Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–20

DOI: 10.1177/20552076241291306
journals.sagepub.com/home/dhj

Sage

MLAU-Net: Deep supervised attention and hybrid loss strategies for enhanced segmentation of low-resolution kidney ultrasound

Rashid Khan[1,2,3], Asim Zaman[4,5], Chao Chen[1,2], Chuda Xiao[6], Wen Zhong[7], Yang Liu[7], Haseeb Hassan[4], Liyilei Su[1,2,3], Weiguo Xie[6], Yan Kang[2,4] and Bingding Huang[1]

Abstract

Objective: The precise segmentation of kidneys from a 2D ultrasound (US) image is crucial for diagnosing and monitoring kidney diseases. However, achieving detailed segmentation is difficult due to US images' low signal-to-noise ratio and low-contrast object boundaries.

Methods: This paper presents an approach called deep supervised attention with multi-loss functions (MLAU-Net) for US segmentation. The MLAU-Net model combines the benefits of attention mechanisms and deep supervision to improve segmentation accuracy. The attention mechanism allows the model to selectively focus on relevant regions of the kidney and ignore irrelevant background information, while the deep supervision captures the high-dimensional structure of the kidney in US images.

Results: We conducted experiments on two datasets to evaluate the MLAU-Net model's performance. The Wuerzburg Dynamic Kidney Ultrasound (WD-KUS) dataset with annotation contained kidney US images from 176 patients split into training and testing sets totaling 44,880. The Open Kidney Dataset's second dataset has over 500 B-mode abdominal US images. The proposed approach achieved the highest dice, accuracy, specificity, Hausdorff distance (HD95), recall, and Average Symmetric Surface Distance (ASSD) scores of 90.2%, 98.26%, 98.93%, 8.90 mm, 91.78%, and 2.87 mm, respectively, upon testing and comparison with state-of-the-art U-Net series segmentation frameworks, which demonstrates the potential clinical value of our work.

Conclusion: The proposed MLAU-Net model has the potential to be applied to other medical image segmentation tasks that face similar challenges of low signal-to-noise ratios and low-contrast object boundaries.

Keywords

Kidney ultrasound segmentation, WD-KUS dataset, attention mechanism, U-Net, deep learning

Submission date: 19 February 2024; Acceptance date: 25 September 2024

[1]College of Big Data and Internet, Shenzhen Technology University, Shenzhen, China
[2]College of Applied Sciences, Shenzhen University, Shenzhen, China
[3]Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen, China
[4]College of Health Science and Environmental Engineering, Shenzhen Technology University, Shenzhen, China
[5]School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen, China
[6]Wuerzburg Dynamics Inc., Shenzhen, China
[7]Department of Urology, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China

Corresponding author:
Bingding Huang, College of Big Data and Internet, Shenzhen Technology University, Shenzhen, 518188, China.
Email: huangbingding@sztu.edu.cn

Yan Kang, College of Applied Sciences, Shenzhen University, Shenzhen, 518060, China and College of Health Science and Environmental Engineering, Shenzhen Technology University, Shenzhen, 518188, China.
Email: kangyan@sztu.edu.cn

Introduction

Ultrasound (US) is one of the mainstays of medical diagnostics for its broad applicability and efficacy. It allows for both a qualitative and quantitative evaluation, offering real-time insights without ionizing radiation. Despite its advantages, US imaging encounters challenges such as artifact presence, noise interference, and subjective interpretation, distinguishing it from modalities like X-ray, computed tomography (CT), and magnetic resonance imaging (MRI).[1] A valid strategy to deal with many of these limitations is the use of computer-aided diagnosis (CAD), which greatly enhances the diagnostic accuracy. Effective image segmentation significantly, enhances the analysis of US images, ensuring precise delineation of anatomical structures for thorough interpretation and clinical diagnosis. Although several techniques have been developed for segmenting US images, their segmentation capability is still inadequate when dealing with relatively complex images.[2] US, one of the most commonly used imaging modalities, is a ubiquitous and effective screening and diagnostic tool for physicians and radiologists. In particular, US imaging is widely used in prenatal screening worldwide due to its safety, low cost, non-invasiveness, real-time imaging, convenience, and user experience.[3]

Recently, there has been a growing interest in automated medical image analysis methods, including kidney ultrasound (KUS) segmentation. Segmentation of KUS images is a difficult task due to the complexity of the kidney and its structure and speckle noise in US images.[4,5] In contrast, accurate segmentation of the kidney region is essential for various clinical applications such as disease diagnosis, surgical planning, and treatment monitoring.[6]Traditionally, KUS image segmentation involved manual contouring of the kidney, which is time-consuming and subject to user variability.[7,8] To address these limitations, several semi-automatic segmentation methods have been proposed.[9–11] Conversely, these methods rely heavily on manual initialization and may produce errors due to unclear boundaries and uneven intensity distribution.[12] Consequently, there is a growing need for automatic and robust KUS segmentation techniques to improve segmentation accuracy and efficiency.

Interest in medical image segmentation (MIS) using deep learning (DL) techniques has grown significantly, and various convolutional neural network (CNN) architectures such as U-Net,[13] FCN,[14] CPFNet,[15] Deeplabv3,[16] and SegNet have been proposed.[17] These approaches have greatly improved segmentation accuracy and reliability compared to traditional segmentation networks. The deep neural networks are used to learn high-level representations of medical images and capture the spatial relationships between complex anatomical structures. The stimulation method includes depth supervised attention (DSA) and multiple loss functions (MLFs). The use of various loss functions and an attention module enhances the flexibility and precision of segmentation results. The observation mechanism allows the network to selectively highlight relevant image regions while using MLFs, allowing the model to cover different aspects of the segmentation task and improve the model and overall performance. For example, Chen et al. proposed a similar approach for US kidney segmentation using a modified U-Net architecture with a deep controlled attention mechanism and multiloss features.[18] They reported a significant improvement in segmentation accuracy compared to existing methods, showing the potential of this approach for KUS segmentation problems. Moreover, Feng et al. presented a CNN-based approach that integrates multi-level contextual information and multi-task learning for re-segmenting US images. Their segmentation methods were more accurately applied than conventional ones. In addition, a DL-based framework using a bidirectional network with multilevel feature fusion was proposed for kidney segmentation in 3D-US volumes, which showed strong performance in segmenting kidneys of different shapes and sizes.[19]

Furthermore, attention mechanisms have been commonly employed in MIS tasks. Attention-guided U-Net for brain tumor segmentation based on MRI images has improved accuracy by incorporating attention mechanisms to guide the network and focus tumor regions.[20] In addition, a self-supervised attention-guided network has been introduced for cardiac image segmentation, which uses self-supervised learning and attention mechanisms to improve segmentation accuracy, especially when labeled data is limited.[21] Multi-loss features have also shown their ability to increase the performance of DL networks in MIS tasks. The accuracy of pancreas segmentation from CT images was improved by incorporating a multiloss attention-guided network with dice loss and focal loss functions.[22] Moreover, a multi-task DL approach with MLFs was shown to improve the segmentation accuracy of liver tumors from CT images.[23] However, kidney segmentation in 2D US images is still a problem in most cases, as the signal-to-noise ratio is low and the contours of the object are poorly contrasted. In most recently proposed CNNs for segmentation,[24] they do not allow capturing the whole high-dimensional structure of a 2D kidney from US entirely, as loss functions based on weak supervision tend to be inefficient to integrate spatial relationships between neighboring pixels; such approaches fail to produce regularly shaped segmentation masks.

To improve the performance of MIS and reduce the complexity of the network structure, we proposed a model called deep tracking MLAU-Net for KUS image segmentation. The model consists of 2D U-Net with two main regulatory components: attention and depth tracking. The main contributions of our work are threefold. (a) The proposed MLAU-Net introduces new depth enhancements and attention gates that enrich the model and allow it to focus on essential image features while ignoring noise. (b) The model includes extensive monitoring as an adjustment method and

encompasses target functions. Across the decoder layers, these target functions deal with small data and deeper networks. The process involves carefully transitioning from deeper blocks to the original segmentation size, ensuring that the model effectively captures high-dimensional structures. (c) A preprocessing pipeline, including custom normalization and efficient data augmentation during training, ensures model reliability and efficiency using various US images. These advances pertain to enhancing the accuracy and efficiency of medical image analysis in CAD. The remaining sections of this work are organized in a uniform manner. In the "Related work" section, various related studies are reviewed and the characteristics and challenges associated with existing approaches are described, while the proposed MLAU-Net approach for KUS image segmentation is detailed in the "Materials and methods" section. The performance of the various measures is reviewed in the "The Experiment" section; finally, the conclusions of the proposed approach are presented in the "Experimental results and discussion" section.

Related work

In recent years, considerable research has been conducted on KUS segmentation using DL methods. Several studies have proposed different methods to achieve accurate and efficient kidney segmentation in US images. U-Net,[13] a pioneering encoder–decoder CNN-based framework, has demonstrated exceptional image segmentation, leading to the development of many U-shaped variants. Weighted Res-UNet[25] uses a weighted attention mechanism to segment small regions, while U-Net++[26] introduces a U-shaped layout with nested dense bypasses to reduce the semantic gap. Dense-UNet uses a densely connected structure to provide optimal separation between intra-output and inter-institution scans.[27] The integration of low- and high-level details is achieved through full-scale skip connections in U-Net3+. ENS-UNet provides a U-shaped architecture with minimal pre- and post-processing requirements,[28] while C-UNet includes inception-like convolutional blocks, recurrent convolutional blocks, and extended convolutional layers to segment skin lesions.[29] Image segmentation tasks often utilize CNN-based techniques that leverage their potent feature extraction capabilities by concentrating on adjacent pixels.[30,31]

An attention-based U-Network, which includes attention mechanisms to improve segmentation accuracy, has been proposed for kidney segmentation in US images.[32] This innovative method is complemented by an efficient kidney segmentation approach that uses an attention-based dual network that efficiently captures contextual information to enable precise segmentation.[23] Additionally, a self-guided attention model adaptable for kidney segmentation using US images has emerged, employing self-supervised learning to enhance network performance.[33] Moreover, a multidimensional attention-guided U-Network tailored for renal tumor segmentation from CT images has been introduced, enhancing segmentation accuracy, particularly for tumor regions.[34] Bidirectional attention-guided U-Net for kidney segmentation from multicontrast CT images has demonstrated the efficacy of bidirectional and attentional mechanisms for accurate segmentation.[35,36] Similarly, a comparable attention module, incorporating two convolutional layers followed by softmax, was integrated into the U-Net hierarchical pooling framework for left atrial segmentation.[37] Recent advancements include the incorporation of additional attention gate modules into the bypass interfaces of the U-Net decoding path, enhancing the model's ability to capture additional information from the encoder.[38,39] Table 1 summarizes previous studies on KUS segmentation using deep-learning approaches. These studies had different goals and specific limitations. Their objectives ranged from applying multitasking CNNs to the precise kidney segmentation of US.

Materials and methods

The main objective of this work was to propose an MLAU-Net model designed for accurate US kidney segmentation of healthy organs to assist the radio-oncologist. This section provides a detailed overview of the proposed segmentation methodology. Figure 1 depicts the framework tailored for KUS segmentation.

KUS dataset

One of the most challenging aspects of deep learning (DL) approaches is developing datasets that require many manually labeled images to train a neural network effectively. The WD-KUS dataset was created via our collaborative research with Guangzhou Medical University and its First Affiliated Hospital. This dataset comprises 44,880 KUS images obtained from patients with a clinical indication for US investigations of their kidneys using TELEMED SmartUs EXT-1 M/3 M. In accordance with privacy measures, all personally identifiable information (PII) was carefully removed during the dataset collection process. The study was conducted in accordance with the Institutional Ethics Committee (IEC) at the First Affiliated Hospital of Guangzhou Medical University.

Deep supervised multi-loss attention 2D U-Net framework (MLAU-Net)

The model uses a single-channel 2D renal US image as input, and the output is a 2D image of identical dimensions depicting the kidney segmentation map. To achieve these segmentation goals, the MLAU-Net approach is proposed with specific modifications, including feature depth enhancement and an attention gate, which enhance focus on important image features

Table 1. A summary of the objectives and limitations of previous studies using deep learning methods for human kidney ultrasound segmentation.

Authors	Architecture	Number of Images	Evaluation Metrics	Limitations of the Method
Shi Yin et al.[2]	Boundary Regression Network	185 US kidney images	• Dice • Mean • Accuracy • IoU	Rely heavily on pre-trained image classification networks like VGG-16; thus, it might not be easily adapted to the subtleties and idiosyncrasies of kidney segmentation in the US images
Deepthy Mary Alex et al.[4]	YSegNet	700 2D US images	• Accuracy • Recall • Precision • Specificity • F1 score • IoU	Dependent on boundary extraction may have problems when segmenting the correct kidney from images with weak boundaries. Weak boundaries are not clearly distinguishable, especially in low contrast or features of noise and speckle potentially bringing uncertain accuracy upon segmentation
Gongping Chen, et al.[40]	Multi-scale inputs pyramid (MSIP)	400 Kidney images	• Accuracy • Dice • Jaccard • Precision • Recall • ASSD	model displays a high performance for kidney segmentation from US images however although it can detect the boundary of the kidney, the boundaries are often blurred and the texture-based model may not work for various heterogeneous structures, not implemented through real-time processing
Peng, et al.[41]	Spider-Net	KiTS2019 300 images	• Dice • PPV (Positive Predictive Value) • Hausdorff95	Complexity of the dual-channel design, Spider-net lies in its computational complexity, as training and inference may require significant computational resources due to the use of multiple attention modules and fusion of CNN and Transformer architectures.
Chen, et al.[18]	Asymmetric U-Shaped Network	300 Kidney US images	• Jaccard • Dice • Accuracy • Recall • Precision • ASSD • AUC	Increased risk of overfitting, especially if the model is trained on a limited dataset, which may limit its performance on diverse or unseen data
Pengceng, et al.[42]	A-PSPNet	1850 annotated ultrasound images	• MIoU • MPA	Modest computational overhead but potential challenges in generalizing to different datasets, with further validation on larger datasets needed. kidney contours exhibit significant variations, it is difficult to accurately segment the renal ultrasound images for A-PSPNet, which may reduce the efficiency of the contrast image for generating the first detection image as well as the identification of ROIs in ultrasonic images.

while reducing noise. In addition, the model includes deep supervision as a regularization technique that embeds objective functions in the decoder layers to address the challenges of small datasets and deeper networks. This approach requires carefully transitioning from deeper blocks to the original segmentation size, ensuring the efficient capture of high-dimensional structures. In Figure 2, MLAU-Net framework for KUS segmentation, emphasizing the integration of attention mechanisms and MLFs. Attention gates selectively focus on relevant features, while various loss functions, including dice and focus loss, optimize training for accurate segmentation.

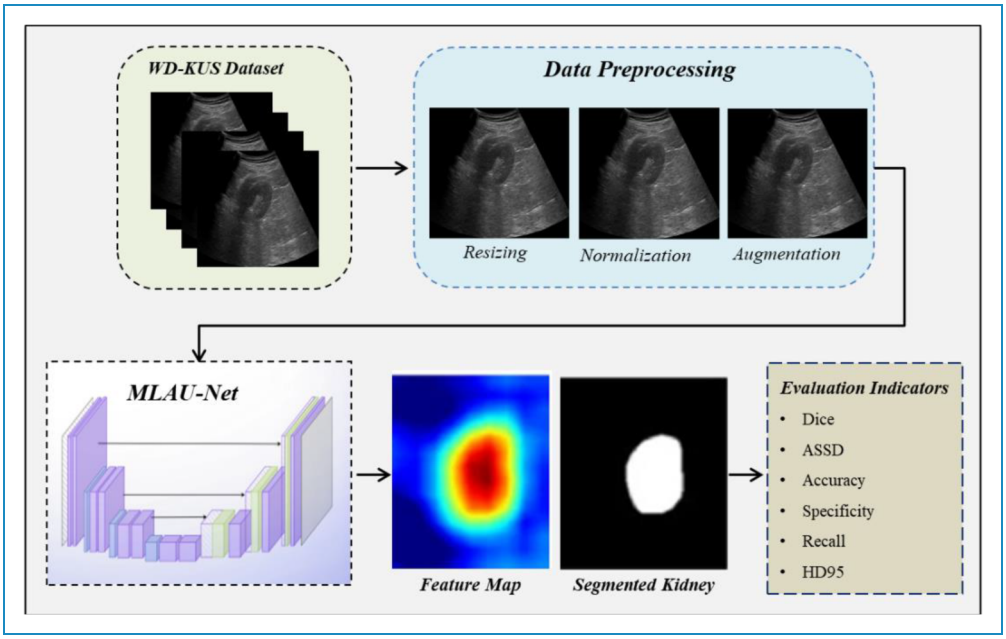

Figure 1. Proposed MLAU-Net pipeline for KUS segmentation.

Depth and the number of initial filters. The model designs seven layers and 16 starting filters instead of the conventional five layers and 64 starting filters. Increasing the depth of the model ensures that the deeper layers capture more complex and abstract information essential for encoding latent details needed for effective feature representation. Concrete concerns about available computational resources recommend using a five-layer model with 16 initial filters for the most efficiency, even though the seven-layer structure could boost performance metrics. Based on the flexible and parametric structure of the model, it is possible to easily modify the number of layers, providing flexibility in balancing computational restrictions with performance. A comprehensive description of the architecture of the proposed MLAU-Net model is given in Table 2. The framework's multilevel attention (MLA) processes enhance its overall performance and capacity to hold intricate features. The essential features, including layer type, filter size, and activation functions, are detailed in each row, corresponding to a specific layer or module in the model. This detailed model summary is invaluable for understanding the proposed MLAU-Net architecture's structural complexity.

Attention mechanism in MLAU-Net. The introduction of attention gates into the U-Net algorithm is done at the level of the concatenation of the skip connections with the up-sampled signal coming from deeper layers in the decoder module. The intricate details of attention gates, elucidating the process, are visualized in Figure 3, which comprises a visual representation of how an attention gate enhances the model's focus on relevant KUS features.[43]

Here, x is the skip connection from the encoder, and g is the decoder feature from the previous block. These two vectors are summed elementwise in the attention unit. In this way, aligned weights become more extensive, while unaligned weights become smaller and, thus, less relevant. The vectors are then computed using the ReLU activation function and a $1 \times 1 \times 1$ convolution followed by a sigmoid layer that ensures all coefficients are in the interval $\alpha_i \in [0, 1]$, so all in all, upon entering a ghost cell, the attention mechanism is determined by the importance of the critical parts of the US image. In the case of image processing using DL, the analogy is that attention helps us to "bring into focus" the informative parts of an image and blank out other artifacts, such as noise, so that their contributions do not enter the final output.[44] The output of attention gates is the element-wise multiplication of input feature maps and attention coefficients:

$$\widehat{x}_{i,c}^{l} = x_{i,c}^{l} \cdot \alpha_i^{l} \tag{1}$$

A single scalar attention value is calculated for every pixel vector in the default configuration. $x_i^l \in \mathbb{R}^{F_l}$. Here, F_l corresponds to the number of feature maps in layer l. Each attention gate is trained to concentrate on a subset of target structures selectively. As labeled in Figure 3, a gating vector $g_i \in \mathbb{R}^{F_g}$ is used for each pixel i to determine focus regions. The formulation of the attention gate is as follows:

$$q_{att}^{l} = \psi^T(\sigma_1(W_x^T x_i^l + W_g^T g_i + b_g)) + b_\psi, \tag{2}$$

$$\alpha_i^l = \sigma_2(q_{att}^l(x_i^l, g_i;\ \Theta_{att})). \tag{3}$$

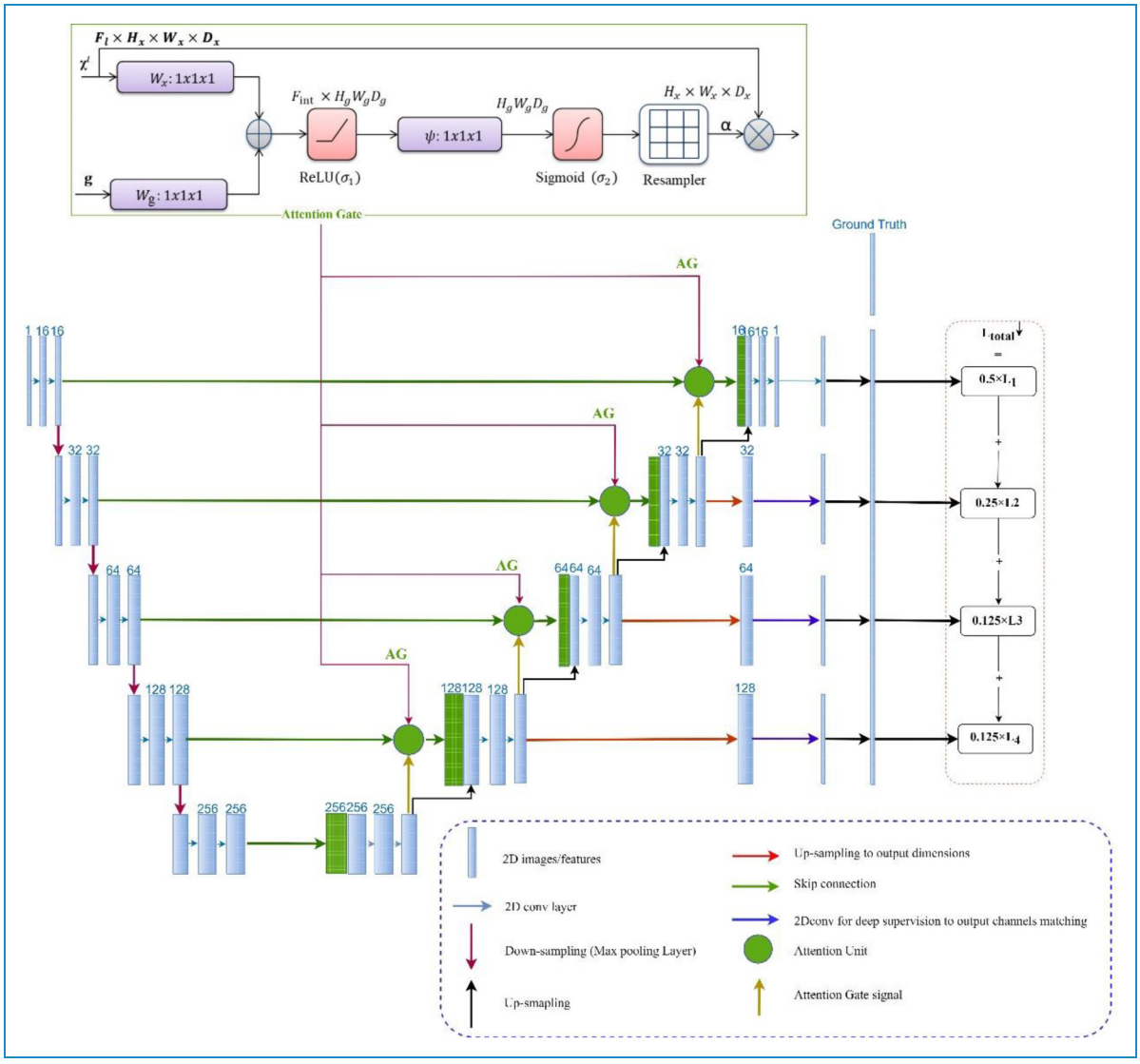

Figure 2. MLAU-Net framework for KUS segmentation incorporated with deep supervision and an attention gate.

Here, σ_2 denotes the sigmoid activation function

$$\sigma_2(x_{i,c}) = \frac{1}{1+\exp(-x_{i,c})}. \tag{4}$$

The attention gate is characterized by a set of parameters Θ_{att} containing, and the linear transformation is

$$W_x \in \mathbb{R}^{F_l \times F_{int}}, \tag{5}$$

$$W_g \in \mathbb{R}^{F_g \times F_{int}}. \tag{6}$$

The linear transformations are calculated through channel-wise $1 \times 1 \times 1$ convolutions applied to the input tensors, where the concatenated features x^l and g are linearly mapped to a $\mathbb{R}^{F_{int}}$ dimensional intermediate space. We observed that the attention gate parameters can be trained using standard back-propagation updates, eliminating the necessity for sampling-based update methods employed in hard-attention.[45]

Deep supervision. Deep supervision involves incorporating companion objective functions into the final three hidden layers in the decoder of the network. The final loss, as illustrated in Figure 5 featuring the proposed MLAU-Net, incorporates dice loss $\mathcal{L}_{\text{dice}}$ and focus loss $\mathcal{L}_{\text{focal}}$. The benefits of deep supervision are twofold. For small training datasets and relatively shallow networks, deep supervision acts as a robust regularization method for training. For somewhat larger datasets, it helps bring deeper networks, avoiding the typical convergence problems associated with such data, such as vanishing or exploding gradients.[46]

Learning rate decay. Learning rate decay is a primarily utilized technique to improve performance in DL models. This involves periodically reducing the learning of the optimizer

Table 2. Summary of the proposed MLAU-Net model.

Layer (Type: Depth-Idx)		Output Shape	Parameter Number
MLAU-Net		[64, 1, 128, 128]	–
ModuleList: 1-1		–	–
	Encoder: 2-1	[64, 16, 128, 128]	–
	DoubleConv: 3-1	[64, 16, 128, 128]	1242
	Encoder: 2-2	[64, 32, 64, 64]	–
	MaxPool2d: 3-2	[64, 16, 64, 64]	–
	DoubleConv: 3-3	[64, 32, 32, 32]	6976
	Encoder: 2-3	[64, 64, 32, 32]	–
	MaxPool2d: 3-4	[64, 32, 32, 32]	–
	DoubleConv: 3-5	[64, 64, 32, 32]	27,776
	Encoder: 2-4	[64, 128, 16, 16]	–
	MaxPool2d: 3-6	[64, 64, 16, 16]	–
	DoubleConv: 3-7	[64, 128, 16, 16]	110,848
	Encoder: 2-5	[64, 256, 8, 8]	–
	MaxPool2d: 3-8	[64, 128, 8, 8]	–
	DoubleConv: 3-9	[64, 256, 8, 8]	442,880
ModuleList: 1-2		–	–
	Decoder_attention: 2-6	[64, 128, 16, 16]	–
	up_conv_block: 3-10	[64, 128, 16, 16]	295,040
	AttentionBlock: 3-11	[64, 128, 16, 16]	16,577
	DoubleConv: 3-12	[64, 128, 16, 16]	443,136
	Decoder_attention: 2-7	[64, 32, 64, 64]	–
	Lup_conv_block: 3-13	[64, 32, 64, 64]	73,792
	AttentionBlock: 3-14	[64, 32, 64, 64]	4193
	DoubleConv: 3-15	[64, 32, 64, 64]	110,976
	Decoder_attention: 2-8	[64, 32, 64, 64]	–
	up_conv_block: 3-16	[64, 32, 64, 64]	18,464

(continued)

Table 2. Continued.

Layer (Type: Depth-Idx)		Output Shape	Parameter Number
	AttentionBlock: 3-17	[64, 32, 64, 64]	1073
	DoubleConv: 3-18	[64, 32, 64, 64]	27,840
	Decoder_attention: 2-9	[64, 16, 128, 128]	–
	up_conv_block: 3-19	[64, 16, 128, 128]	4624
	AttentionBlock: 3-20	[64, 16, 128, 128]	281
	DoubleConv: 3-21	[64, 16, 128, 128]	7008
Conv2d: 1-3		[64, 1, 128, 128]	17
ReLU: 1-4		[64, 1, 128, 128]	–
Total parameters: 1,592,743			
Trainable prams: 1,592,743			
Non-trainable prams: 0			
Total multi-adds (G): 58.02			
Input size (M.B.): 4.19			
Forward/backward pass size (M.B.): 2372.53			

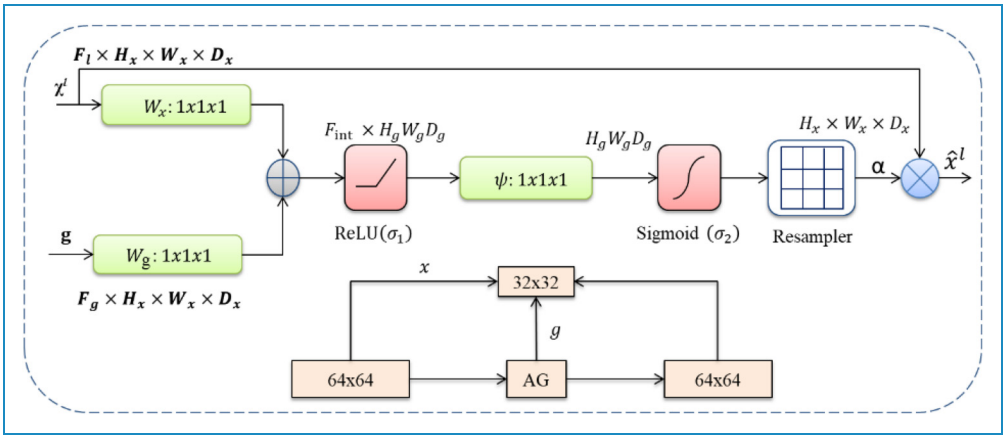

Figure 3. The attention gate works by adjusting the input functions (x^1) using attention factors (α) calculated from the attention gate component. To determine the ratio, we carefully analyze the activations and contextual information conveyed by the input signal (g) on a larger scale. Then, we reorganize the attention coefficients using interpolation. This step-by-step process enables us to precisely refine and focus on regions of space, taking advantage of how input features, attention coefficients, and contextual details interact with each other in the input signal.

when the validation metric plateaus for a certain number of time steps (e.g., 20). This adjustment helps the model approach the actual minimum of the loss function, which can improve overall performance. Implementing a learning rate decay policy is often valuable for optimizing DL frameworks.

The experiment. The input provided to the network is the US 2D image containing the renal US data, and the output will be the 2D kidney segmented image. The following pipeline is created to train and test image classification models on the WD-KUS dataset depending on the different types of data augmentation needed. It also includes using the AdamW optimizer and two loss function training sets with different weights to make the model work appropriately.

Preprocessing

Normalization methods. In the proposed MLAU-Net, we employed min-max normalization to scale the data between 0 and 1, ensuring consistency across all input images. This approach was chosen due to the absence of significant outliers in our US image dataset. Specifically, we applied the following formula to normalize the data:

$$I_{\text{out}} = \frac{I_{\text{in}} - \min(I_{\text{in}})}{\max(I_{\text{in}}) - \min(I_{\text{in}})}. \quad (7)$$

This min-max normalization technique was selected to facilitate convergence during model training and to ensure that the model effectively learns the relationships between input and output.

Resizing. Finally, every image is resized to a standard size of [128, 128] pixels since all the input/output images must have the same size to be fed into the DL model. The values chosen are explained in terms of powers of 2. Since we are implementing a U-Net architecture that gradually down samples the volumes by a factor of 2, using a size that is a power of 2 will avoid size errors in the inner layers of the model. This value is a parameter that should be low enough (limited by the computational resources) to be able to perform the training and high enough to have an amicable spatial resolution. This value can be optimized as a parameter and increased for future finer outputs.

Data augmentation. Regarding data augmentation, the most common techniques for medical images are as follows:

Affine transformations:

- Rotations: Slight rotations in a range $(-10°, 10°)$ will considerably increase the number of cases and make the model more robust.
- Scalings: Slight scaling are applied to the input volume that zooms the patient's body by 10%.

Training for several epochs on a limited dataset requires us to take specific precautions against overfitting. To combat overfitting, we use many different data augmentation methods. Through the training cycle, the following data augmentation methods are applied to the data on the fly: random rotations, random scaling, random skewing, gamma correction augmentation, and mirroring. These methods augment the original training dataset and improve the model on new data that may look different. Annotated data additions are fed into the data loader of the DL model and applied with predefined probabilities. In other words, at each step of each epoch, each data sample is transformed according to defined transformations and probabilities. The training speed is also not compromised because monetary transformations allow us to complete these steps efficiently. Other transformations for data augmentation, such as noise addition, flipping, or

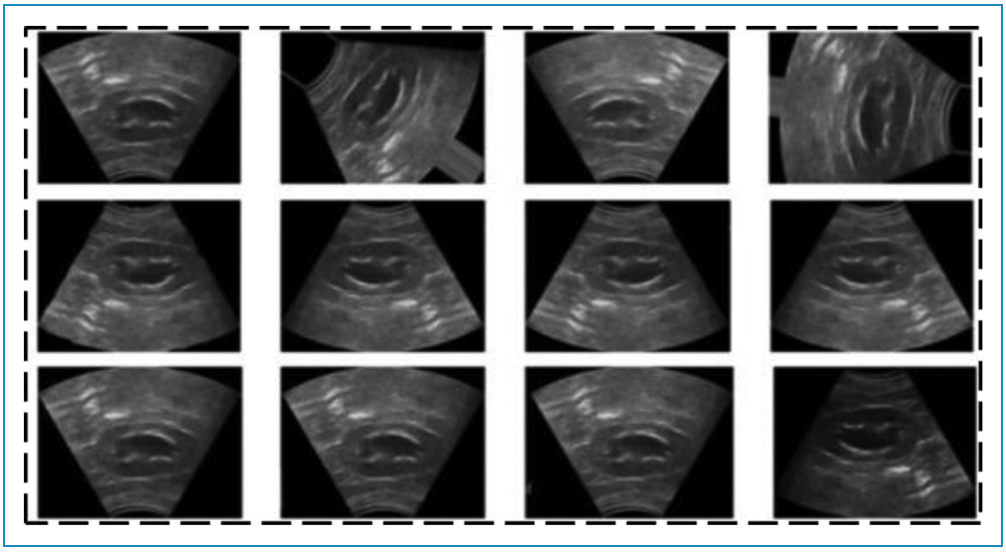

Figure 4. Data augmentation effects on a single image. Various transformations, including random rotations, scaling, skewing, gamma correction, and mirroring, applied during on-the-fly augmentation contribute to the enhanced training dataset.

other distortion transformations, should not be included since, in the medical image domain, it is essential to preserve the original relationships and orientations for the model to learn contextual anatomical information. Figure 4 represents the results of data augmentation on a single image.

Dataset splitting

Wuerzburg-Dynamic Kidney Ultrasound (WD-KUS): A collaboration was established with the First Affiliated Hospital of Guangzhou Medical University to generate the Wuerzburg-Dynamic Kidney Ultrasound (WD-KUS) dataset from patients with a clinical indication requiring US investigation of their kidneys. WD-KUS dataset includes 44,880 images from 176 patients. The data is divided into training, validation, and testing. The training set consists of 33,395 images with ground truth labels from 131 patients. The validation set contains 1357 images with ground truth labels from 13 patients. The test set comprises 10,128 images with ground truth labels from 32 patients. The specifics of this data-splitting process are detailed in Table 3.

K-fold cross-validation is a technique used to assess a model's performance and robustness. It involves splitting the dataset into equally sized folds of *k*. The model is trained on $k-1$ folds and tested on the remaining folds. This process is repeated *k* times, with each fold used as the test set once. The final performance metric is obtained by averaging the results from all *k* iterations.

Implementation and evaluation methods

Loss function and model training. All experiments were executed in Pytorch, and the GPU was an RTX 4090. The DL process uses the WD-KUS and Open Kidney Dataset's[47] dataset for training networks to classify renal US images. The final results of the trained models are tested with an independent test set. In the training phase, training frames are applied to various random data augmentation strategies such as auto-brightness, contrast, gamma, Gaussian blur, Gaussian noise, arbitrary rotation, elastic deformation, random clipping, and scaling. The AdamW optimizer, with 100 epochs, a batch size of 16, and a weight decay of 0.001, is utilized to optimize the networks and reduce overfitting. The final loss used in the model is calculated as shown in Figure 5.

Two loss functions, dice loss L_{dice} and focus loss L_{focal}, with varying weights (W_1 and W_2), are used to train the networks, as formulated in equation (8). The dice loss calculates the overlap between predicted and actual labels and is frequently employed in segmentation tasks. The focus loss is a modified version of the focal loss function that emphasizes difficult-to-classify samples by assigning

Table 3. Detailed overview of dataset splitting for kidney ultrasound (WD-KUD) images. This table outlines the division of patients' kidney ultrasound images into training and test sets, including the number of patients, total images, ground truth annotations, and allocation for training and validation subsets.

Split Name	Number of Patients	Total Images	Ground Truth Available
Training Set	131	33,395	Yes
Validation Set	13	1357	Yes
Test Set	32	10,128	Yes

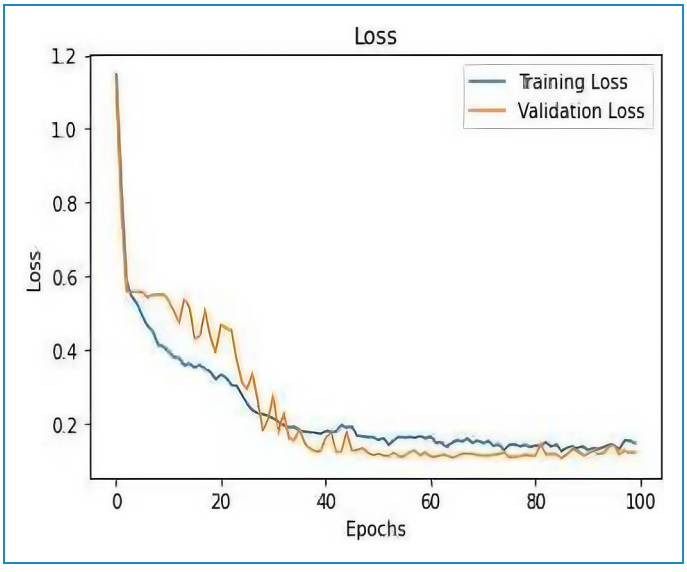

Figure 5. Calculation of the loss in the proposed approach.

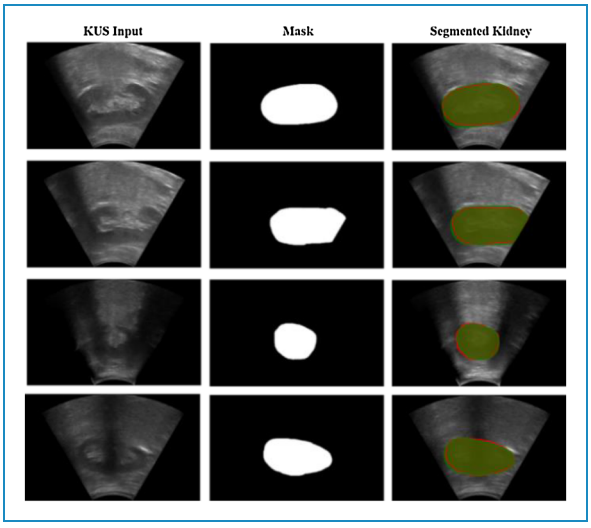

Figure 6. Visual analysis of MLAU-Net results; red denotes the label, and green represents the model's predictions.

them higher weights, thereby directing the network to pay more attention to them during training.

$$\mathcal{L}_{\text{total}} = W_1 \times \mathcal{L}_{\text{dice}} + W_2 \times \mathcal{L}_{\text{focal}} . \tag{8}$$

In equation (8), weights $_1$ and W_2 are pre-determined and set manually to balance the dice loss and focus loss contributions during training. These weights are not learned or trained but are chosen to optimize both loss functions.

The weighted dice loss was selected based on a quantitative evaluation of loss functions. The comprehensive loss function for the network can be expressed as presented in equation (9).

$$\mathcal{L}_{\text{dice}} = -\frac{2}{|P|}\sum_{k\in P}\frac{\sum_{h\in I} v_h^k u_h^k}{\sum_{h\in I} v_h^k + \sum_{h\in I} u_h^k}, \tag{9}$$

where u is a one-hot encoding of the ground truth segmentation map, and v is the network's softmax output. With $h \in I$ indicating the number of pixels in the training patch/batch and $k \in P$ being the classes, m,n have the shape $I \times P$.

Focal loss ($\mathcal{L}_{\text{focal}}$) is derived from the cross-entropy loss and aims to tackle the issue of category imbalance by assigning additional weights to challenging or easily misclassified objects. Examples include backgrounds with noisy textures, partial objects, or the specific objects under focus. The focal loss is defined in equation (10) as follows:

$$\mathcal{L}_{\text{focal}}(p_t) = -(1-p_t)^s \log(b_t), \tag{10}$$

where s is the focusing parameter and b_t is the model's estimation probability for ground truth $y \in \{\pm 1\}$, $b_t = \begin{cases} b & y = 1 \\ 1-b & y = 0 \end{cases}$. Using the focal loss can improve training stability when dealing with a situation with an imbalance in the classes.

We chose various standard evaluation metrics to assess the segmentation performance of our network. These metrics include dice score, average symmetric surface distance (ASSD in mm), accuracy, specificity, recall, and 95th percentile of Hausdorff distance (HD95 in mm). The formulations for these metrics are provided in equations (11)–(18), as detailed in Ref. 48 The metrics used for evaluation are based on true positives (TPs). The segmentation model correctly identifies these pixels as belonging to the target class (e.g., lesion, organ, etc.). True negatives (TNs this term isn't commonly used because the focus is on identifying positive (target) regions rather than classifying the entire image as negative), false positives (FPs), and false negatives (FNs). The metrics determined should range from 0 to 1 or 0 to 100%, and the performance of the proposed model increases when the calculated metrics' values are high.

$$Dice = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \tag{11}$$

$$ASSD = \frac{\sum_{c\in R}\min_{d\in B} \| c-d \| + \sum_{d\in B}\min_{c\in A} \| c-d \|}{N_R + N_S}. \tag{12}$$

R and S represent the boundaries of segmented and reference images, and a and b denote locations on R and S accordingly.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{13}$$

$$\text{Specificity} = \frac{T_N}{T_N + F_P}. \tag{14}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{15}$$

$$HD_{95}(A, B) = \max\left(d_{95}(A, B), d_{95}(B, A)\right), \tag{16}$$

$$d_{95}(A, B) = \max\left(K^{95}\left(\begin{matrix} dis(a, B) \\ a \in A \end{matrix}\right)\right), \tag{17}$$

$$dis(a, B) = \underbrace{\min}_{b\in B} ||a - b||. \tag{18}$$

The distance between a and b is indicated by $\|c - d\|$. The numbers $N_{\text{R.}}$ and $N_{\text{S.}}$ refer to the number of positions on R and S.

Experimental results and discussion. In this study, we adopted six frequently employed metrics to quantitatively compare various methods for KUS image segmentation performance. The six evaluation indicators include accuracy, dice, HD95, recall, specificity, and ASSD. Accuracy, dice, recall, and specificity vary from 0 to 1. A higher score indicates superior method quality, whereas lower ASSD scores correspond to improved segmentation results for the method. Figure 6 shows results in images

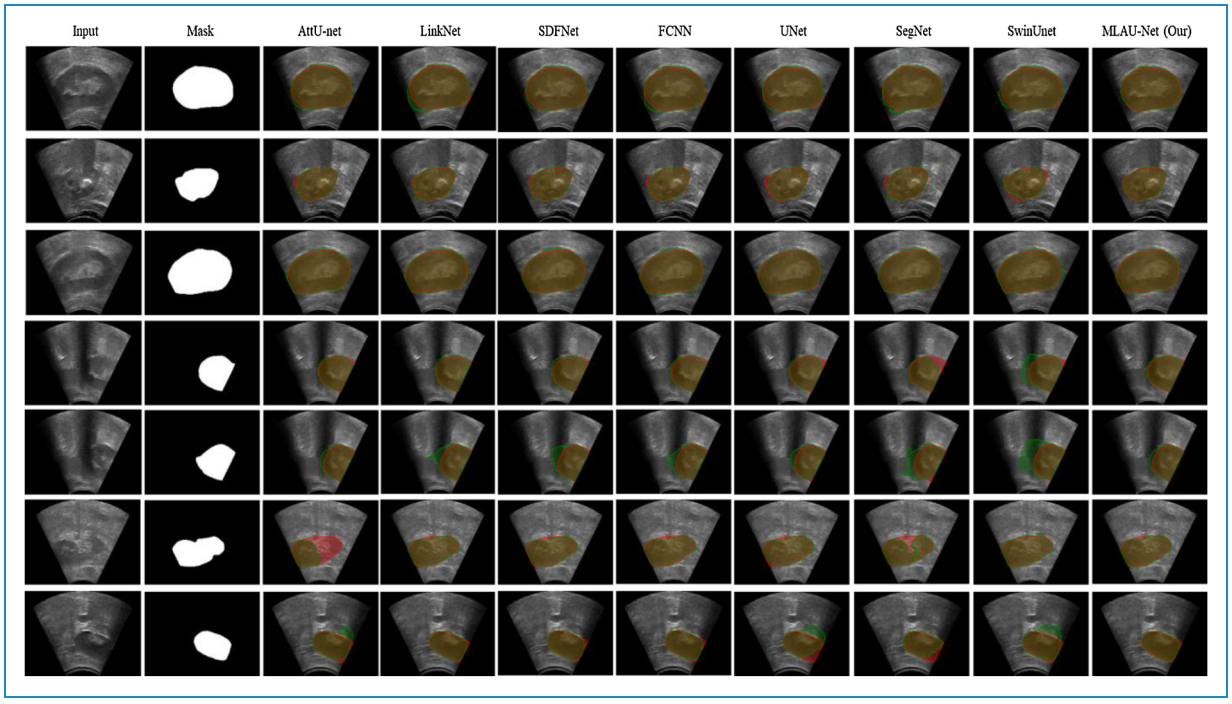

Figure 7. Qualitative analysis of MLAU-Net and different baseline methods on the WD-KUD dataset. Red indicates the label, and green represents the model's predictions. Column 1 displays the input image, Column 2 shows the mask, Columns 3–9 exhibit the outputs of baseline models, and the final column depicts the prediction of the proposed model.

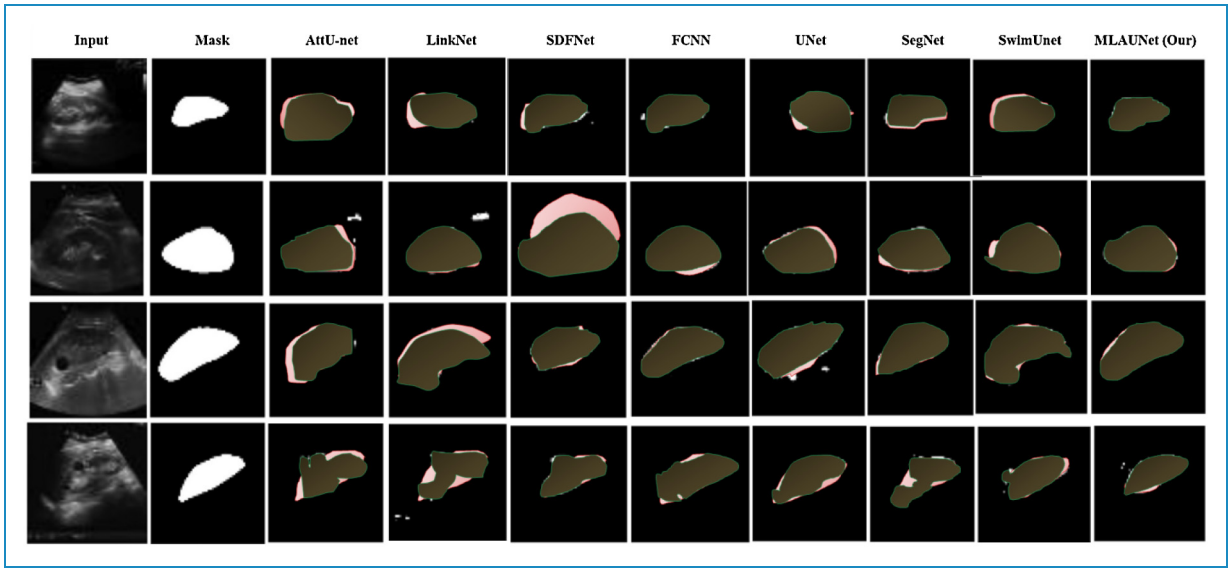

Figure 8. The open kidney dataset was used to perform a qualitative analysis of MLAU-Net and various baseline methods. This figure is represented by the red color, which symbolizes the label, and the green color denotes the model's predictions. Column 1 depicts the input image, column 2 shows the ground truth mask, and columns 3–9 illustrate results of different baseline models. The last one gives an output for MLAU-Net predicted thus.

featuring both the input images and the segmented output masks obtained using the proposed MLAU-Net framework. In this representation, the red color indicates the label, while the green color shows the prediction. Figure 6 show that the images the proposed framework predicted closely resemble the original mask.

Table 4. Quantitative analysis comparing MLAU-Net with seven state-of-the-art segmentation approaches. A comprehensive assessment of performance metrics was conducted on our collected WD-KUS dataset in a consistent experimental environment. The table showcases the results of ablation experiments, highlighting the efficacy of MLAU-Net against established segmentation methods.

Models	Dice (%) Mean ± SD	Accuracy (%) Mean ± SD	Specificity (%) Mean ± SD	HD95 (mm) Mean ± SD	Recall (%) Mean ± SD	ASSD (mm) Mean ± SD
AttU-Net[43]	88.62 ± 0.17	97.18 ± 0.04	98.91 ± 0.09	9.66 ± 0.28	90.62 ± 1.07	2.98 ± 0.08
LinkNet[49]	88.35 ± 0.60	97.2 ± 0.04	98.85 ± 0.01	9.82 ± 0.57	91.13 ± 0.63	3.02 ± 0.19
SDFNet[50]	87.16 ± 0.25	97.07 ± 0.05	98.86 ± 0.05	9.89 ± 0.42	89.99 ± 0.48	3.10 ± 0.09
FCNN[51]	88.79 ± 0.79	96.02 ± 0.16	98.91 ± 0.14	10.54 ± 0.84	88.92 ± 1.88	3.19 ± 0.25
U-Net[13]	86.02 ± 0.75	95.73 ± 0.11	98.56 ± 0.06	15.96 ± 1.10	88.65 ± 0.73	4.18 ± 0.20
SegNet[17]	86.47 ± 0.58	97.00 ± 0.07	98.78 ± 0.05	10.40 ± 0.35	89.96 ± 0.28	3.23 ± 0.08
SwinUnet[52]	85.22 ± 0.34	96.20 ± 0.05	98.44 ± 0.13	13.90 ± 0.28	85.08 ± 1.23	4.39 ± 0.03
MLAU-Net (our)	90.21 ± 0.62	98.26 ± 0.11	98.93 ± 0.05	8.90 ± 0.15	91.78 ± 1.03	2.87 ± 0.05

Qualitative analysis

In particular, Figures 7 and 8 present a qualitative analysis of the WD-KUS dataset and Open Kidney Dataset[47] using the proposed MLAU-Net model, contrasting it with seven other state-of-the-art approaches. For visual comprehension, segmented maps of sample images from the test dataset are depicted for both the proposed models and the competing models in Figure 7.

The suggested framework consistently outperforms the competing models, even in scenarios involving missing boundaries and shape variability. The boundary of the target area appears indistinct in U-Net and its variants. In contrast, the proposed approach demonstrates resilience to noise and other factors in US images, resulting in segmentations that closely align with the mask.

Quantitative analysis

Numerous kidney segmentation techniques have recently been developed and implemented across various studies. To assess the performance of these methods in comparison to our proposed approach, we conducted a thorough quantitative analysis using metrics such as dice, specificity, accuracy, HD95, recall, and ASSD.[48] The results of this comprehensive evaluation are presented in Table 4, which offers a quantitative comparative analysis of MLAU-Net and seven state-of-the-art segmentation approaches. All the ablation experiments were conducted in a consistent environment using our collected WD-KUS dataset, and in Table 5, a comprehensive assessment of performance metrics was conducted on an Open Kidney Dataset in a consistent experimental environment. The table showcases the results of ablation experiments, highlighting the efficacy of MLAU-Net against established segmentation methods.

For quantitative analysis, comparative experiments were conducted with seven widely employed segmentation methods, specifically AttU-Net,[43] Seg-Net,[17] FCNN,[51] U-Net,[13] SDFNet,[50] SwinUnet,[52] and LinkNet.[49] To ensure a fair and unbiased comparison, each method underwent complete training, and its segmentation results were not subjected to any post-processing. The segmentation accuracy of U-Net, SegNet, SwinNet, and SDFNet does not match that of FCNN, but their results exhibit superior visual effects. Conversely, FCNN, Linknet, and AttUnet display less favorable visual outcomes, characterized by noticeably jagged contours, suggesting a weakness in the learning ability of these methods, particularly along the kidney's edge. Two conjectures arise: the methods struggle to extract finer kidney features comprehensively, and substantial loss of kidney information occurs during deconvolution. Through a comprehensive analysis of evaluation matrices, significance tests, and segmentation results across various networks, our proposed MLAU-Net demonstrates a significant competitive advantage. Notably, it reduces false and missed detection rates on the WD-KUS and datasets. Computational performance trade-offs are illustrated in Figure 9 by comparing accuracy against the number of parameters for various baseline models and the suggested MLAU-Net on the WD-KUS dataset.

To enhance the assessment of segmentation methods on KUS images, we generated curves for accuracy, dice, recall,

Table 5. Quantitative study of MLAU-Net against SOTA techniques using an open kidney dataset.

Models	Dice (%) Mean ± SD	Accuracy (%) Mean ± SD	Specificity (%) Mean ± SD	HD95 (mm) Mean ± SD	Recall (%) Mean ± SD	ASSD (mm) Mean ± SD
AttU-Net[43]	92.22 ± 0.17	98.20 ± 0.03	98.11 ± 0.08	9.62 ± 0.31	91.72 ± 0.09	2.88 ± 0.07
LinkNet[49]	87.93 ± 0.38	96.2 ± 0.06	98.85 ± 0.01	9.82 ± 0.57	92.39 ± 0.43	2.09 ± 0.20
SDFNet[50]	89.21 ± 0.27	96.03 ± 0.07	98.76 ± 0.10	11.39 ± 0.12	90.63 ± 0.51	3.21 ± 0.17
FCNN[51]	90.39 ± 0.69	97.12 ± 0.11	98.21 ± 0.17	13.24 ± 0.64	90.12 ± 1.82	3.43 ± 0.27
U-Net[13]	91.12 ± 0.74	95.77 ± 0.17	98.16 ± 0.06	14.91 ± 1.12	89.45 ± 0.79	4.68 ± 0.21
SegNet[17]	91.47 ± 0.54	97.17 ± 0.10	98.12 ± 0.06	11.40 ± 0.41	91.92 ± 0.34	4.13 ± 0.09
SwinUnet[52]	89.31 ± 0.36	98.11 ± 0.07	98.57 ± 0.11	12.88 ± 0.31	89.09 ± 1.21	3.36 ± 0.05
MLAU-Net (our)	93.43 ± 0.59	98.31 ± 0.10	98.96 ± 0.06	8.92 ± 0.16	93.81 ± 1.05	2.91 ± 0.06

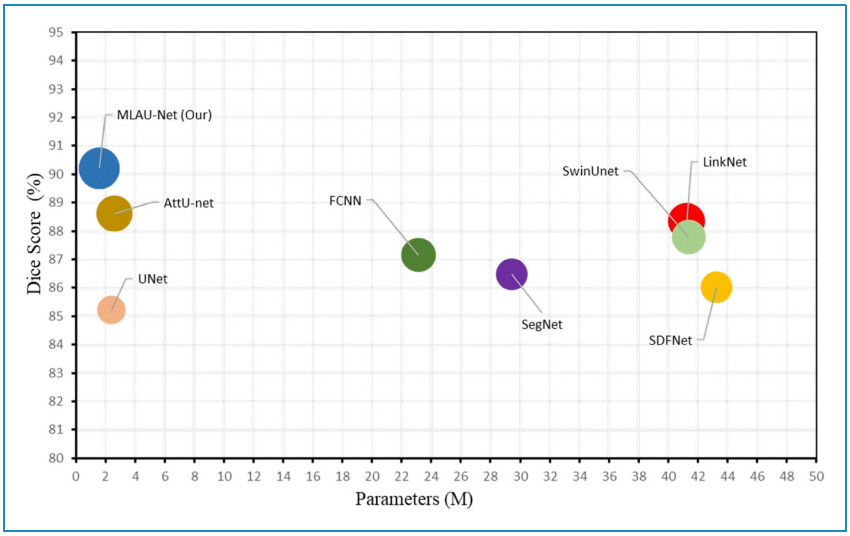

Figure 9. The proposed model effectively balances performance efficiency and computational cost, showcasing the lowest number of parameters among all baseline approaches. Presetting these parameters enables users to tailor computational resources and choose the appropriate encoder–decoder for feature extraction, depending on their specific requirements in MIS.

and specificity, as depicted in Figure 10. The visual representation emphasizes the superior performance of our approach to others, underscoring its aptness for WD-KUS image segmentation.

In addition to assessing traditional metrics, we conducted a comprehensive performance comparison between different models using HD95 and ASSD, as shown in Figure 11. The evaluated networks included AttU-Net, LinkNet, SDFNet, FCNN, U-Net, SegNet, SwinUnet, and the proposed MLAU-Net. Significantly, the suggested framework outperformed all others across HD95 and ASSD metrics, affirming its exceptional segmentation accuracy and dice for both datsets.

Through qualitative and quantitative analyses of our proposed framework, we demonstrate the efficacy of each designed component. Comparisons with state-of-the-art segmentation methodologies reveal that our suggested network consistently outperforms competitors across six widely used evaluation indicators for two different datasets, as displayed in Figure 12.

Despite instances of false and missed detection in the segmentation outcomes, our method demonstrates impressive performance compared to alternative approaches. Our

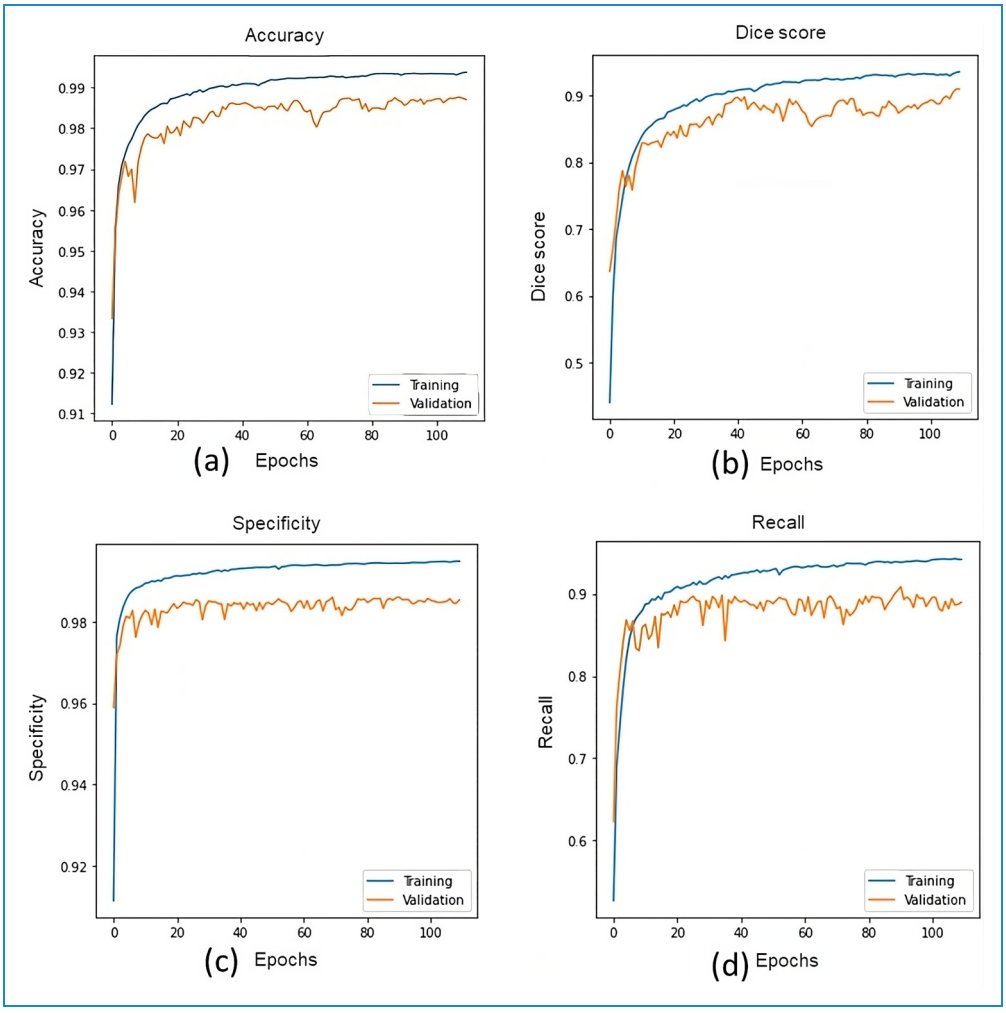

Figure 10. Proposed approach performance evaluation curves for accuracy, dice, recall, and specificity.

method exhibits enhanced robustness to WD-KUS images, showing resilience against various influencing factors. In conclusion, the segmentation approach presented in this study effectively addresses the challenges associated with the automatic segmentation of the WD-KUS dataset, marking a substantial advancement in this domain.

Ablation experiments

In order to assess the effect of each component in the proposed MLAU-Net framework, we performed a detailed ablation study on three major modules: hybrid loss, deep supervision and attention mechanism. K-fold cross-validation with $K = 5$ was used for carrying out the study to ensure its robustness and generalization. In our initial experiment, hybrid loss, which consists of Dice loss as well as cross-entropy loss, was removed, consequently resulting in marked drops in Dice score and accuracy, which indicated that it played a vital role.

In the second experiment, we eliminated deep supervision, leaving us with lower performance measurements that highlighted its contribution to reducing FNs and improving the segmentation accuracy. The last one turned off an attention mechanism, which brought about evident worsening in HD95, especially ASSD, showing how it enhanced boundary delineation and reduced segmentation errors. Table 6 presents average figures (mean $\pm$ SD) for some essential performance metrics obtained via the k-fold cross-validation method applied in ablation studies.

Computational complexity analysis

The comparative analysis proposed study compared the computational complexity levels for MLAU-Net with other state-of-art models in terms of parameters and floating-point operations per second. This comparison is important as it determines how well a model performs in relation to its computational efficiency. The segmentation models that we evaluated were AttU-Net, Seg-Net, FCNN, U-net, SDFNet, SwinUnet, and LinkNet. We trained all these methods from scratch and fairly assessed

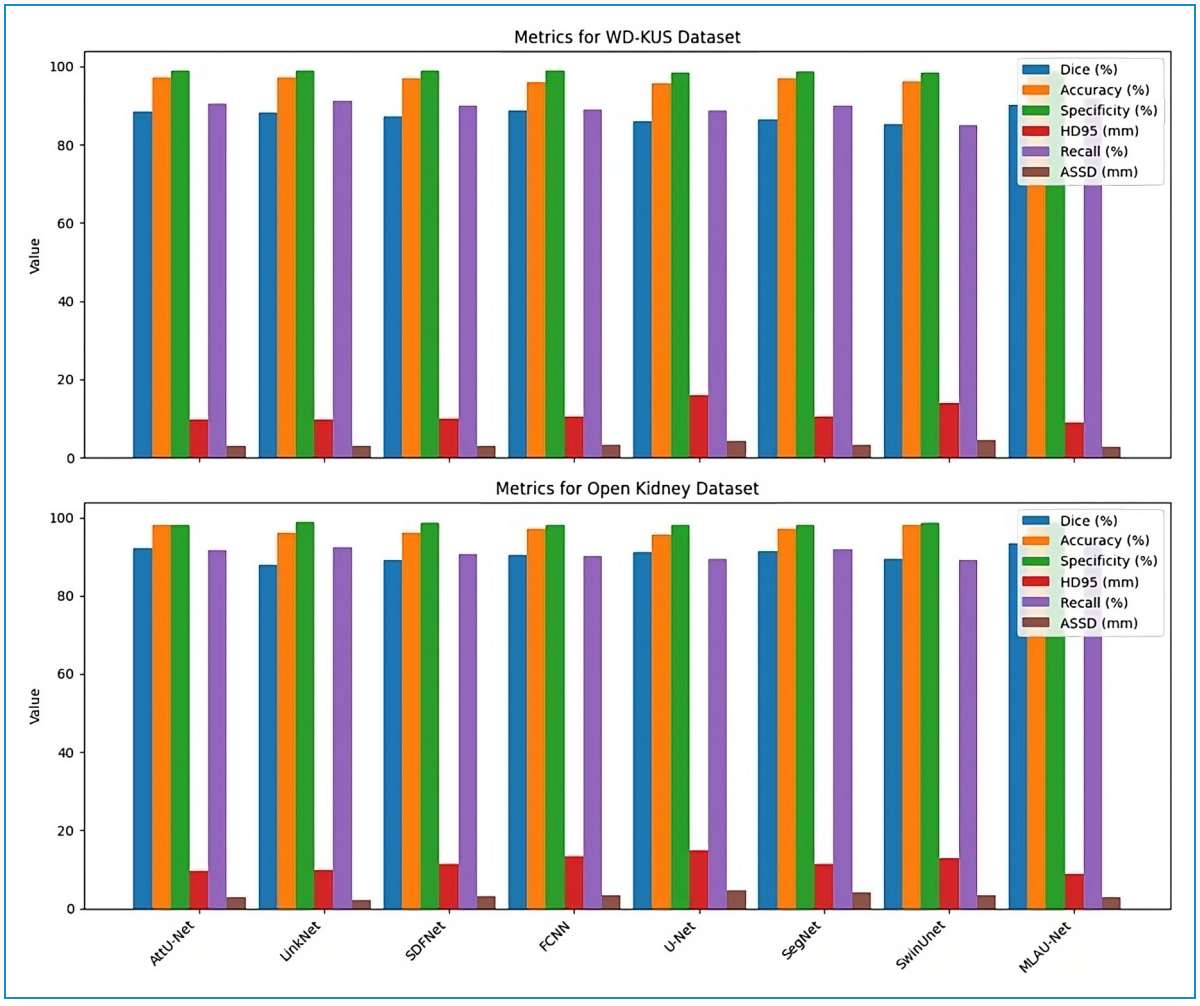

Figure 11. Quantitative analysis of models using segmentation metrics for different SOTA models.

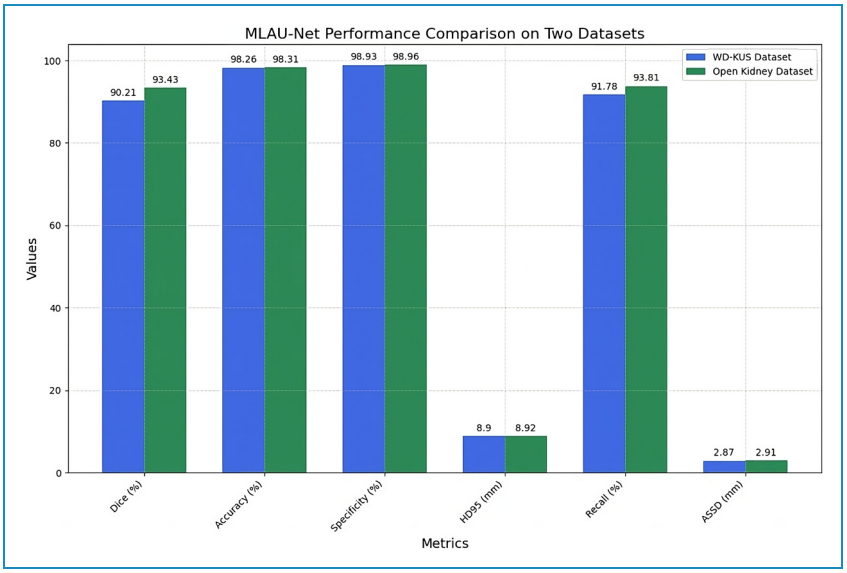

Figure 12. Performance comparison of MLAU-Net across two datasets.

Table 6. Ablation study of MLAU-Net with K-fold cross-validation results for different modules.

Model Variation	Dice (%) Mean ± SD	Accuracy (%) Mean ± SD	Specificity (%) Mean ± SD	HD95 (mm) Mean ± SD	Recall (%) Mean ± SD	ASSD (mm) Mean ± SD
MLAU-Net	90.21 ± 0.62	98.26 ± 0.11	98.93 ± 0.05	8.90 ± 0.15	91.78 ± 1.03	2.87 ± 0.05
W/o Hybrid Loss	87.35 ± 0.58	97.85 ± 0.14	98.62 ± 0.07	10.25 ± 0.22	89.64 ± 1.25	3.14 ± 0.07
W/o Deep Supervision	88.47 ± 0.65	97.98 ± 0.13	98.71 ± 0.06	9.85 ± 0.18	90.35 ± 1.20	3.05 ± 0.06
W/o Attention	86.22 ± 0.72	97.12 ± 0.09	98.48 ± 0.09	12.30 ± 0.25	87.50 ± 1.15	3.50 ± 0.08

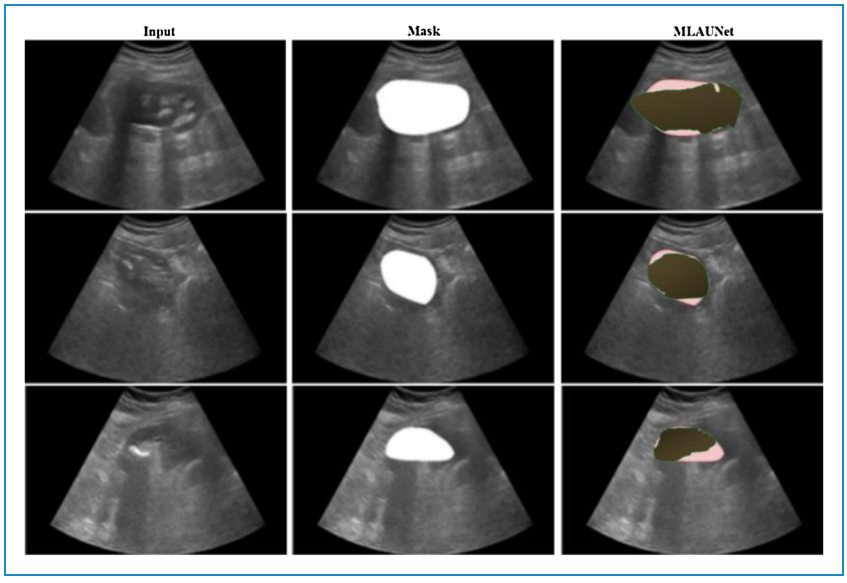

Figure 13. Segmentation challenges in different scenarios: (Row 1) Stone shadow obscuring the target, (Row 2) Unclear shapes caused by distortions or artifacts, and (Row 3) Irregular kidney shape due to anatomical variations

their segmentations without post-processing. As shown in Figures 7–10 and Tables 4–5, our results demonstrate that MLAU-Net achieves better segmentation performances with less parameter numbers and lower computational budget. Table 4 highlights that MLAU-Net outperforms U-Net and FCNN on the WD-KUS dataset, with a Dice Score of 90.21%, while maintaining fewer parameters (Column 3) than other models. Similarly, Table 5 presents the results on the Open Kidney Dataset, where MLAU-Net also achieves a Dice Score of 93.43%, surpassing state-of-the-art models in accuracy and computational efficiency. This proves the effectiveness of MLAU-Net, significantly reducing false alarms and missed segmentations on both the WD-KUS dataset and public datasets. Quantitative comparisons are made in Table 4 (WD-KUS) and Table 5 (Open Kidney Dataset), while Figure 7 provides qualitative analysis for the WD-KUS dataset and Figure 8 for the Open Kidney Dataset. The performance evaluation curves in Figure 9 further confirm MLAU-Net's ability to balance accuracy and computational cost, reinforcing its robustness in segmentation tasks across multiple datasets.

Analysis of segmentation challenges in MLAU-Net

While MLAU-Net achieves overall significant performance, here are certain cases when the situation is different. This includes images with very low signal-to-noise ratio and those that differ significantly from what was used for training. In such situations, the attention mechanism that is responsible for highlighting crucial details might find it difficult to distinguish between noise and subtle renal contours. Furthermore, the deep supervision part may not completely overcome poor contrast, thereby making it difficult to draw exact boundaries. In future research, more advanced techniques of denoising or adversarial training can be studied in order to make handling of such images more robust.

Dealing with small or nonuniform kidneys is another problem. The model could encourage larger kidneys over others because of its inherent class imbalance in the data.

Research into more advanced data augmentation methods and loss functions that are class-weighted might help improve the ability of neural networks to interpret kidney shapes and sizes.

Moreover, anatomical variations or artifacts like renal cysts, stones, or US beam distortions may cause uncertainties, leading to less effective segmentation results. The proposed approach often struggles to distinguish the target from the background in these cases accurately. For example, stone shadows may obscure important structures, distortions or artifacts can make shapes unclear, and irregular anatomical structures complicate the segmentation process. Addressing these issues may require incorporating domain-specific knowledge and multi-modal information. In Figure 13, we visualize some challenging cases where MLAU-Net fails to accurately separate the target from the background due to these factors.

Conclusion

Automatic segmentation of human KUS images is essential in helping urologists diagnose and treat kidney diseases in clinical practice. Nevertheless, factors such as the image quality, kidney morphology, and heterogeneous structures present challenges for accurate and automatic segmentation. In this study, we introduced MLAU-Net, a novel framework that leverages well-controlled attention mechanisms and a hybrid loss strategy to enhance the segmentation of low-resolution renal US images. Key components of MLAU-Net, including attention gates, deep supervision, and a meticulous preprocessing pipeline, significantly improve existing methods. Our results demonstrate that MLAU-Net excels in producing accurate segmentations thanks to its advanced features. Including attention, gates ensure that the model focuses on critical regions, while deep supervision aids in refining segmentation outputs. This makes MLAU-Net a valuable tool for precise medical image analysis and diagnosis, addressing the inherent challenges of low-resolution renal US imaging. In addition, through performing more evaluation on MLAU-Net using Open Kidney Dataset which is an open access dataset used to further confirm its ability. According to results, regardless of various metrics such as dice coefficient, accuracy, specificity, HD95, recall or ASSD this technique outperforms all modern techniques. This extensive evaluation underscores the robustness and generalizability of MLAU-Net in different clinical scenarios. The proposed methodology, incorporating domain information integration and weighted feature fusion, yielded superior results, particularly for segmenting malignant masses. MLAU-Net demonstrates high accuracy and efficiency in KUS image processing and holds promise for broader applications in other MIS tasks. In the future, we aim to refine MLAU-Net by incorporating specific pig KUS data and CT scan results as more data becomes available. This continuous improvement will increase its applicability and effectiveness, overcoming difficulties of low signal-to-noise ratios and weakly contrasted boundaries between objects. Developing MLAU-Net further presently aims at producing an all-encompassing solution towards MIS thereby facilitating improved clinical practice diagnosis and treatment planning.

List of abbreviations

MLAU-Net	Multiloss attention U-Net
DL	Deep learning
KUS	Kidney ultrasound
PII	Personally identifiable information
MRI	Magnetic resonance imaging
IoU	Intersection over union
TELEMED	Telemedicine
LSTM	Long short-term memory
CNN	Convolutional neural network
SVM	Support vector machine
GLCM	Gray-level co-occurrence matrix
SIFT	Scale-invariant feature transform
CT	Computed tomography
ROI	Region of interest
GPU	Graphics processing unit
WD-KUS	Wuerzburg-dynamic kidney ultrasound
ASSD	Average symmetric surface distance
HD95	95th percentile of Hausdorff distance
SGD	Stochastic gradient descent
MSE	Mean squared error

Acknowledgment: This study was supported by the Guangzhou Science and Technology Project (202201020535), Guangzhou Medical University (2024SRP077), the National Natural Science Foundation of China (82100805), Guangzhou Science and Technology Planning Project (202102021129), and the Educational Commission of Guangdong Province (2022ZDJS113). The authors also acknowledge the support from Wuerzburg Dynamics Inc. for the Weiding Joint Laboratory of Medical Artificial Intelligence at Shenzhen Technology University.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

The authors state that they have no known competing financial interests or close personal ties that could have influenced the research work presented in this paper.

Ethical approval: This study was conducted in accordance with the ethical guidelines and approval of the Institutional Ethics Committee (IEC) at the First Affiliated Hospital of Guangzhou Medical University (No. ES-2024–046-02).

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Natural Science Foundation of China, The Educational Commission of Guangdong Province of China, the School-Enterprise Cooperation Fund provided

by Wuerzburg Dynamics Inc. to the Weiding Joint Laboratory of Medical Artificial Intelligence, Shenzhen Technology University, Guangzhou Science and Technology Project, Guangzhou Science and Technology Planning Project, Guangzhou Medical University Scientific Research Capacity Improvement Program (grant number: 82100805, 2022ZDJS113, 202201020535, 202102021129, and 2024SRP077).

Guarantor: Rashid Khan and Bingding Huang.

Informed consent: All study participants provided written informed consent or had their legally authorized representatives do so prior to the initiation of the study. The study design and informed consent procedures were reviewed and approved by the Institutional Ethics Committee at the First Affiliated Hospital of Guangzhou Medical University.

ORCID iD: Rashid Khan https://orcid.org/0000-0002-2410-044X

Data availability: Data will be made available upon request from the corresponding author. Code: https://github.com/qumais/MLAUNet-for-Kidney-Ultrasound-Segmentation.git

References

1. Jiang H, Diao Z, Shi T, et al. A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation. *Comput Biol Med* 2023; 157: 106726.
2. Yin S, Peng Q, Li H, et al. Automatic kidney segmentation in ultrasound images using subsequent boundary distance regression and pixelwise classification networks. *Med Image Anal* 2020; 60: 101602.
3. Noble JA and Boukerroui D. Ultrasound image segmentation: a survey. *IEEE Trans Med Imaging* 2006; 25: 987–1010.
4. Alex DM, Abraham Chandy D, Hepzibah Christinal A, et al. YSegnet: a novel deep learning network for kidney segmentation in 2D ultrasound images. *Neural Comput Appl* 2022; 34: 22405–22416.
5. Yu H, Yang LT, Zhang Q, et al. Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives. *Neurocomputing* 2021; 444: 92–110.
6. Hesamian MH, Jia W, He X, et al. Deep learning techniques for medical image segmentation: achievements and challenges. *J Digit Imaging* 2019; 32: 582–596.
7. Guo Z, Li X, Huang H, et al. Deep learning-based image segmentation on multimodal medical imaging. *IEEE Trans Radiat Plasma Med Sci* 2019; 3: 162–169.
8. Yang X, Le Minh H, Cheng K-TT, et al. Renal compartment segmentation in DCE-MRI images. *Med Image Anal* 2016; 32: 269–280.
9. Levienaise-Obadia B and Gee A. Adaptive segmentation of ultrasound images. *Image Vis Comput* 1999; 17: 583–588.
10. Xie J, Jiang Y and Tsui H-t. Segmentation of kidney from ultrasound images based on texture and shape priors. *IEEE Trans Med Imaging* 2005; 24: 45–57.
11. Shim H, Chang S, Tao C, et al. Semiautomated segmentation of kidney from high-resolution multidetector computed tomography images using a graph-cuts technique. *J Comput Assist Tomogr* 2009; 33: 893–901.
12. De Jesus-Rodriguez HJ, Morgan MA and Sagreiya H. Deep learning in kidney ultrasound: overview, frontiers, and challenges. *Adv Chronic Kidney Dis* 2021; 28: 262–269.
13. Ronneberger O, Fischer P and Brox T. U-Net: convolutional networks for biomedical image segmentation. In: *Proceedings of medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Part III 18*, Munich, Germany, October 5–9, 2015, Springer, 2015, pp. 234–241.
14. Lin T-Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, Hawaii, USA, July 21–27, 2017, pp. 2117–2125.
15. Feng S, Zhao H, Shi F, et al. CPFNet: context pyramid fusion network for medical image segmentation. *IEEE Trans Med Imaging* 2020; 39: 3008–3018.
16. Chen L-C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation, arXiv preprint arXiv:1706.05587, 2017.
17. Badrinarayanan V, Kendall A and Cipolla R. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell* 2017; 39: 2481–2495.
18. Chen G-P, Zhao Y, Dai Y, et al. Asymmetric U-shaped network with hybrid attention mechanism for kidney ultrasound images segmentation. *Expert Syst Appl* 2023; 212: 118847.
19. Yu C, Li S, Ghista D, et al. Multi-level multi-type self-generated knowledge fusion for cardiac ultrasound segmentation. *Inf Fusion* 2023; 92: 1–12.
20. Chen B, Liu Y, Zhang Z, et al. TransAttUnet: multi-level attention-guided U-net with transformer for medical image segmentation. *IEEE Trans. Emerg. Top. Comput. Intell* 2023; 8: 55–68.
21. El-Taraboulsi J, Cabrera CP, Roney C, et al. Deep neural network architectures for cardiac image segmentation. *Artif Intell Life Sci* 2023; 4: 100083.
22. Sinha A and Dolz J. Multiscale self-guided attention for medical image segmentation. *IEEE J Biomed Health Inform* 2020; 25: 121–130.
23. Wu J, Zhou S, Zuo S, et al. U-Net combined with multiscale attention mechanism for liver segmentation in C.T. Images. *BMC Med Inform Decis Mak* 2021; 21: 1–12.
24. Valente S, Morais P, Torres HR, et al. A deep learning method for kidney segmentation in 2D ultrasound images. In: *2022 44th annual international conference of the IEEE engineering in medicine & biology society (EMBC)*, IEEE, Glasgow, Scotland, UK, July 11–15, 2022, pp. 3911–3914.
25. Xiao X, Lian S, Luo Z, et al. Weighted res-unet for high-quality retina vessel segmentation. In: 2018 9th international conference on information technology in medicine and education (ITME), IEEE, 2018, pp. 327–331.
26. Al Suman A, Sarda S, Asikuzzaman M, et al. Two-stage u-net ++ for medical image segmentation. In: *2021 digital image computing: techniques and applications (DICTA)*, IEEE, 2021, pp. 01–06.

27. Li X, Chen H, Qi X, et al. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from C.T. Volumes. *IEEE Trans Med Imaging* 2018; 37: 2663–2674.
28. Meng Z, Fan Z, Zhao Z, et al. ENS-Unet: end-to-end noise suppression U-Net for brain tumor segmentation. In: *2018 40th annual international conference of the ieee engineering in medicine and biology society (EMBC)*, IEEE, Honolulu, Hawaii, USA, July 18–21, 2018, pp. 5886–5889.
29. Wu J, Chen EZ, Rong R, et al. Skin lesion segmentation with C-UNet. In: *2019 41st Annual International Conference of the IEEE engineering in medicine and biology society (EMBC)*, IEEE, Berlin, Germany, July 23-27, 2019, pp. 2785–2788.
30. Ghosh S, Das N, Das I, et al. Understanding deep learning techniques for image segmentation. *ACM Comput Surv (CSUR)* 2019; 52: 1–35.
31. Wang R, Zhou H, Fu P, et al. A multiscale attentional unet model for automatic segmentation in medical ultrasound images. *Ultrason Imaging* 2023; 45: 159–174.
32. Chen G, Li L, Dai Y, et al. AAU-net: an adaptive attention U-net for breast lesions segmentation in ultrasound images. *IEEE Trans Med Imaging* 2022; 42: 1289–1300.
33. Yu M, Han M, Li X, et al. Adaptive soft erasure with edge self-attention for weakly supervised semantic segmentation: thyroid ultrasound image case study. *Comput Biol Med* 2022; 144: 105347.
34. Zhang Q, Liang Y, Zhang Y, et al. A comparative study of attention mechanism based deep learning methods for bladder tumor segmentation. *Int J Med Inf* 2023; 171: 104984.
35. Peng J and Wang Y. Medical image segmentation with limited supervision: a review of deep network models. *IEEE Access* 2021; 9: 36827–36851.
36. Vakanski A, Xian M and Freer PE. Attention-enriched deep learning model for breast tumor segmentation in ultrasound images. *Ultrasound Med Biol* 2020; 46: 2819–2833.
37. He A, Wang K, Li T, et al. H2Former: an efficient hierarchical hybrid transformer for medical image segmentation. *IEEE Trans Med Imaging* 2023; 42: 2763–2775.
38. Xu Z, Tian B, Liu S, et al. Collaborative attention guided multiscale feature fusion network for medical image segmentation. *IEEE Trans Netw Sci Eng* 2023; 11: 1–15.
39. Xie L, Cai W and Gao Y. Dmcgnet: a novel network for medical image segmentation with dense self-mimic and channel grouping mechanism. *IEEE J Biomed Health Inform* 2022; 26: 5013–5024.
40. Chen G, Yin J, Dai Y, et al. A novel convolutional neural network for kidney ultrasound images segmentation. *Comput Methods Programs Biomed* 2022; 218: 106712.
41. Peng Y, Hu X, Hao X, et al. Spider-Net: High-resolution multiscale attention network with full-attention decoder for tumor segmentation in kidney, liver and pancreas. *Liver Pancreas* 2024; 93: 106163.
42. Wen P, Guan Y, Li J, et al. A-PSPNet: a novel segmentation method of renal ultrasound image. In: *2021 IEEE international conference on systems, man, and cybernetics (SMC),* IEEE, Melbourne, Australia, October 17–20, 2021, pp. 40–45.
43. Oktay O, Schlemper J, Folgoc LL, et al. Attention u-net: Learning where to look for the pancreas, arXiv preprint arXiv:1804.03999, 2018.
44. Mnih V, Heess N and Graves A. Recurrent models of visual attention. *Adv Neural Inf Process Syst* 2014; 27: 2204–2212.
45. Guo M-H, Lu C-Z, Liu Z-N, et al. Visual attention network. *Comput Vis Media* 2023; 9: 733–752.
46. Chen G, Liu Y, Qian J, et al. DSEU-net: a novel deep supervision SEU-net for medical ultrasound image segmentation. *Expert Syst Appl* 2023; 223: 119939.
47. Singla R, Ringstrom C, Hu G, et al. *The open kidney ultrasound data set, international workshop on advances in simplifying medical ultrasound*. Springer, Melbourne, Australia, October 17–20, 2023, pp. 155–164.
48. Mohammadi R, Shokatian I, Salehi M, et al. Deep learning-based auto-segmentation of organs at risk in high-dose rate brachytherapy of cervical cancer. *Radiother Oncol* 2021; 159: 231–240.
49. Chaurasia A and Culurciello E, Linknet: exploiting encoder representations for efficient semantic segmentation. In: *2017 IEEE visual communications and image processing (VCIP)*, IEEE, St. Petersburg, Florida, USA, December 10–13, 2017, pp. 1–4.
50. Chen G, Dai Y, Li R, et al. SDFNet: automatic segmentation of kidney ultrasound images using multiscale low-level structural feature. *Expert Syst Appl* 2021; 185: 115619.
51. Long J, Shelhamer E and Darrell T. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.
52. Cao H, Wang Y, Chen J, et al. Swin-unet: unet-like pure transformer for medical image segmentation. In: *Proceedings of computer vision–ECCV 2022 workshops, Part III, Tel Aviv, Israel, October 23–27, 2022*. Springer, 2023, pp. 205–218.