi-scale feature extraction and attention refinement, referred to as MDA-Net. Inspired by TransUNet [17], the proposed network adopts a hybrid CNN–Transformer backbone. The main contributions of this study are summarized as follows:

- This study proposes a novel U-shaped segmentation network for accurate kidney tumor segmentation. To address the limitation of conventional skip connections that rely on simple feature concatenation, the network embeds a customized Multi-Scale Feature Extraction Block (MSFE), which enriches decoder representations with multi-scale detail and texture cues and strengthens multi-scale texture modeling.
- This study designs a Deformable Pyramid Feature Refinement (DPFR) Block at the bottleneck, which serially combines deformable convolution for adaptive sampling with an ASPP module, thereby improving the modeling of complex tumor morphologies and enhancing robustness to variations in tumor shape, location, and scale.
- This study introduces a Channel and Spatial Attention (CASA) Block into the decoder and applies coordinated channel and spatial attention to refine features during upsampling reconstruction. The CASA module suppresses background noise and amplifies tumor-relevant responses, thereby improving tumor contour recovery and localization accuracy.

The remainder of this paper is organized as follows. Section 2 describes the dataset and the proposed methodology. Section 3 presents the datasets and evaluation metrics used to assess model performance and reports the experimental results. Section 4 discusses the limitations of the proposed model and outlines potential directions for future improvement. Finally, Section 5 concludes the paper by summarizing the present study.

2. Materials and Methods

2.1. Dataset Description

This study uses two public datasets, KiTS19 [26] and KiTS21 [27], for model training and evaluation. The first dataset is KiTS19, which contains abdominal CT scans from 210 patients and provides expert-annotated segmentation labels for the kidney and tumor regions. All images and annotations are provided in NIfTI format. The second dataset is KiTS21, which contains 300 contrast-enhanced abdominal CT scans from patients who underwent partial or radical nephrectomy for suspected kidney malignancy. The definitions of the kidney and tumor regions in KiTS21 are consistent with those in KiTS19.

Raw CT volumes were preprocessed before model training. Following commonly used window-width and window-level settings for abdominal CT images, HU values were clipped to $[-74,\ 300]$ to increase the contrast between renal tumors and surrounding tissues. The clipped images were then normalized to $[0, 1]$ to ensure input consistency and improve the stability of model training [28]. For each patient, axial slices were extracted as the model input [29]. Each input image was resized to 512×512. Slices without corresponding mask annotations were excluded, and the remaining slices are retained as valid inputs to the network. Before training, data augmentation techniques, including random rotation, horizontal flipping, and vertical flipping, are applied to improve generalization and reduce the risk of overfitting.

The KiTS19 dataset is randomly split at the patient level into the training set, validation set, and independent test set at a ratio of 6:2:2, which include 126, 42, and 42 patients, respectively. The test set is reserved exclusively for final evaluation, while the training and validation sets are used for optimal model selection. The preprocessing procedure for the KiTS21 dataset was consistent with that used for KiTS19, and KiTS21 was employed as an independent dataset to evaluate model generalization.

2.2. Overall Architecture

Figure 2 illustrates the overall architecture of MDA-Net, which comprises a hybrid encoder, skip connections, a bottleneck stage, and a decoder. Unlike conventional variants of U-Net, MDA-Net integrates CNNs and Transformers within the encoder, where the CNN branch extracts fine-grained local texture and structural information, while the Transformer branch uses self-attention to model long-range dependencies and capture global contextual information. The MDA-Net adopts ResNet-50 as the CNN backbone and further incorporates a Vision Transformer encoder based on CNN features to extract global semantic information. The Vision Transformer encoder follows the 12-layer Transformer configuration used in TransUNet, with a patch size of 16×16, 12 attention heads, and a hidden feature dimension of 768.

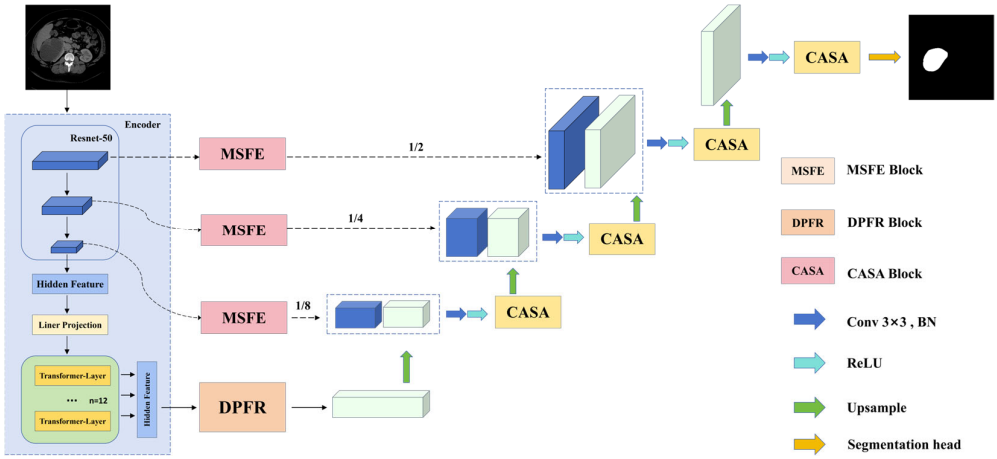


Figure 2. Overall architecture of the proposed MDA-Net.

MDA-Net introduces several targeted architectural refinements. First, a DPFR Block is inserted at the bottleneck between the Vision Transformer encoder and the decoder, which substantially enlarges the effective receptive field without reducing feature resolution, thereby enhancing global discriminability and regional consistency. Second, in the skip connections, the proposed MSFE Block is incorporated to extract and fuse multi-scale features from shallow CNN stages, providing the decoder with rich multi-scale detail and texture information. Finally, the decoder extends the conventional stage-wise U-shaped decoding pathway by embedding CASA Blocks throughout the decoding process, where coordinated channel attention and spatial attention improve detail recovery and semantic representation during resolution reconstruction.

2.3. Multi-Scale Feature Extraction (MSFE) Block

Although skip connections in conventional U-shaped architectures help preserve shallow spatial details, they often lack explicit multi-scale refinement and therefore cannot fully exploit such detailed information. This limitation may prevent the decoder from receiving the spatial and semantic cues required for accurate resolution recovery. To address this limitation, we design a Multi-Scale Feature Extraction Block and embed it into the skip connections between the shallow encoder stages and the decoder, thereby enabling more effective multi-scale feature extraction and providing the decoder with features enriched by edge information and texture details. The architecture of MSFE is illustrated in Figure 3.

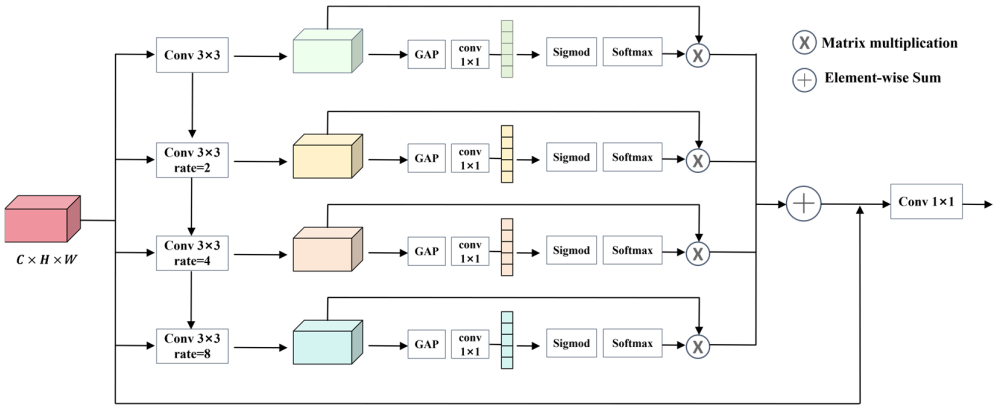


Figure 3. Architecture of the MSFE Block.

Given an input feature map $X \in R^{C \times H \times W}$, where C denotes the number of channels and H and W represent the height and width, respectively, MSFE employs four dilated

convolutions with distinct dilation rates to capture multi-scale features Y_i, thereby obtaining representations under different receptive-field sizes:

$$Y_i = \begin{cases} Cov3 \times 3_{rate=1}(X) & i = 1 \\ Cov3 \times 3_{rate=2(i-1)}(X + Y_{i-1}) & 1 < i <= n \end{cases} \quad (1)$$

where n denotes the number of dilated convolution branches and is set to 4 in this study. The dilation rates of the four branches are set to 1, 2, 4, and 8, respectively, to progressively enlarge the receptive field while preserving fine-grained spatial details. For a 3×3 dilated convolution with dilation rate r, the effective kernel size is given by $2r + 1$. Accordingly, the adopted dilation rates correspond to effective receptive fields of 5×5, 9×9, and 17×17. This configuration enables the MSFE block to hierarchically capture local boundary details, neighborhood-level structural continuity, medium-scale tumor morphology, and broader contextual cues. Since MSFE is embedded in the skip connections between the shallow encoder stages and the decoder, excessively large dilation rates may introduce sparse sampling patterns and weaken boundary continuity, particularly for small or low-contrast tumor regions.

After feature extraction from the four branches, the resulting multi-scale features are each passed through global average pooling followed by a 1×1 convolution to generate channel-wise attention weights. Next, a sigmoid activation is applied to map these weights into the range of $[0, 1]$, and a softmax operation is subsequently used to normalize the weights across branches, thereby emphasizing informative signals. Finally, for each branch, the normalized weight is multiplied by the corresponding multi-scale feature map to obtain the reweighted feature representation. The output feature of the i-th branch, denoted by Z_i, is given as follows:

$$\mathrm{Z_i} = \mathrm{Y_i} \otimes \mathrm{Softmax}(\mathrm{Sigmoid}(\mathrm{GAP}(\mathrm{Y_i}))), (i \leq n) \quad (2)$$

where $\otimes$ denotes element-wise multiplication. Subsequently, the outputs from all branches are fused through element-wise addition. A residual connection is then introduced by adding the original input feature map to the aggregated reweighted features, followed by an additional 1×1 convolution to obtain the final output feature map. By leveraging dilated convolutions with different dilation rates, MSFE captures tumor-related cues at multiple scales and effectively integrates multi-scale representations, thereby enriching spatial information and enhancing semantic expressiveness.

2.4. Deformable Pyramid Feature Refinement (DPFR) Block

Conventional convolutions often perform effectively when segmenting objects with regular shapes because they aggregate features within fixed local windows determined by the kernel size. However, their effectiveness may decline in the segmentation of complex renal tumors, as abdominal CT images exhibit substantial variability in tumor morphology, size, and location. Therefore, to enhance the ability of the model to capture tumors with complex shapes, we introduce the proposed DPFR Block, as illustrated in Figure 4.

The encoder output feature map produced by the proposed module has dimensions of $C \times H \times W$, where C denotes the number of high-dimensional feature channels, and H and W represent the spatial resolution. This feature map is first processed by a deformable convolution with a 3×3 kernel and is then fed into several parallel branches for multi-scale feature extraction, ultimately producing a refined feature map with the same spatial dimensions as the input.

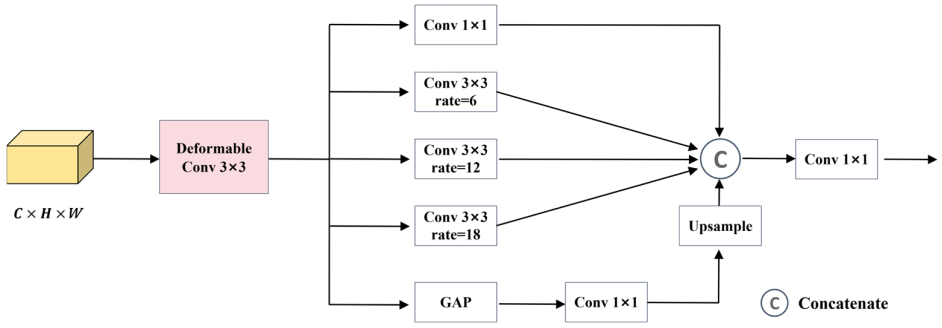


Figure 4. Structure of the proposed DPFR block.

Specifically, the parallel branches include three dilated convolution branches with dilation rates of 6, 12, and 18, a 1×1 convolution branch, and a global average pooling branch followed by upsampling. Together, these branches provide receptive fields of different sizes to capture contextual information at multiple scales. Deformable convolution (DC) introduces additional offsets into the convolution operation to adaptively adjust the sampling locations of standard convolution, thereby expanding the sampling coverage and receptive field in the spatial domain. In this study, these offsets are learned from the kidney tumor segmentation task, which enables the module to better account for variations in tumor shape.

2.5. Channel and Spatial Attention (CASA) Block

In prior studies, the decoders of U-shaped networks commonly use transposed convolution or bilinear interpolation for resolution reconstruction. Although computationally efficient, these operations may cause a mismatch between semantic information and fine-grained details, leading to over-smoothed boundaries, loss of fine structures, and background interference from noisy skip connection features. These limitations can introduce redundant channel responses during decoding and weaken the localization of low-contrast tumors and small lesions.

Channel and spatial attention mechanisms provide an effective way to address these issues by selectively refining features from complementary dimensions. Channel attention emphasizes task-relevant semantic responses and suppresses redundant feature channels, whereas spatial attention highlights lesion-related regions and reduces interference from irrelevant background areas. This refinement is particularly suitable for the decoder, where spatial resolution is progressively restored and pixel-level segmentation masks are reconstructed by integrating high-level semantic features with shallow spatial details. Therefore, attention-guided feature selection during decoding can directly support boundary recovery and tumor localization. The CASA Block [30] has shown effectiveness in feature refinement in previous segmentation studies; we embed it into the stage-wise reconstruction process of the decoder to enhance tumor-relevant responses while suppressing background interference. The CASA Block combines parallel channel attention and spatial attention to capture contextual features in a more comprehensive manner. Through the coordinated interaction of these two attention mechanisms, CASA enables more precise reconstruction of fine-grained tumor morphology and boundary structures while preserving global semantic consistency, thereby reducing the risk of missed detections and boundary deviations and improving robustness in the presence of small tumors and low-contrast conditions. The architecture of CASA is illustrated in Figure 5.

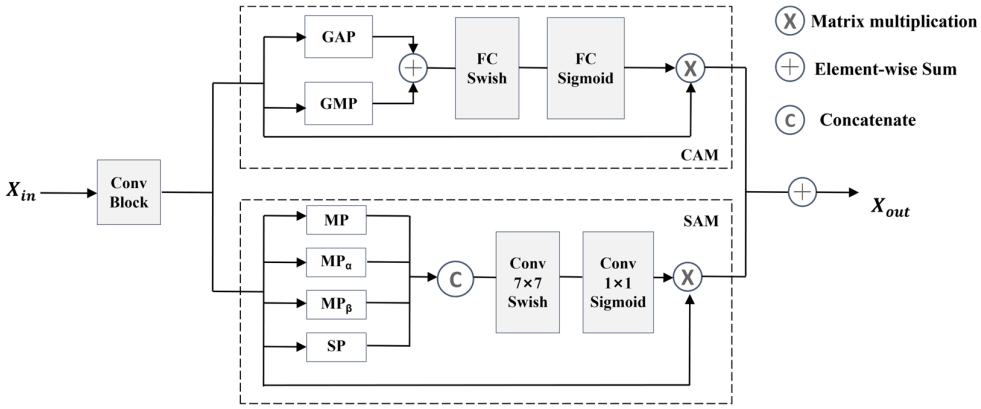


Figure 5. Architecture of the CASA Block.

The CASA module consists of a feature refinement unit, a channel-attention module, and a spatial-attention module. Given an input feature map X_{in}, it is first processed by two serial convolutional layers with kernel sizes of 3×3 and 1×1, respectively, each followed by batch normalization and a ReLU activation. The resulting feature map is then fed in parallel into the channel-attention branch and the spatial-attention branch.

In the channel-attention module, global average pooling and global max pooling are first applied to capture salient details and global contextual information. The resulting pooled descriptors are then combined and passed through two fully connected layers arranged in series, with the Swish activation function [31] used to introduce nonlinearity and preserve smooth gradients. The reduction ratio is set to 16. Specifically, a sigmoid function is applied after the second fully connected layer to generate attention weights, thereby emphasizing the channels that are most relevant to the target task. The output of CAM can be expressed as follows:

$$\mathrm{CAM}(\mathrm{x}) = \mathrm{x} \otimes \sigma(\mathrm{FC}(\mathrm{Swish}(\mathrm{FC}(\mathrm{GAP}(\mathrm{x}) + \mathrm{GMP}(\mathrm{x}))))) \quad (3)$$

where σ denotes the sigmoid activation function, $\otimes$ denotes matrix multiplication, and x represents the feature map refined by the feature refinement module.

In the spatial-attention module, the input feature map is processed by average pooling (MP), max pooling (MP_α), min pooling (MP_β), and sum pooling (SP). The resulting descriptors are then concatenated and fed into a 7×7 convolution layer for further feature refinement. Finally, Swish and sigmoid activations are applied to emphasize informative spatial patterns and generate attention weights. The output of SAM can be expressed as follows:

$$SAM(x, x_{pool}) = x \otimes \sigma(Conv_{1\times1}(Swish(Conv_{7\times7}(x_{pool})))). \quad (4)$$

Here, x and x_{pool} denote the original feature map fed into SAM and the pooled feature descriptor obtained after concatenation, respectively. $\mathrm{Conv}_{1\times1}$ and $\mathrm{Conv}_{7\times7}$ denote the 1×1 and 7×7 convolution operations, respectively. The final output of the CASA module is formulated as

$$X_{out} = CAM(ConvBlock(X_{in})) + SAM(ConvBlock(X_{in})) \quad (5)$$

where ConvBlock denotes a feature refinement unit composed of a convolutional layer, batch normalization, and a ReLU activation function.

2.6. Loss Function

This study adopts a hybrid loss function that combines binary cross-entropy loss with Dice loss [32–34]. Class imbalance is common in medical imaging, and reliance on binary

cross-entropy loss alone may bias the training process toward the majority class, causing minority regions to be overlooked. In contrast, the Dice similarity coefficient effectively alleviates this issue and can also help reduce overfitting. The Dice loss is defined as follows:

$$\mathcal{L}_{Dice} = 1 - \frac{2|A \cap B|}{|A| + |B|} \tag{6}$$

where $|A \cap B|$ denotes the number of overlapping foreground pixels between the predicted segmentation map and the ground-truth label, and $|A|$ and $|B|$ denote the numbers of foreground pixels in the prediction and the ground truth (GT), respectively. The binary cross-entropy loss used in this study is defined as follows:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{7}$$

where N denotes the total number of pixels in an image, y_i denotes the ground-truth label of the i-th pixel, and p_i denotes the model prediction for the i-th pixel. The final loss function of MDA-Net is defined as follows:

$$Loss = \alpha \times \mathcal{L}_{Dice} + \beta \times \mathcal{L}_{\mathrm{BCE}} \tag{8}$$

where α and β are hyperparameters that control the relative contributions of the Dice loss and the BCE loss, respectively. Based on extensive empirical experiments, we found that the model achieves its best segmentation performance when $\alpha = 0.5$ and $\beta = 1$.

3. Results

This section presents the evaluation protocol, implementation details of MDA-Net, and the corresponding visualization results.

3.1. Experimental Setup

All models in the comparative experiments were trained under identical hyperparameter and implementation settings to ensure a fair evaluation. Specifically, all experiments were conducted using Python 3.10 on an Intel(R) Xeon(R) Gold 6226R CPU @ 2.90 GHz and an NVIDIA RTX A6000 48 GB GPU. Adam was used as the optimizer, and the training objective combined BCE loss and Dice loss. The initial learning rate was set to 1×10^{-4}, the weight decay coefficient was set to 5×10^{-4}, the batch size was set to 8, and each model was trained for 100 epochs. Early stopping was triggered when the validation loss did not improve for 10 consecutive epochs, which prevented unnecessary computation and reduced the risk of overfitting. For each model, the optimal checkpoint was selected based on the validation set performance, and the final performance was reported on the test set.

3.2. Evaluation Metrics

To comprehensively evaluate the segmentation performance of the proposed MDA-Net, we adopt four metrics: Dice coefficient, Intersection over Union (IoU), Recall, and Precision. The Dice coefficient quantifies the overlap agreement between the predicted segmentation and the ground-truth annotation, and is computed as follows:

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{9}$$

where TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively. In the kidney tumor segmentation task, tumor pixels are treated as the positive class, whereas all non-tumor pixels are regarded as the negative class.

IoU quantifies the ratio between the intersection and the union of the predicted region and the ground-truth region, providing a more stringent measure of spatial agreement and region-level matching accuracy. It is computed as follows:

$$IoU = \frac{TP}{TP + FP + FN} \tag{10}$$

Recall measures the proportion of ground-truth positive samples that are correctly predicted as positive, thereby characterizing the model's detection sensitivity and its tendency to miss target regions. It is computed as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

Precision is defined as the proportion of predicted positive samples that are truly positive, reflecting the reliability of positive predictions and the extent of false alarms. It is computed as follows:

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

Collectively, these four metrics provide a comprehensive assessment of segmentation quality: higher values indicate closer agreement between the predicted masks and the GT, with greater spatial overlap and, consequently, more accurate delineation.

3.3. Ablation Experiments

Kidney tumor segmentation remains a highly challenging task. To evaluate the segmentation performance of MDA-Net, we conduct ablation studies to quantify the individual and combined effects of MSFE, DPFR, and CASA on the overall segmentation performance. Section 3.3.1 presents the ablation results of the proposed model on the KiTS19 dataset, and Section 3.3.2 reports the sensitivity analysis of the loss-function parameters on the same dataset. To ensure statistical robustness and reduce the influence of random variation during training, each experiment is independently repeated five times. The results are reported as the mean ± standard deviation across these repeated runs.

3.3.1. Ablation Experiment on Model Components

Table 1 presents the ablation results on the KiTS19 dataset. As shown, the introduction of these components consistently improves performance across all evaluation metrics. Starting from the baseline results of 0.7698 Precision, 0.7895 Recall, 0.7789, and 0.6929, the addition of MSFE yields improvements of 0.0488, 0.0329, 0.0268, and 0.0449 in Precision, Recall, Dice, and IoU, respectively. This improvement is primarily attributed to the MSFE block, which enhances multi-scale feature fusion within skip connections and effectively supplements shallow texture cues and boundary details.

Table 1. Ablation results for renal tumor segmentation on the KiTS19 dataset (Precision, Recall, Dice, and IoU). Note: Each metric is reported as mean ± standard deviation.

MSFE	DPFR	CASA	Precision	Recall	Dice	IoU
×	×	×	0.7698 ± 0.0122	0.7895 ± 0.0130	0.7789 ± 0.0125	0.6929 ± 0.0142
√	×	×	0.8186 ± 0.0015	0.8224 ± 0.0024	0.8057 ± 0.0032	0.7378 ± 0.0028
√	√	×	0.8205 ± 0.0017	0.8267 ± 0.0024	0.8126 ± 0.0035	0.7429 ± 0.0029
√	√	√	0.8379 ± 0.0176	0.8438 ± 0.0195	0.8325 ± 0.0183	0.7489 ± 0.0102

Building on this configuration, the incorporation of the DPFR block further improves Precision, Recall, and Dice by 0.19%, 0.43%, and 0.69%, respectively. By combining de-

formable sampling with pyramid-based contextual refinement, DPFR enlarges the effective receptive field and strengthens the internal representation of encoder features. In the final stage, the integration of CASA introduces coordinated channel attention and spatial attention into the decoder, thereby suppressing redundant background responses while enhancing lesion-relevant regions. As a result, both false positives and false negatives are reduced, and the overlap between predicted masks and ground truth is further improved. Overall, the complete MDA-Net outperforms the baseline by 6.81%, 5.43%, 5.36%, and 5.60% percentage points in Precision, Recall, Dice, and IoU, respectively. These results confirm the effectiveness of the three modules and indicate a clear synergistic effect. By capturing complementary cues from multi-scale context and discriminative feature channels, the proposed design improves both segmentation accuracy and robustness.

Table 2 further presents the ablation results of the DPFR Block. Compared with the baseline TransUNet, the use of deformable convolution alone improves Precision, Recall, Dice, and IoU from 0.7698, 0.7895, 0.7789, and 0.6929 to 0.7812, 0.7946, 0.7861, and 0.7018, respectively. This improvement indicates that deformable convolution enhances the ability of the model to adapt to irregular tumor morphology and local structural variations. Similarly, the use of ASPP alone also improves all evaluation metrics, with Dice and IoU reaching 0.7874 and 0.7031, respectively, suggesting that multi-scale contextual aggregation contributes to more effective representation of tumor regions. When deformable convolution and ASPP are further integrated into the proposed DPFR Block, the model achieves the best overall performance, with Precision, Recall, Dice, and IoU reaching 0.8016, 0.8073, 0.8042, and 0.7258, respectively. These results demonstrate that the DPFR Block effectively combines local deformation modeling and multi-scale context perception, thereby improving the segmentation of tumor boundaries, morphological variations, and low-contrast regions.

Table 2. Ablation results of DPFR block (Precision, Recall, Dice, and IoU). Note: Each metric is reported as mean ± standard deviation.

Model	Precision	Recall	Dice	IoU
TransUNet	0.7698 ± 0.0028	0.7895 ± 0.0031	0.7789 ± 0.0026	0.6929 ± 0.0034
TransUNet + DC	0.7812 ± 0.0125	0.7946 ± 0.0029	0.7861 ± 0.0023	0.7018 ± 0.0031
TransUNet + ASPP	0.7838 ± 0.0027	0.7937 ± 0.0019	0.7874 ± 0.0124	0.7031 ± 0.0132
TransUNet + DPFR	0.8016 ± 0.0021	0.8073 ± 0.0026	0.8042 ± 0.0022	0.7258 ± 0.0028

3.3.2. Sensitivity Analysis of Loss Function Parameters

To analyze the influence of BCE and Dice loss weights on segmentation performance, this study compares three representative weighting configurations. These configurations correspond to BCE-dominant weighting, balanced weighting, and Dice-dominant weighting, respectively. For each setting, MDA-Net is trained and tested under the same experimental configuration. Table 3 reports the experimental results obtained under different weight-parameter settings. The results show that the setting $\alpha = 0.50$ and $\beta = 1.00$ achieves the best and most balanced performance in terms of Dice and IoU. This result suggests that assigning a relatively higher weight to BCE loss provides more stable pixel-level supervision, which helps improve the discrimination between tumor and non-tumor pixels. Meanwhile, the Dice loss term still contributes to optimizing the overlap between the predicted mask and the ground-truth annotation, which is important for handling the foreground–background imbalance in kidney tumor segmentation. Therefore, the setting $\alpha = 0.50$ and $\beta = 1.00$ achieves a better balance between pixel-level classification accuracy and region-level overlap consistency, and is adopted in the final loss function.

Table 3. Sensitivity analysis of loss-function weights on the KiTS19 dataset. Note: Each metric is reported as mean ± standard deviation.

α	β	Precision	Recall	Dice	IoU
1.0	1.0	0.8296 ± 0.0189	0.8367 ± 0.0206	0.8241 ± 0.0194	0.7398 ± 0.0115
1.0	0.5	0.8312 ± 0.0193	0.8275 ± 0.0211	0.8179 ± 0.0202	0.7346 ± 0.0121
0.5	1.0	0.8379 ± 0.0176	0.8438 ± 0.0195	0.8325 ± 0.0183	0.7489 ± 0.0102

3.4. Comparison of the Latest Networks

To further validate the effectiveness and practicality of MDA-Net for kidney tumor segmentation, we compare it with several representative segmentation methods, including U-Net [10], U-Net++ [35], DeepLabv3+ [36], PIDNet-L [37], Swin-Unet [18], Spider-Net [38], CMUNeXt [39], and UCTransNet [40]. These baselines span classic CNN encoder–decoder designs, multi-scale context modeling, boundary-aware parsing, and Transformer-based U-shaped architectures. U-Net introduces skip connections to fuse semantic and spatial cues, whereas U-Net++ strengthens feature fusion through nested and dense skip pathways. DeepLabv3+ combines atrous spatial pyramid pooling with an encoder–decoder design and atrous separable convolution to capture multi-scale contextual information and refine object boundaries. PIDNet-L adopts a three-branch architecture that parses detail, context, and boundary information, and uses boundary attention to guide the fusion of detail and context features. Swin-Unet constructs a U-shaped encoder–decoder architecture based on hierarchical Swin Transformer blocks with shifted-window self-attention, thereby enabling long-range dependency modeling within a pure Transformer framework. Spider-Net employs a high-resolution multi-scale attention encoder together with a full-attention decoder for tumor segmentation. UCTransNet revisits the original skip connections of U-Net from a channel-wise perspective. It uses the CTrans module, which consists of Channel-wise Cross Fusion Transformer and Channel-wise Cross Attention, to reduce the semantic gap between encoder and decoder features. CMUNeXt designs a lightweight fully convolutional U-shaped segmentation network. It uses large-kernel depthwise convolution and inverted bottleneck structures to capture global contextual information, and introduces a Skip-Fusion block to enable efficient feature fusion between encoder and decoder representations.

Through comparison with these established methods, the performance and potential of MDA-Net for kidney tumor segmentation are further evaluated, and the results provide practical insights for both research and clinical application in this field.

All models are evaluated on the KiTS19 dataset, and performance is assessed using four metrics: Dice coefficient, Intersection over Union (IoU), Precision, and Recall. To ensure statistical robustness and reduce the influence of random variation during training, each experiment is independently repeated five times. The results are reported as the mean ± standard deviation across these repeated runs. The quantitative results are summarized in Table 4.

The quantitative comparison results for kidney tumor segmentation on the KiTS19 dataset show that MDA-Net achieves competitive and relatively balanced performance across the four evaluation metrics, reaching 83.79% Precision, 84.38% Recall, 83.25% Dice, and 74.89% IoU. It ranks first in Precision, Dice, and IoU, while achieving near-best Recall, which indicates improved overlap quality without compromising sensitivity to tumor regions.

Table 4. Quantitative comparison results for renal tumor segmentation on the KiTS19 dataset. Note: Each metric is reported as mean ± standard deviation.

Model	Precision	Recall	Dice	IoU
U-Net	0.7451 ± 0.0124	0.7535 ± 0.0141	0.7246 ± 0.0136	0.6562 ± 0.0152
U-Net++	0.7849 ± 0.0108	0.8083 ± 0.0126	0.7749 ± 0.0119	0.7026 ± 0.0133
Deeplabv3+	0.8019 ± 0.0097	**0.8442 ± 0.0112**	0.8067 ± 0.0105	0.7316 ± 0.0121
PIDNet-L	0.8275 ± 0.0089	0.8322 ± 0.0104	0.8172 ± 0.0096	0.7453 ± 0.0113
Swin-Unet	0.7924 ± 0.0115	0.8276 ± 0.0121	0.8252 ± 0.0108	0.7138 ± 0.0137
TransUNet	0.7698 ± 0.0122	0.7895 ± 0.0130	0.7789 ± 0.0125	0.6929 ± 0.0142
Spider-Net	0.7731 ± 0.0118	0.8079 ± 0.0124	0.7721 ± 0.0121	0.6982 ± 0.0139
CMUNeXt	0.7643 ± 0.0127	0.8147 ± 0.0135	0.7665 ± 0.0131	0.6965 ± 0.0146
UCTransNet	0.7833 ± 0.0111	0.8108 ± 0.0128	0.7781 ± 0.0117	0.7027 ± 0.0134
Ours	**0.8379 ± 0.0176**	0.8438 ± 0.0195	**0.8325 ± 0.0183**	**0.7489 ± 0.0102**

Underlined values indicate the second-best performance in each column. Bold values denote the best performance in each column.

Compared with the baseline TransUNet, MDA-Net improves Precision by 6.81% and Recall by 5.43%, while Dice and IoU increase by 5.36% and 5.60%, respectively. Beyond the baseline comparison, MDA-Net shows modest gains over several strong competing methods in specific metrics. Relative to PIDNet-L, which ranks second in both Precision and IoU, MDA-Net further improves Precision by 1.04% and IoU by 0.36%. Relative to Swin-Unet, which achieves the second-highest Dice score, MDA-Net improves Dice by 0.73%. Notably, although DeepLabv3+ attains the highest Recall of 84.42%, MDA-Net achieves a comparable Recall of 84.38% together with a substantially higher Precision of 83.79%, compared with 80.19% for DeepLabv3+, which suggests fewer false positives at a nearly equivalent detection level. Compared with CMUNeXt, MDA-Net achieves improvements of 7.36, 2.91, 6.60, and 5.24 percentage points in Precision, Recall, Dice, and IoU, respectively. Although CMUNeXt employs large-kernel depthwise convolution and skip-fusion operations to enhance contextual representation within a lightweight fully convolutional framework, its segmentation performance remains limited in this task, especially in Precision, Dice, and IoU. This result suggests that lightweight convolutional feature fusion alone may be insufficient to capture the complex boundary variations and heterogeneous tumor appearances associated with renal tumor segmentation. Compared with UCTransNet, MDA-Net also achieves consistent gains across all evaluation metrics, with improvements of 5.46, 3.30, 5.44, and 4.62 percentage points in Precision, Recall, Dice, and IoU, respectively. UCTransNet refines skip connections from a channel-wise Transformer perspective and reduces the semantic gap between encoder and decoder features. However, its relatively lower Dice and IoU indicate that channel-wise feature interaction alone may not be sufficient to address multi-scale tumor morphology and boundary ambiguity. In contrast, MDA-Net integrates multi-scale feature enhancement, deformable pyramid refinement, and coordinated channel–spatial attention, thereby enabling more accurate lesion localization and mask reconstruction.

Figure 6 presents representative segmentation examples produced by MDA-Net, the baseline model, and eight competing methods on the KiTS19 dataset. Owing to variations in tumor size and morphology, indistinct boundaries, and interference from adjacent tissues, different models exhibit distinct segmentation behaviors. U-Net and U-Net++ often produce incomplete masks with coarse contours, indicating limited boundary recovery. TransUNet and Swin-Unet tend to generate over-smoothed predictions and may show local mis-segmentation when tumor boundaries are ambiguous. DeepLabv3+, PIDNet-L, Spider-Net, CMUNeXt, and UCTransNet produce relatively continuous masks in some cases, but they still suffer from boundary deviations, under-segmentation, or inaccurate recovery of small and irregular lesions. In contrast, MDA-Net delineates tumor bound-

aries more precisely and remains stable under complex shapes and scale variations. By jointly capturing local details and global context with attention-guided feature refinement, MDA-Net improves discrimination in regions with uncertain boundaries and produces segmentation results that are more consistent with the ground-truth annotations.

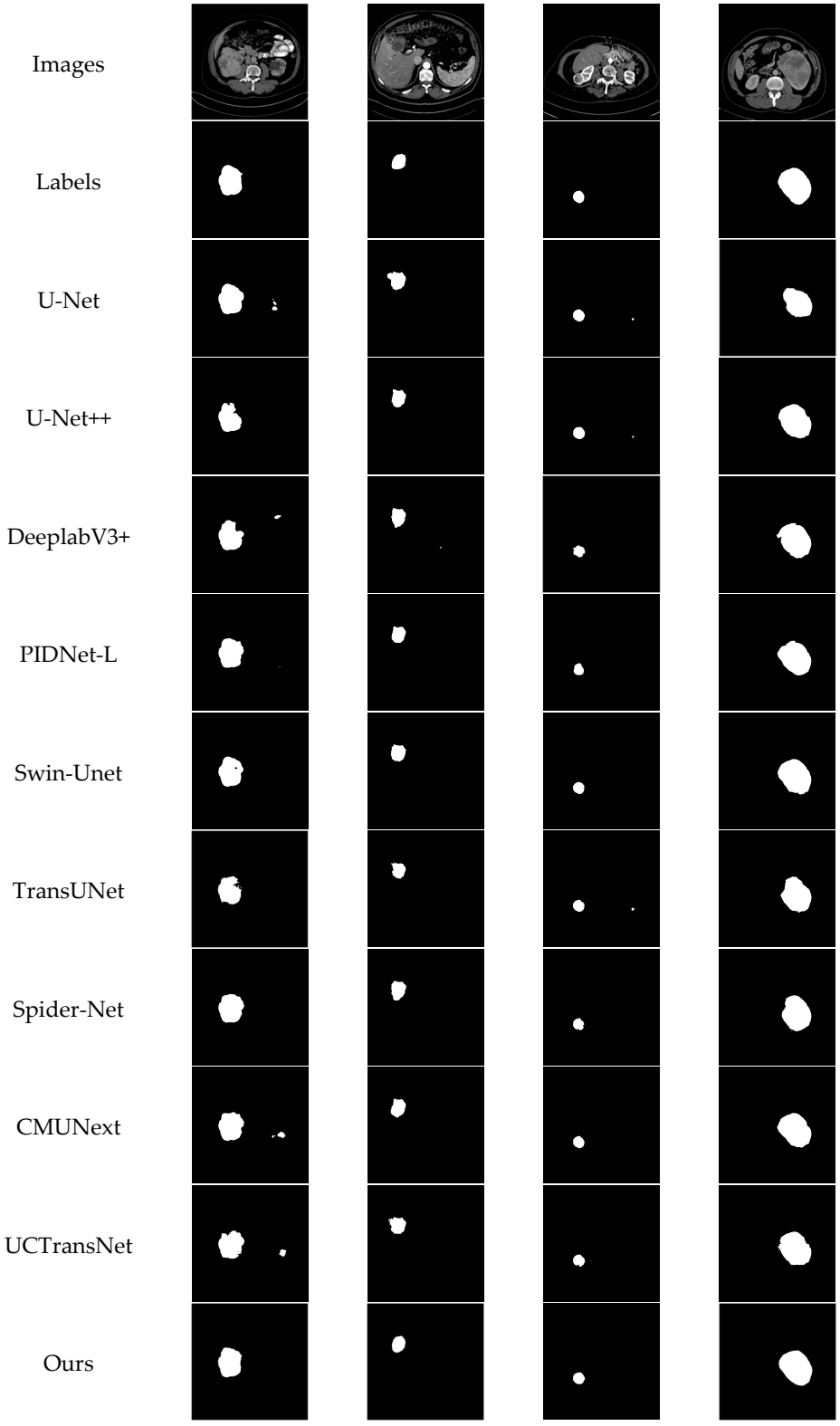


Figure 6. Example segmentation results of different models on the KiTS19 dataset.

Table 5 presents the independent evaluation results of different models on the KiTS21 dataset. Compared with the evaluation on KiTS19, this experiment provides a more direct assessment of the generalization performance of each model in an external test setting. As shown in Table 5, MDA-Net achieves the best overall performance across all four evaluation metrics, with 78.51% Precision, 81.40% Recall, 78.83% Dice, and 70.87% IoU. These results indicate that the proposed method maintains stable segmentation performance on an independent dataset.

Table 5. Quantitative comparison results for renal tumor segmentation on the KiTS21 dataset. Note: Each metric is reported as mean ± standard deviation.

Model	Precision	Recall	Dice	IoU
U-Net	0.7474 ± 0.0313	0.7820 ± 0.0325	0.7392 ± 0.0311	0.6614 ± 0.0308
U-Net++	0.7397 ± 0.0315	0.7797 ± 0.0325	0.7330 ± 0.0310	0.6530 ± 0.0307
Deeplabv3+	0.7623 ± 0.0292	<u>0.7903 ± 0.0293</u>	0.7632 ± 0.0285	0.6827 ± 0.0288
PIDNet-L	0.7716 ± 0.0981	0.7981 ± 0.0302	<u>0.7692 ± 0.0292</u>	<u>0.6933 ± 0.0292</u>
Swin-Unet	0.7671 ± 0.0911	0.7961 ± 0.0287	0.7603 ± 0.1817	0.6808 ± 0.1285
TransUNet	0.7329 ± 0.0331	0.7306 ± 0.0334	0.7037 ± 0.0320	0.6195 ± 0.0314
Spider-Net	<u>0.7731 ± 0.0289</u>	0.7961 ± 0.0286	0.7603 ± 0.0817	0.6867 ± 0.0285
CMUNeXt	0.7217 ± 0.0328	0.7853 ± 0.0327	0.7274 ± 0.0318	0.6495 ± 0.0315
UCTransNet	0.7360 ± 0.0308	0.7771 ± 0.0316	0.7287 ± 0.0303	0.6441 ± 0.03018
Ours	**0.7851 ± 0.0283**	**0.8140 ± 0.0282**	**0.7883 ± 0.0273**	**0.7087 ± 0.0279**

Underlined values indicate the second-best performance in each column. Bold values denote the best performance in each column.

Compared with the strongest competing methods, MDA-Net also shows consistent advantages. Specifically, MDA-Net improves Precision by 1.20 percentage points over Spider-Net, which obtains the second-best Precision. It improves Recall by 1.59 percentage points over PIDNet-L, which achieves the second-best Recall. For Dice and IoU, MDA-Net outperforms PIDNet-L by 1.91 and 1.54 percentage points, respectively. These results suggest that MDA-Net improves tumor-region sensitivity while enhancing region-level overlap consistency. In addition, compared with TransUNet, MDA-Net achieves improvements of 5.22, 8.34, 8.46, and 8.92 percentage points in Precision, Recall, Dice, and IoU, respectively. This further demonstrates the effectiveness of the proposed multi-scale feature enhancement, deformable pyramid refinement, and channel–spatial attention modules.

Overall, the independent evaluation results on KiTS21 show that MDA-Net achieves better generalization than the compared methods and provides more accurate tumor localization and mask reconstruction in cross-dataset evaluation.

Figure 7 presents representative segmentation examples from different models on the KiTS21 dataset. Compared with KiTS19, KiTS21 provides an independent evaluation setting and includes tumors with diverse sizes, irregular shapes, and ambiguous boundaries, further increasing the difficulty of accurate segmentation. Visually, U-Net and TransUNet tend to produce incomplete masks and frequently miss substantial tumor regions, especially when lesion boundaries are unclear or tumor morphology is complex. U-Net++, CMUNeXt, and UCTransNet can capture parts of tumor regions, but their predictions often exhibit under-segmentation and fail to recover the full tumor extent. DeepLabv3+, PIDNet-L, Swin-Unet, and Spider-Net generate relatively continuous masks in some cases, but they still suffer from boundary deviations, over-smoothed contours, or small false-positive regions. In contrast, MDA-Net produces segmentation masks that are more consistent with the ground-truth annotations in most examples. The predicted regions preserve more complete tumor structures and show better agreement with irregular lesion boundaries. These qualitative results further support the robustness and generalization ability of MDA-Net in cross-dataset evaluation.

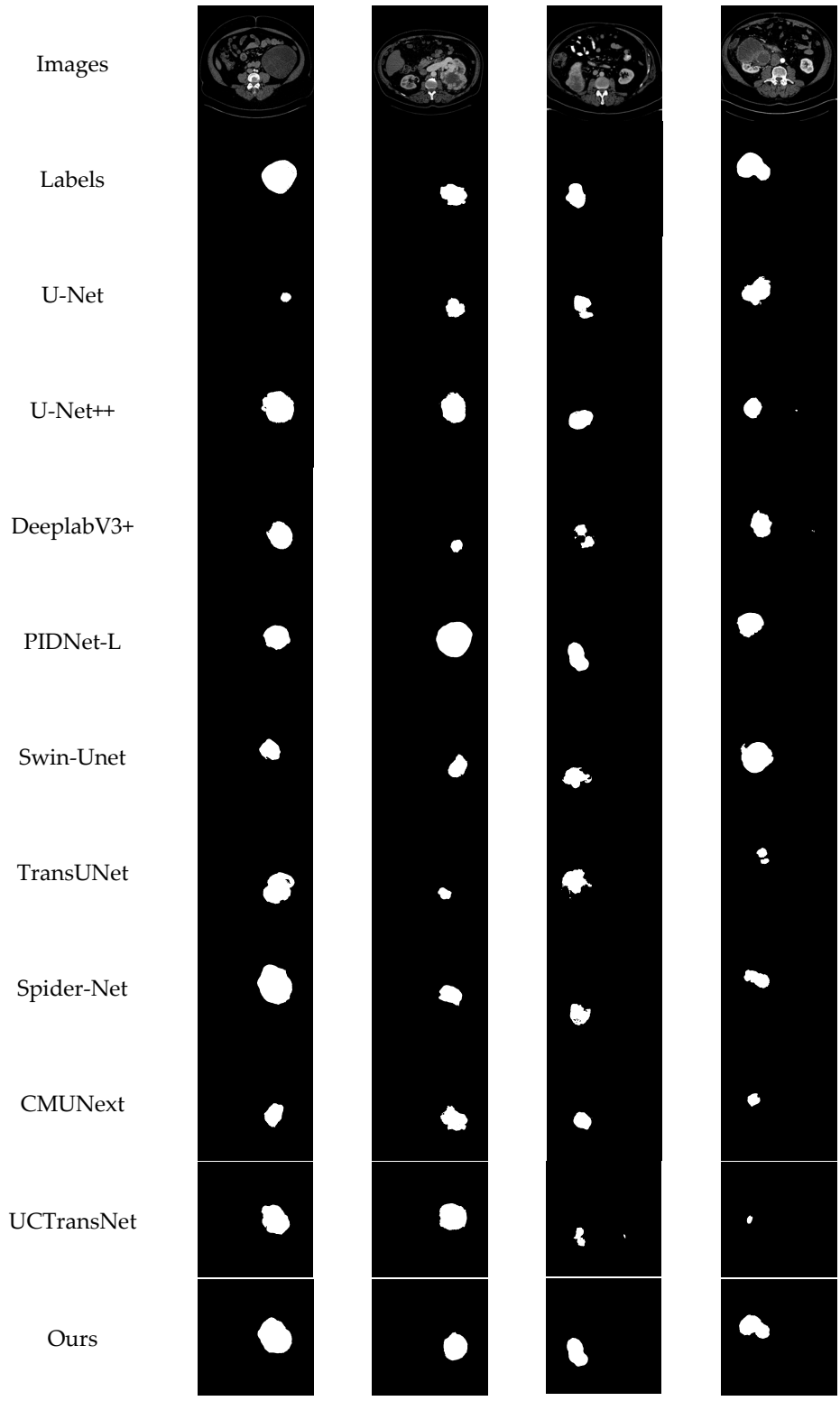


Figure 7. Example segmentation results of different models on the KiTS21 dataset.

3.5. Grad-CAM Based Interpretability Analysis

We further employ Grad-CAM as an interpretability technique to visualize the activations of the final convolutional layer in the proposed model. This technique highlights the image regions that make stronger contributions to model prediction, thereby providing visual evidence for understanding the decision basis of the model and improving model transparency [41–43]. To strengthen the interpretability analysis beyond qualitative visual-

ization, we further conduct a quantitative experiment to measure the spatial agreement between the Grad-CAM attention maps and the ground-truth tumor masks in the KiTS19 test set. Specifically, each Grad-CAM heatmap is first resized to the same spatial resolution as the corresponding tumor mask and normalized to the range of $[0, 1]$. The normalized heatmap is then binarized using a threshold of 0.5, and the Intersection over Union is computed between the binarized attention map and the corresponding ground-truth tumor mask. Across three random-sampling validation runs with a sampling rate of 0.3, the mean IoU is 0.5857, with a standard deviation of 0.0026. This metric provides a quantitative measure of whether the regions attended to by the model are spatially consistent with the clinically relevant tumor regions. In addition, three cases are randomly selected from the test set for qualitative visualization, where the Grad-CAM heatmaps are presented together with the input images and tumor masks to provide an intuitive interpretation of model behavior.

Figure 8 presents Grad-CAM heatmaps, GT, and the corresponding predicted masks of MDA-Net on representative KiTS19 cases. In each example, the heatmap visualizes the image regions that contribute most strongly to the segmentation decision, where warmer colors indicate higher relevance. GT denotes the tumor regions manually annotated by clinical experts. Overall, the predicted masks closely match the GT in both extent and shape. Meanwhile, Grad-CAM activations are concentrated on tumor regions and areas adjacent to tumor boundaries, with limited responses in irrelevant background tissues. These qualitative results indicate that the predictions of MDA-Net are primarily driven by tumor-related cues. They also suggest that the proposed components improve the overall ability of the model to extract tumor-relevant information and to attend to features associated with tumor boundaries and salient regions in medical images.

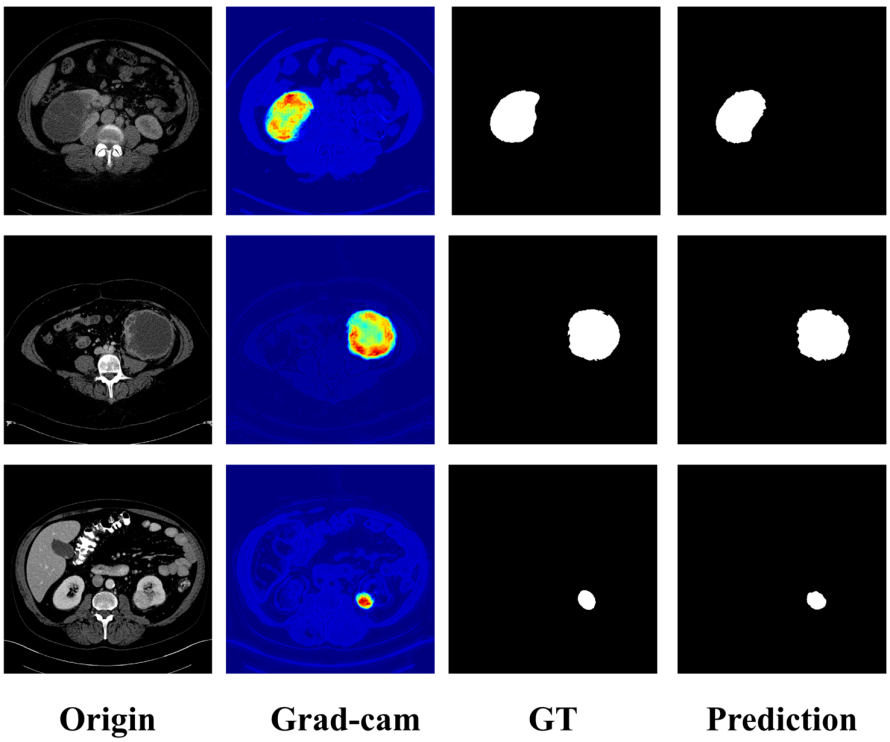


Figure 8. Grad-CAM visualizations for three representative test samples in KiTS19.

3.6. Computational Efficiency Comparison

In the complexity analysis experiments, this study evaluates key efficiency indicators, including inference efficiency, parameter count, and floating-point operations (FLOPs). Inference efficiency is defined as the average time required for each model to generate

predictions on the target dataset. Parameter count reflects the total number of learnable parameters in the model. FLOPs are computed using an input tensor size of $[1, 3, 512, 512]$ to characterize the computational cost of a single forward pass. The KiTS21 dataset is used for evaluation, and all experiments are conducted under identical hardware settings to ensure fair and reliable comparisons.

As shown in Table 6, CMUNeXt has the lowest computational cost, with 59.34 G FLOPs, and the smallest parameter count, with 3.15 M parameters. PIDNet-L achieves the fastest inference speed, with an inference time of 1.82 ms. UCTransNet has the highest FLOPs among the compared methods, requiring 344.02 G FLOPs and 158.27 ms for inference. The proposed MDA-Net requires 328.06 G FLOPs and contains 113.03 M parameters, reflecting the additional computational cost introduced by the multi-scale feature enhancement, deformable pyramid refinement, and channel–spatial attention modules. Although MDA-Net is not the lightest model, it achieves an inference time of 31.21 ms, which is faster than those of TransUNet, Spider-Net, and UCTransNet by 12.02 ms, 56.77 ms, and 127.06 ms, respectively. Compared with Transformer-based and attention-enhanced models, these results indicate that MDA-Net maintains acceptable inference efficiency while improving segmentation accuracy. Therefore, the proposed model achieves a reasonable trade-off between computational complexity and segmentation performance.

Table 6. Complexity study on the KiTS21 dataset.

Model	FLOPs (G)	Parameter Count (M)	Inference Time (ms)
U-Net	248.95	13.40	20.72
U-Net++	279.23	<u>9.16</u>	30.07
Deeplabv3+	114.09	46.62	<u>7.16</u>
PIDNet-L	72.04	37.31	**1.82**
Swin-Unet	<u>61.79</u>	41.39	13.26
TransUNet	252.85	105.34	43.23
Spider-Net	221.82	81.06	87.98
CMUNeXt	**59.34**	**3.15**	28.45
UCTransNet	344.02	67.23	158.27
Ours	328.06	113.03	31.21

Underlined values indicate the second-best performance in each column. Bold values denote the best performance in each column.

4. Discussion

Experimental results on the KiTS19 and KiTS21 dataset support the effectiveness of MDA-Net for kidney tumor segmentation and show competitive performance compared with existing methods. By integrating a hybrid encoder with the multi-scale feature enhancement (MSFE) Block, MDA-Net is able to capture fine-grained texture cues and multi-scale contextual information, which is essential for separating low-contrast tumors from surrounding tissues. The deformable pyramid feature refinement (DPFR) Block further strengthens the ability of the model to accommodate tumors with varying shapes and sizes while preserving high-resolution representations, making the network particularly effective for lesions with irregular boundaries or pronounced intensity variation. In the decoder, the channel-and-spatial attention (CASA) Block refines boundary delineation and suppresses background interference, thereby enabling the network to focus more precisely on tumor regions. Compared with the baseline model, MDA-Net shows improved segmentation performance for tumors with irregular shapes, small volumes, or low contrast. Compared with strong competing methods, the numerical gains are modest, indicating that further validation is needed to determine whether these improvements translate into clinically meaningful benefits. These findings highlight the importance of multi-scale feature integration and attention-guided refinement in medical image segmentation, where accu-

rate boundary detection and reliable tumor localization are of critical importance. Moreover, Grad-CAM-based interpretability analysis improves clinical transparency by highlighting tumor-relevant regions and providing insights into the decision-making process of the model.

Despite its strong performance, MDA-Net still has several limitations. First, the inference efficiency of the model may be insufficient to satisfy real-time clinical requirements, particularly for high-resolution CT scans. Second, dependence on large-scale fully annotated datasets raises concerns about training scalability and annotation cost. In addition, the generalization ability of MDA-Net across other imaging modalities, such as PET and MRI, has not been thoroughly evaluated, which may limit its applicability in multimodal clinical settings. Addressing these issues is essential for translating MDA-Net from a research prototype into broader clinical practice. In this regard, the exploration of lightweight variants through model pruning or knowledge distillation may improve deployment feasibility in time-sensitive clinical workflows.

5. Conclusions

This paper proposes MDA-Net, an end-to-end network for kidney tumor segmentation based on enhanced multi-scale feature extraction and attention-guided refinement. The proposed framework integrates the MSFE, DPFR, and CASA modules to improve multi-scale contextual modeling, adapt to irregular tumor morphology, and refine tumor-relevant features during decoder reconstruction. By combining these complementary components, MDA-Net enhances fine-grained boundary delineation, lesion localization, and mask reconstruction for renal tumor segmentation.

Extensive experiments on the KiTS19 dataset demonstrate that MDA-Net achieves competitive segmentation performance compared with representative CNN-based, Transformer-based, and attention-enhanced segmentation networks. Independent evaluation on the KiTS21 dataset further shows that the proposed method maintains stable performance under cross-dataset testing, indicating improved robustness and generalizability. Ablation experiments confirm the individual and complementary contributions of MSFE, DPFR, and CASA, while Grad-CAM-based visualization provides additional evidence that the model focuses on tumor-related regions and boundary-sensitive structures. Computational efficiency analysis also shows that, although MDA-Net introduces additional computational cost, it maintains acceptable inference speed compared with several Transformer-based and attention-enhanced methods.

Despite these advantages, MDA-Net still has several limitations. The relatively large parameter count and computational cost of the model may restrict its deployment in real-time or resource-constrained clinical scenarios. In addition, the current validation is limited to CT-based kidney tumor segmentation, and the applicability of the method to other imaging modalities or multi-center clinical data requires further investigation. Future work will focus on developing lightweight variants through model compression, pruning, or knowledge distillation, extending validation to larger and more diverse datasets, and exploring multimodal integration to further improve clinical applicability and generalization.

Author Contributions: Methodology, S.L.; Resources, J.C.; Software, Y.C.; Supervision, S.L.; Writing—original draft, Y.C.; Writing—review and editing, S.L., J.C. and L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by Beijing Natural Science Foundation (No. 4222022) and National Key Research and Development Program of China (Grant No. 2020YFB2104402).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are obtained from the KiTS19 and KiTS21 challenges, which are publicly available at their official websites.

Acknowledgments: The author expresses deep gratitude to the anonymous reviewers.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gharaibeh, M.; Alzu'bi, D.; Abdullah, M.; Hmeidi, I.; Al Nasar, M.R.; Abualigah, L.; Gandomi, A.H. Radiology Imaging Scans for Early Diagnosis of Kidney Tumors: A Review of Data Analytics-Based Machine Learning and Deep Learning Approaches. *BDCC* **2022**, *6*, 29. [CrossRef]
2. Xia, K.; Yin, H.; Zhang, Y. Deep Semantic Segmentation of Kidney and Space-Occupying Lesion Area Based on SCNN and ResNet Models Combined with SIFT-Flow Algorithm. *J. Med. Syst.* **2019**, *43*, 2. [CrossRef]
3. Bello, A.K.; Okpechi, I.G.; Levin, A.; Ye, F.; Damster, S.; Arruebo, S.; Donner, J.-A.; Caskey, F.J.; Cho, Y.; Davids, M.R.; et al. An Update on the Global Disparities in Kidney Disease Burden and Care across World Countries and Regions. *Lancet Glob. Health* **2024**, *12*, e382–e395. [CrossRef]
4. Rose, T.L.; Kim, W.Y. Renal Cell Carcinoma: A Review. *JAMA* **2024**, *332*, 1001. [CrossRef]
5. Kan, H.-C.; Fan, G.-M.; Wei, M.-H.; Lin, P.-H.; Shao, I.-H.; Yu, K.-J.; Chien, T.-H.; Pang, S.-T.; Wu, C.-T.; Peng, S.-J. Automated Kidney Tumor Segmentation in CT Images Using Deep Learning: A Multi-Stage Approach. *Acad. Radiol.* **2025**, *32*, 7193–7203. [CrossRef] [PubMed]
6. Xu, J.; Gao, C.; Zhang, J.; Lu, J.; Xuan, Y.; Wang, S.; Bu, C. Advancements in Imaging Technologies and AI Integration for Neurodegenerative Disease Management: A Narrative Review. *Mol. Imaging* **2025**, *24*, 15353508251393056. [CrossRef]
7. Kuai, H.; Huang, J.X.; Tao, X.; Pasi, G.; Yao, Y.; Liu, J.; Zhong, N. Web Intelligence (WI) 3.0: In Search of a Better-Connected World to Create a Future Intelligent Society. *Artif. Intell. Rev.* **2025**, *58*, 265. [CrossRef]
8. Deng, Y.; Liu, X.; Shao, L.; Wang, K.; Wang, X.; Gao, Q. Kidney Tumor Segmentation Based on DWR-SegFormer. *Electronics* **2024**, *13*, 3226. [CrossRef]
9. Zhang, Y.; Wang, Y.; Hou, F.; Yang, J.; Xiong, G.; Tian, J.; Zhong, C. Cascaded Volumetric Convolutional Network for Kidney Tumor Segmentation from CT Volumes. *arXiv* **2019**, arXiv:1910.02235.
10. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2015; Volume 9351, pp. 234–241, ISBN 978-3-319-24573-7.
11. Yang, G.; Li, G.; Pan, T.; Kong, Y.; Wu, J.; Shu, H.; Luo, L.; Dillenseger, J.-L.; Coatrieux, J.-L.; Tang, L.; et al. Automatic Segmentation of Kidney and Renal Tumor in CT Images Based on 3D Fully Convolutional Neural Network with Pyramid Pooling Module. In *Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018*; IEEE: New York, NY, USA, 2018; pp. 3790–3795.
12. Zhao, Z.; Chen, H.; Li, J.; Wang, L. Boundary Attention U-Net for Kidney and Kidney Tumor Segmentation. In *Proceedings of the 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, UK, 11–15 July 2022*; IEEE: New York, NY, USA, 2022; pp. 1540–1543.
13. Guo, J.; Zeng, W.; Yu, S.; Xiao, J. RAU-Net: U-Net Model Based on Residual and Attention for Kidney and Kidney Tumor Segmentation. In *Proceedings of the 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, 15–17 January 2021*; IEEE: New York, NY, USA, 2021; pp. 353–356.
14. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 38–45.
15. Yao, D.; Shao, Y. A Data Efficient Transformer Based on Swin Transformer. *Vis. Comput.* **2024**, *40*, 2589–2598. [CrossRef]
16. Shen, T.; Xu, H. Medical Image Segmentation Based on Transformer and HarDNet Structures. *IEEE Access* **2023**, *11*, 16621–16630. [CrossRef]
17. 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 2021. *arXiv* **2021**, arXiv:2102.04306.
18. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In *Computer Vision—ECCV 2022 Workshops*; Karlinsky, L., Michaeli, T., Nishino, K., Eds.; Lecture Notes in Computer Science; Springer Nature: Cham, Switzerland, 2023; Volume 13803, pp. 205–218, ISBN 978-3-031-25065-1.

19. Shen, Z.; Yang, H.; Zhang, Z.; Zheng, S. Automated Kidney Tumor Segmentation with Convolution and Transformer Network. In *Kidney and Kidney Tumor Segmentation*; Heller, N., Isensee, F., Trofimova, D., Tejpaul, R., Papanikolopoulos, N., Weight, C., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2022; Volume 13168, pp. 1–12, ISBN 978-3-030-98384-0.
20. Hou, B.; Zhang, G.; Liu, H.; Qin, Y.; Chen, Y. Dcctnet: Kidney Tumors Segmentation Based on Dual-Level Combination of Cnn and Transformer. In *Proceedings of the 2024 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 27 October 2024*; IEEE: New York, NY, USA, 2024; pp. 3112–3116.
21. Hu, W.; Yang, S.; Guo, W.; Xiao, N.; Yang, X.; Ren, X. STC-UNet: Renal Tumor Segmentation Based on Enhanced Feature Extraction at Different Network Levels. *BMC Med. Imaging* **2024**, *24*, 179. [CrossRef]
22. Yuan, F.; Zhang, Z.; Fang, Z. An Effective CNN and Transformer Complementary Network for Medical Image Segmentation. *Pattern Recognit.* **2023**, *136*, 109228. [CrossRef]
23. Mazher, M.; Qayyum, A.; Puig, D.; Abdel-Nasser, M. Deep Learning-based Survival Prediction of Brain Tumor Patients Using Attention-guided 3D Convolutional Neural Network with Radiomics Approach from Multimodality Magnetic Resonance Imaging. *Int. J. Imaging Syst. Tech.* **2024**, *34*, e23010. [CrossRef]
24. Akter, A.; Nosheen, N.; Ahmed, S.; Hossain, M.; Yousuf, M.A.; Almoyad, M.A.A.; Hasan, K.F.; Moni, M.A. Robust Clinical Applicable CNN and U-Net Based Algorithm for MRI Classification and Segmentation for Brain Tumor. *Expert Syst. Appl.* **2024**, *238*, 122347. [CrossRef]
25. Rasool, N.; Bhat, J.I.; Aoun, N.B.; Alharthi, A.; Wani, N.A.; Chopra, V.; Anwar, M.S. ResMHA-Net: Enhancing Glioma Segmentation and Survival Prediction Using a Novel Deep Learning Framework. *CMC* **2024**, *81*, 885–909. [CrossRef]
26. Humpire-Mamani, G.E. *KiTS19: Kidney and Kidney Tumor Segmentation in CT Scans Using a 3D U-Net Based Network with Additional Tasks*; Radboud University Medical Center: Nijmegen, The Netherlands, 2019.
27. Heller, N.; Isensee, F.; Trofimova, D.; Tejpaul, R.; Zhao, Z.; Chen, H.; Wang, L.; Golts, A.; Khapun, D.; Shats, D.; et al. The KiTS21 Challenge: Automatic Segmentation of Kidneys, Renal Tumors, and Renal Cysts in Corticomedullary-Phase CT 2023. *arXiv* **2023**, arXiv:2307.01984.
28. Mo, Y.; Wu, Y.; Yang, X.; Liu, F.; Liao, Y. Review the State-of-the-Art Technologies of Semantic Segmentation Based on Deep Learning. *Neurocomputing* **2022**, *493*, 626–646. [CrossRef]
29. Sun, P.; Mo, Z.; Hu, F.; Song, X.; Mo, T.; Yu, B.; Zhang, Y.; Chen, Z. 2.5D MFFAU-Net: A Convolutional Neural Network for Kidney Segmentation. *BMC Med. Inform. Decis. Mak.* **2023**, *23*, 92. [CrossRef]
30. Wazir, S.; Kim, D. Rethinking Decoder Design: Improving Biomarker Segmentation Using Depth-to-Space Restoration and Residual Linear Attention. In *Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10 June 2025*; IEEE: New York, NY, USA, 2025; pp. 30861–30871.
31. Ramachandran, P.; Zoph, B.; Le, Q.V. Searching for Activation Functions 2017. *arXiv* **2017**, arXiv:1710.05941.
32. Ji, Z.; Mu, J.; Liu, J.; Zhang, H.; Dai, C.; Zhang, X.; Ganchev, I. ASD-Net: A Novel U-Net Based Asymmetric Spatial-Channel Convolution Network for Precise Kidney and Kidney Tumor Image Segmentation. *Med. Biol. Eng. Comput.* **2024**, *62*, 1673–1687. [CrossRef]
33. Yuan, N.; Zhang, Y.; Lv, K.; Liu, Y.; Yang, A.; Hu, P.; Yu, H.; Han, X.; Guo, X.; Li, J.; et al. HCA-DAN: Hierarchical Class-Aware Domain Adaptive Network for Gastric Tumor Segmentation in 3D CT Images. *Cancer Imaging* **2024**, *24*, 63. [CrossRef] [PubMed]
34. Silveira, J.T.; Girisha, S.; Kundapur, P.P. Automated Lumbar Spine Segmentation in MRI Using an Enhanced U-Net with Inception Module and Dual-Output Mechanism. *Sci. Rep.* **2025**, *15*, 39215. [CrossRef]
35. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J.M.R.S., Bradley, A., Papa, J.P., Belagiannis, V., et al., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11045, pp. 3–11, ISBN 978-3-030-00888-8.
36. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11211, pp. 833–851, ISBN 978-3-030-01233-5.
37. Xu, J.; Xiong, Z.; Bhattacharyya, S.P. PIDNet: A Real-Time Semantic Segmentation Network Inspired by PID Controllers. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023*; IEEE: New York, NY, USA, 2023; pp. 19529–19539. [CrossRef]
38. Peng, Y.; Hu, X.; Hao, X.; Liu, P.; Deng, Y.; Li, Z. Spider-Net: High-Resolution Multi-Scale Attention Network with Full-Attention Decoder for Tumor Segmentation in Kidney, Liver and Pancreas. *Biomed. Signal Process. Control* **2024**, *93*, 106163. [CrossRef]

39. Tang, F.; Ding, J.; Wang, L.; Ning, C.; Zhou, S.K. CMUNeXt: An Efficient Medical Image Segmentation Network Based on Large Kernel and Skip Fusion 2023. In Proceedings of the 2024 IEEE International Symposium on Biomedical Imaging (ISBI), Athens, Greece, 27–30 May 2024.
40. Wang, H.; Cao, P.; Wang, J.; Zaiane, O.R. UCTransNet: Rethinking the Skip Connections in U-Net from a Channel-Wise Perspective with Transformer. *AAAI* **2022**, *36*, 2441–2449. [CrossRef]
41. Kuai, H.; Chen, J.; Tao, X.; Cai, L.; Imamura, K.; Matsumoto, H.; Liang, P.; Zhong, N. Never-Ending Learning for Explainable Brain Computing. *Adv. Sci.* **2024**, *11*, 2307647. [CrossRef]
42. Tang, D.; Chen, J.; Ren, L.; Wang, X.; Li, D.; Zhang, H. Reviewing CAM-Based Deep Explainable Methods in Healthcare. *Appl. Sci.* **2024**, *14*, 4124. [CrossRef]
43. Gunashekar, D.D.; Bielak, L.; Hägele, L.; Oerther, B.; Benndorf, M.; Grosu, A.-L.; Brox, T.; Zamboglou, C.; Bock, M. Explainable AI for CNN-Based Prostate Tumor Segmentation in Multi-Parametric MRI Correlated to Whole Mount Histopathology. *Radiat. Oncol.* **2022**, *17*, 65. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.