

SE-UNet: Spatial and Channel Refinement for Robust Multi-Anatomical Structure Segmentation in Abdominal Ultrasound

[Md. Rashed](#) , [Muhammad Jamil](#) , [Md. Imran Hossain](#) ^{*} , [Adnan Kavak](#) ^{*} , Md. Riad Hassan , [Özgür Çakır](#) , [Hossein Fotouhi](#) ^{*}

Posted Date: 21 July 2026

doi: 10.20944/preprints2026071481.v1

Keywords: ultrasound segmentation; multi-anatomical structure segmentation; U-Net; ResNet-34; squeeze-and-excitation; spatial attention; dice-focal loss; clinical decision support



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

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.

Article

SE-UNet: Spatial and Channel Refinement for Robust Multi-Anatomical Structure Segmentation in Abdominal Ultrasound

Md. Rashed ¹, Muhammad Jamil ^{2,3}, Md. Imran Hossain ^{1,*}, Adnan Kavak ^{2,3,*},
Md. Riad Hassan ⁴, Özgür Çakır ⁵ and Hossein Fotouhi ^{6,*}

- ¹ Department of Information and Communication Engineering, Pabna University of Science and Technology, Pabna 6600, Bangladesh
- ² Department of Computer Engineering, Kocaeli University, 41001, İzmit, Türkiye
- ³ Wireless Information and Intelligent Systems (WINS) Research Center, Kocaeli University, 41001, İzmit, Türkiye
- ⁴ Department of Computer Science and Engineering, Green University of Bangladesh, Dhaka, Bangladesh
- ⁵ Faculty of Medicine, Department of Radiology, Research and Application Hospital, Kocaeli University, 41001 İzmit, Türkiye
- ⁶ Department of Computer Science and Engineering, Mälardalen University (MDU), 72123 Västerås, Sweden
- * Correspondence: imran05ice@pust.ac.bd (M.I.H.); akavak@kocaeli.edu.tr (A.K.); hossein.fotouhi@mdu.se (H.F.)

Abstract

Automated segmentation of multiple anatomical structures in abdominal ultrasound could enhance clinical decision-making by providing consistent quantitative measurements for disease monitoring. However, ultrasound-specific artifacts including speckle noise, low contrast and weak boundaries challenge existing deep learning models. We propose SE-UNet, a novel architecture that integrates spatial attention gates and squeeze-and-excitation blocks to improve boundary localization and tissue discrimination. On a multi-anatomical dataset combining kidney and spleen ultrasound images (six tissue classes), SE-UNet achieved a mean Dice score of 0.7165, outperforming nine baseline methods. Clinical validation with an expert radiologist on 50 cases demonstrated 0.78 Dice agreement with manual annotations, with 88% of segmentations deemed clinically acceptable for measurement purposes. The model processes images in 0.028 seconds, supporting real-time clinical workflow integration. Our approach demonstrates how combined spatial and channel refinement addresses key ultrasound segmentation challenges with direct clinical applicability.

Keywords: ultrasound segmentation; multi-anatomical structure segmentation; U-Net; ResNet-34; squeeze-and-excitation; spatial attention; dice-focal loss; clinical decision support

1. Introduction

Ultrasound imaging is one of the most widely used diagnostic tools worldwide, alongside X-ray, CT, MRI and PET [1,2]. Its advantages include real-time imaging, portability and no radiation exposure, making it ideal for routine abdominal exams and repeated monitoring [3]. However, interpreting ultrasound images remains challenging even for experts [4–6]. The images contain speckle noise, acoustic shadows and low contrast between tissues [7,8]. These artifacts make it difficult to identify organ boundaries consistently.

Accurate multi-anatomical segmentation in abdominal ultrasound enables quantitative assessment of organ morphology, which is crucial for diagnosing and monitoring conditions like chronic kidney disease, spleen enlargement and abdominal pathologies. Automated tools could standardize measurements currently subject to inter-observer variability, potentially improving diagnostic consistency and reducing radiologist workload.

In clinical practice, radiologists often need to measure multiple anatomical structures during abdominal scans. Accurate segmentation of kidney substructures (cortex, medulla) and spleen regions supports quantitative assessment of diseases like chronic kidney disease [10,11]. Manual segmentation

is time-consuming and varies between operators [12]. Automated tools could reduce this variability and save valuable clinical time.

Despite recent advances in deep learning, multi-anatomical structure (MAS) segmentation in ultrasound remains difficult. Models trained on cleaner modalities like MRI often fail on ultrasound due to its unique artifacts [9]. Even specialized ultrasound segmentation methods struggle with key problems: missing boundaries in low-contrast regions, leakage into background areas and confusion between visually similar tissues [13–15]. These limitations stem from inadequate handling of both spatial context and channel-wise feature importance [17,18].

Compared with MRI and CT, ultrasound shows substantial speckle noise, acoustic shadowing, signal dropouts, intensity changes and weak boundaries between tissues [7,8]. This affects how well models can find organ boundaries. Standard models often produce masks with missing pieces. Performance also varies across patients and imaging devices [9]. The challenge is greater in MAS settings because anatomical regions may look similar, vary in size and appear partially hidden. This requires robust multi-scale representation while keeping fine spatial details.

This study has two main motivations. First, MAS segmentation in abdominal ultrasound can support clinically meaningful measurements like organ localization, thickness estimation and volume approximation. These are relevant for disease screening, treatment planning and monitoring over time [10,11]. Second, reliable MAS segmentation reduces differences between operators in manual drawing. This improves consistency and enables computer-aided diagnosis and decision support without adding much work for radiologists [12]. Specifically, accurate segmentation of kidney substructures (cortex and medulla) and spleen regions can help in assessing chronic kidney disease, inflammation and other conditions where organ shape and echo patterns matter clinically.

From an application view, we aim for radiology workflow automation and point-of-care ultrasound. Here, fast and stable results are needed across different machines and operator skill levels. In such settings, segmentation models must handle low signal-to-noise ratios and keep clinically important boundaries clear. They also need to be efficient enough for hospital systems or edge devices. So, robustness to artifacts and balanced accuracy across classes are key needs, along with good overall performance on easier regions.

Many studies have explored deep learning for ultrasound segmentation, including U-Net variants, multi-scale decoders and attention-based designs [13–16]. Recent methods improve context gathering or efficiency through multi-scale attention, lightweight decoders, or transformer-like parts. They show strong results in specific organ segmentation tasks. However, many current solutions are made for single-organ cases and do not fully address MAS needs in ultrasound. Models that focus on global context may lose boundary details. Models that focus only on local features may still be sensitive to speckle and background noise. Attention designs help with localization but often do not adjust channel importance well enough. This can lead to unstable results when multiple similar regions must be separated in one prediction [17,18]. Also, class imbalance (lots of background versus small tissue regions) can bias training toward easier cases. This hurts segmentation of smaller structures unless the loss function is designed to counter this effect [19,20].

Thus, the main limitations of existing methods are: (i) incomplete boundary recovery in low-contrast, noisy conditions, (ii) background leakage due to poor filtering of irrelevant skip features, (iii) reduced accuracy across classes when tissues look alike and (iv) limited robustness across different imaging conditions and label sets. These issues point to the need for an architecture that combines spatial focus (where to look) and channel adjustment (what to emphasize) within an encoder-decoder setup, along with a loss that handles class imbalance.

Based on these gaps, our paper answers these research questions:

- **RQ1:** How can we design a segmentation architecture to improve boundary finding and reduce background errors for MAS in ultrasound?
- **RQ2:** Can using spatial attention followed by channel recalibration improve balanced accuracy across different anatomical regions?

- **RQ3:** How should we set the training loss to handle both overlap quality and class imbalance in multi-class ultrasound segmentation?
- **RQ4:** Does the proposed framework show consistent gains over strong baselines under the same experimental conditions?

Our study objectives are: (i) to develop a robust MAS segmentation method for ultrasound that improves feature learning and location under noise and weak edges; (ii) to create a combined multi-class dataset by merging kidney and spleen ultrasound labels for better MAS evaluation; (iii) to design a training approach and loss function that boost stability and class-balanced accuracy; and (iv) to compare thoroughly against current segmentation networks using the same preprocessing, training and metrics.

To meet these goals, we propose *SE-UNet*, a ResNet-34 encoder-decoder model that improves feature learning and localization through step-by-step refinement. First, decoder-guided spatial attention gates filter skip features to highlight anatomically relevant areas and reduce clutter before merging. Second, squeeze-and-excitation (SE) blocks adjust channel responses to strengthen useful feature channels and cut down redundant ones. This attention-SE combination is added at each decoder stage to keep fine details while improving multi-scale understanding. We also train the model with a combined Dice and focal loss to improve overlap and handle class imbalance together.

This paper contributes the following:

- SE-UNet for multi-class ultrasound segmentation, using spatial attention gates and SE blocks.
- A merged MAS dataset by combining kidney and spleen ultrasound labels into six classes.
- Training with a Dice and focal combined loss to manage overlap quality and class imbalance.
- Comparison with nine baselines using identical training and test setups.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 presents the SE-UNet design. Section 4 describes experiments and evaluation. Section 5 presents results. Section 6 discusses findings. Section 7 concludes the paper.

2. Related Works

Deep learning has significantly improved ultrasound image segmentation in recent years. Many studies have proposed methods to increase accuracy and robustness under ultrasound-specific artifacts. Chen et al. (2024) [13] proposed MSCA-UNet for medical ultrasound image segmentation. The model uses a multi-scale module to generate feature maps at different scales and applies channel attention to adjust channel weights. A global feature module helps preserve edge details. The network improved Dice and IoU on several datasets. The authors also highlighted common ultrasound challenges, including noise, low contrast and uneven grayscale distribution, which make edge extraction difficult. In addition, the dataset size was limited. The study in [14] proposed a two-stage framework for breast ultrasound image segmentation. The first U-Net detects the lesion region and the second U-Net segments the lesion within the cropped region. This method performed better than a single-stage U-Net in their experiments. However, the final segmentation depends on correct detection in the first stage. If detection fails, the second-stage segmentation also fails. The method is also limited by the small size of the available dataset.

EMCAD [15] introduced an efficient attention decoder that reduces model complexity for multi-organ ultrasound segmentation. It reported a segmentation accuracy of 83.63% and outperformed several baseline methods. However, the study did not include real-world clinical validation. It also provided limited discussion of robustness under common clinical ultrasound artifacts. MISSFormer [16] and ConvTrans-Net [21] combined transformer-based designs with CNN components to capture global and local context. These architectures can model complex anatomical structures in ultrasound imaging. However, discriminating visually similar structures remains challenging in some cases. Another study [17] developed an Attention U-Net model for detecting carotid plaques in ultrasound images of the internal and common carotid arteries. It reported better performance than baseline models such as U-Net, U-Net++ and U-Net3P. However, the authors highlighted data selection limitations and suggested using multicenter datasets and stronger augmentation to improve generalization.

U-Net and Attention U-Net were originally designed for MRI and can achieve strong boundary delineation in kidney segmentation [22]. These architectures are also relevant for ultrasound segmentation. However, performance can degrade in ultrasound due to thick edges, low contrast boundaries and variations in organ size. CMUNeXt [18] was designed to capture global image context efficiently. It uses inductive bias, large kernels and inverted bottlenecks to combine distant and local spatial information. CMUNeXt reported strong performance compared with both heavyweight and lightweight segmentation models. Its efficiency can be useful when training data are limited. UNETR++ [19] is a hybrid model for 3D medical image segmentation. It introduced an efficient paired-attention block with a linear spatial attention module and shared parameters to achieve high performance with reduced computational cost. The authors noted limitations when segmenting multiple structures with rare abdominal shapes and suggested future work using improved augmentation strategies.

ResUNet++ [23] was proposed to improve polyp segmentation in colonoscopy data and achieved strong performance on that domain. However, it does not explicitly address the typical artifacts found in ultrasound images. FCNN [24], originally developed for CT, has been adapted to other modalities such as MRI and achieved strong results in multi-organ segmentation. However, it can still struggle on ultrasound images due to speckle noise and low contrast. The deep supervised CNN framework in [25] incorporated multi-scale inputs and pyramid pooling for kidney ultrasound segmentation and achieved high Dice scores. However, it was not designed for multi-organ or multi-anatomical structure segmentation.

BACANet [26] is a liver ultrasound segmentation model that integrates convolutional attention, boundary supervision and large kernels to achieve high accuracy with a lightweight design. However, capturing small anatomical structures remains challenging. Residual U-Net [27] introduced residual blocks and attention fusion to improve breast tumor ultrasound segmentation. It improved segmentation accuracy and gradient stability during training. However, the work does not clearly describe how the method handles difficult cases with strong ultrasound artifacts. The multi-modality framework in [28] distinguished between single-modality and multi-modality information through feature fusion. However, limitations can arise during alignment and high-level semantic fusion may reduce modality-specific details.

Kolmogorov-Arnold Networks [29] addressed limitations of linear modeling in standard transformer and CNN architectures. The method introduced a two-phase encoder-decoder design, with a convolutional phase for feature extraction and a tokenized KAN phase for modeling complex patterns. It achieved an IoU of 78.69% and an F1 score of 87.22%. However, a key gap remains in integrating KANs into general-purpose and real-time systems beyond validation tasks. MCAFT [30] combined transformer-based attention with frequency-domain analysis using a Discrete Wavelet Transform to capture both low-frequency structures and high-frequency edge details. It achieved a Dice score of 83.87 on the Synapse abdominal multi-organ dataset. However, processing high-resolution images can introduce substantial computational overhead.

PMFSNet [31] aimed to balance global and local features while reducing redundancy by simplifying the U-Net design and self-attention mechanisms. It achieved strong performance across several public datasets. However, reduced complexity may limit the ability to capture very fine anatomical details. Karimi et al. [32] designed a two-stage framework for spleen ultrasound segmentation. First, a SegFormerB0 model generates a coarse segmentation. Then, a Pix2Pix model refines boundaries and reduces errors. The framework achieved a mean Dice of 96.82% and a mean IoU of 94.17%, outperforming traditional U-Net and SS-Net baselines. However, the approach focuses on a limited region of interest and may not generalize well to broader MAS settings.

In a nutshell, these studies show steady progress in ultrasound segmentation, but challenges remain in robust boundary delineation, feature discrimination across similar tissues and generalization under diverse acquisition conditions. To address these limitations, we propose SE-UNet, which combines spatial attention with squeeze-and-excitation blocks in a pretrained ResNet-34 encoder-decoder. Table 1 summarizes representative methods, their main contributions, limitations and reported performance.

Table 1. Summary of recent AI-based segmentation models, highlighting their main contributions, limitations, application domains and reported performance/efficiency.

References	Method	Key contributions and limitations	Dataset	Performance / efficiency
[13]	MSCA-UNet	Multi-scale feature extraction with channel attention and depth-separable convolutions to reduce parameters. Limitation: robustness may decrease under severe ultrasound noise and low-contrast boundaries.	Multi-organ ultrasound (4 datasets)	Outperformed prior methods; 29.82% fewer parameters than U-Net
[14]	Two-stage U-Net pipeline	Two U-Nets are used sequentially for lesion detection and fine segmentation in cropped regions. Limitation: segmentation performance depends on successful detection in the first stage; limited by small dataset size.	Breast ultrasound segmentation	Dice improvement: 1.8% overall; 14.5% for hard cases (Dice < 70%)
[15]	EMCAD	Efficient attention decoder with depth-wise convolutions and gated attention for reduced complexity. Limitation: lacks real-world clinical validation and provides limited analysis of robustness under typical clinical artifacts.	Multi-organ ultrasound (12 datasets)	1.91M parameters, 0.381G FLOPs; ~80% reduction in parameters/FLOPs
[16]	MISSFormer	Transformer-based segmentation that captures global and local context. Limitation: feature discrimination between visually similar structures can remain challenging.	Multi-organ and cardiac segmentation	State-of-the-art performance across tasks
[17]	Attention U-Net (cardiovascular)	Attention-augmented U-Net for carotid plaque segmentation in ultrasound. Limitation: generalization may be affected by data selection; multicenter data and stronger augmentation are recommended.	Carotid ultrasound	Correlation: 0.96; AUC: 0.97; surpassed U-Net, U-Net++ and U-Net3P
[18]	CMUNeXt	Efficient global-context modeling using large kernels and inverted bottlenecks. Limitation: may require careful tuning on very small datasets to avoid overfitting.	General medical image segmentation	Effective global context capture with low computational cost
[19]	UNETR++	Hybrid 3D segmentation model with an efficient paired-attention block and linear spatial attention. Limitation: multi-structure segmentation can be difficult for rare abdominal shapes; improved augmentation is suggested.	Synapse, BTCV, ACDC, BraTS, Decathlon Lung	Dice: 87.2%
[21]	ConvTrans-Net	CNN-Transformer hybrid with attention and MLP-based feature preselection. Limitation: separating visually similar tissues can still be challenging in ultrasound.	Breast ultrasound	Jaccard: 85.21%; Precision: 85.17%; Recall: 89.65%
[22]	U-Net and Attention U-Net (MRI)	Benchmark U-Net variants evaluated with standard and focal cross-entropy on MRI kidney segmentation. Limitation: performance can degrade in ultrasound due to thick edges and low-contrast boundaries.	Kidney segmentation in MRI	Dice: 0.966; strong boundary delineation
[23]	ResUNet++	Residual U-Net variant that improves boundary modeling for polyp segmentation. Limitation: not tailored to ultrasound-specific artifacts such as speckle and shadowing.	Kvasir-SEG, CVC-612	Outperformed U-Net and ResUNet; high Dice and mIoU
[24]	FCNN (CT → MRI)	Modality-robust FCNN originally developed for CT and adapted to MRI. Limitation: performance may degrade on ultrasound due to speckle noise and low contrast.	Multi-organ MRI segmentation	High Dice; modality-independent performance
[25]	Deep-supervised CNN (kidney)	Multi-scale inputs, pyramid pooling and deep supervision for kidney ultrasound segmentation. Limitation: not designed for multi-organ or MAS segmentation settings.	Kidney ultrasound	Dice: 95.86%; robust under typical ultrasound noise
[26]	BACANet	Lightweight liver ultrasound segmentation with convolutional attention, boundary supervision and large kernels. Limitation: small structures and fine details can be difficult to capture.	Liver ultrasound	Lightweight with high segmentation accuracy
[27]	Residual attention U-Net	Residual blocks with attention fusion to improve breast tumor ultrasound segmentation and training stability. Limitation: limited discussion of failure cases under strong ultrasound artifacts.	Breast ultrasound	Improved accuracy and gradient stability
[28]	Multi-modality segmentation framework	Fusion framework that separates single-modality and multi-modality information for segmentation. Limitation: high-level semantic alignment may reduce modality-specific details.	Multi-modality medical images	Improved over single-modality approaches
[29]	Kolmogorov—Arnold networks (KAN)	Two-phase encoder-decoder combining convolutional feature extraction with tokenized KAN for complex pattern modeling. Limitation: limited evidence for real-time or general-purpose deployment beyond validation studies.	Medical image data	IoU: 78.69%; F1-score: 87.22%
[30]	MCAFT	Transformer-based segmentation with frequency-domain analysis via Discrete Wavelet Transform for edge/structure decomposition. Limitation: computational overhead can be high for high-resolution images.	Synapse abdominal multi-organ data	Dice: 83.87%
[31]	PMFSNet	Simplified U-Net and self-attention to balance global/local features with reduced redundancy. Limitation: reduced capacity may limit very fine anatomical detail capture.	Multi-organ and multi-task segmentation (7 datasets)	Strong performance across all evaluated datasets
[32]	SegFormerB0—Pix2Pix (two-stage)	Two-stage pipeline with SegFormerB0 for coarse masks and Pix2Pix refinement for fine boundaries. Limitation: region-of-interest focus may limit generalization to broader MAS settings.	Spleen ultrasound	Mean Dice: 96.82%; Mean IoU: 94.17%

Synthesis of Research Gaps

Reviewing these methods reveals a clear progression in ultrasound segmentation. Early CNN-based approaches like U-Net [22] established strong baselines but lacked mechanisms to handle ultrasound-specific artifacts. Attention mechanisms [17] and transformer-based designs [16,21] improved context modeling but often at high computational cost or with limited focus on channel discrimination. Several gaps remain unaddressed for multi-anatomical segmentation in ultrasound. First, most methods treat spatial attention and channel recalibration separately, missing their complementary benefits [17,18]. Second, many approaches are validated on single-organ tasks but struggle with the class imbalance and anatomical diversity of MAS segmentation [19,20]. Third, few studies systematically evaluate both overlap metrics and boundary accuracy under consistent experimental conditions [13,15]. Our work addresses these gaps by integrating spatial attention gates with squeeze-and-excitation blocks in a unified framework. This sequential refinement approach allows the model to first decide *where* to focus (spatial attention) and then determine *what* features to emphasize (channel recalibration). We validate this design on a challenging MAS dataset with rigorous comparisons against multiple baselines.

3. Proposed Methodology

We designed SE-UNet to tackle the specific challenges of MAS segmentation in ultrasound images. The architecture follows an encoder-decoder structure with two key innovations: spatial attention gates and squeeze-and-excitation blocks. The complete architecture is shown in Figure 1. The spatial attention process is detailed in Fig. 2, while the channel refinement via squeeze-and-excitation is shown in Fig. 3.

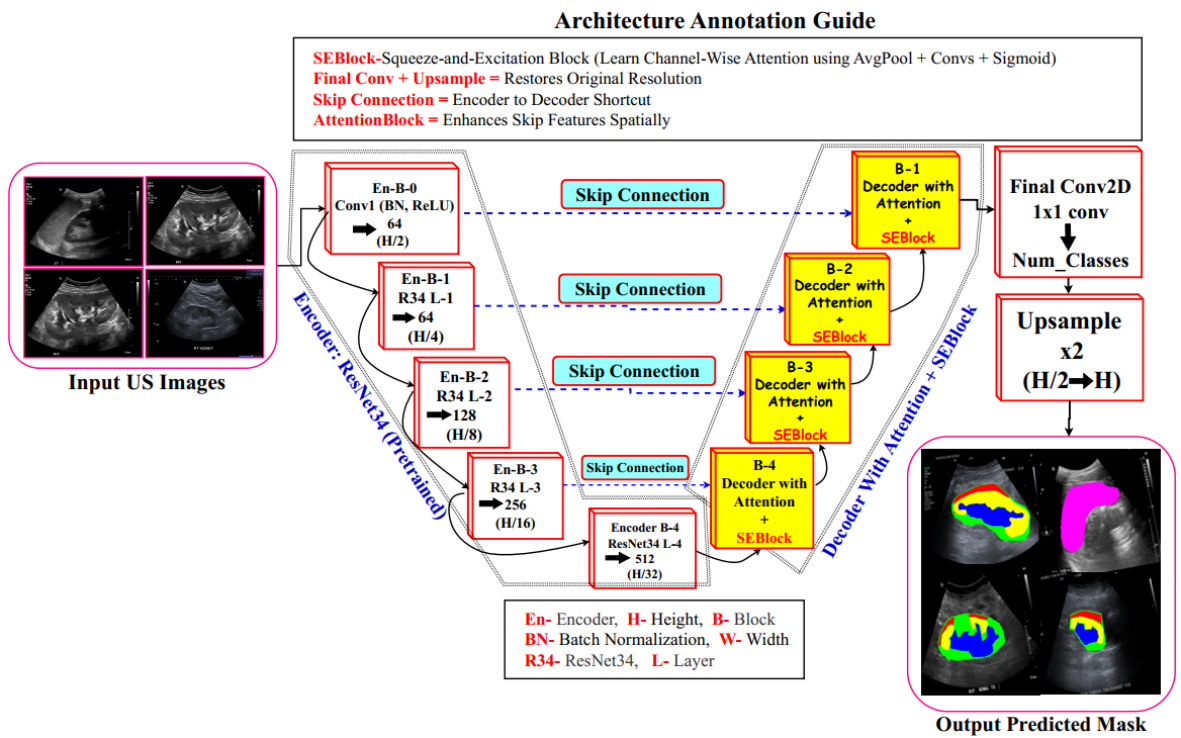


Figure 1. Proposed SE-UNet architecture for multi-anatomical structure segmentation in ultrasound images. The model integrates a ResNet-based encoder with attention and squeeze-and-excitation (SE) enhanced decoder stages.

We chose ResNet-34 as our encoder backbone for three reasons. First, its residual connections help with gradient flow in deep networks, which is crucial for learning complex anatomical patterns in noisy ultrasound data. Second, its intermediate feature maps capture multi-scale information

important for organs of different sizes. Third, ImageNet-pretrained weights provide a good starting point, which is valuable given our limited ultrasound training data [13].

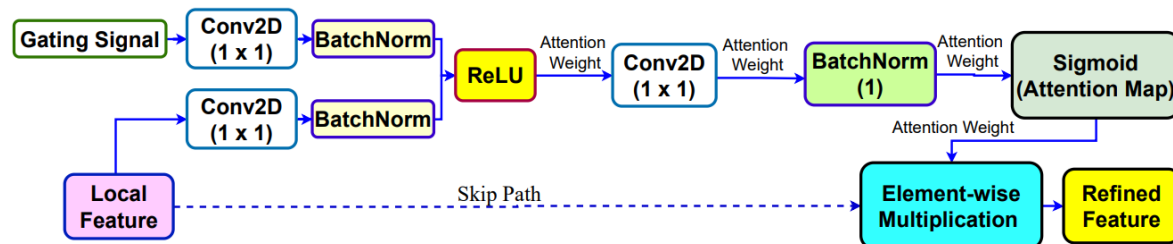


Figure 2. Spatial attention gate used to refine encoder skip features using decoder gating signals.

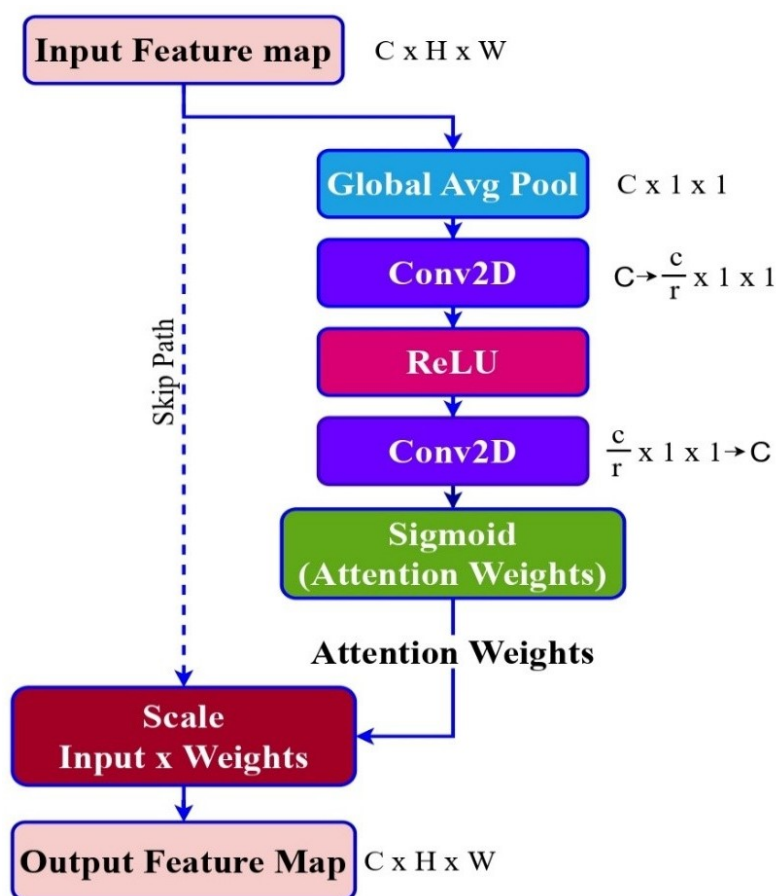


Figure 3. Squeeze-and-excitation (SE) block for channel-wise feature recalibration.

The decoder reconstructs segmentation masks from these features. Simply copying encoder features via skip connections can propagate noise and irrelevant details [17]. To solve this, we introduce spatial attention gates that use decoder context to filter encoder features. This keeps important anatomical regions while suppressing background clutter.

Even with spatial filtering, different feature channels may respond inconsistently to similar tissues. We address this with squeeze-and-excitation blocks that learn to emphasize informative channels and suppress redundant ones [13]. This combination of spatial and channel refinement forms the core of our approach to handling ultrasound-specific challenges like speckle noise and weak boundaries [7,8].

Class definition:

We use a multi-class segmentation setup with $K = 6$ classes. The classes are {background, capsule, central echo complex, medulla, cortex, spleen}. In the results section, per-tissue Dice scores are reported for the five foreground classes.

3.1. Notation and End-to-End Mapping

Let the input ultrasound image be denoted as

$$\mathbf{x} \in \mathbb{R}^{H \times W \times C}, \quad (1)$$

where H , W and C represent image height, width and channels, respectively. The model outputs a pixel-wise segmentation probability map

$$\hat{\mathbf{Y}} \in [0, 1]^{H \times W \times K}, \quad (2)$$

where K is the number of anatomical classes and $\sum_{k=1}^K \hat{Y}_{i,j,k} = 1$ for each pixel (i, j) . The encoder generates hierarchical feature maps

$$\{\mathbf{E}_s\}_{s=0}^S, \quad \mathbf{E}_s \in \mathbb{R}^{H_s \times W_s \times d_s}, \quad (3)$$

which are progressively reconstructed by the decoder into

$$\{\mathbf{D}_s\}_{s=0}^{S-1}, \quad \mathbf{D}_s \in \mathbb{R}^{H_s \times W_s \times d'_s}. \quad (4)$$

3.2. ResNet-34 Encoder

The encoder is based on ResNet-34 and employs residual learning to stabilize gradient propagation in deep networks. Each residual block is defined as

$$\text{ResBlock}(\mathbf{u}) = \sigma(\mathcal{F}(\mathbf{u}) + \mathcal{S}(\mathbf{u})), \quad (5)$$

where $\sigma(\cdot)$ denotes ReLU activation, $\mathcal{F}(\cdot)$ is the residual transformation and $\mathcal{S}(\cdot)$ is the shortcut mapping. The residual transformation consists of two convolutional layers:

$$\mathcal{F}(\mathbf{u}) = \text{BN}_2(\text{Conv}_2(\sigma(\text{BN}_1(\text{Conv}_1(\mathbf{u}))))), \quad (6)$$

while the shortcut is defined as

$$\mathcal{S}(\mathbf{u}) = \begin{cases} \mathbf{u}, & \text{if dimensions match,} \\ \text{Conv}_{1 \times 1}(\mathbf{u}), & \text{otherwise.} \end{cases} \quad (7)$$

The encoder produces multi-scale features $\{\mathbf{E}_s\}$, which are forwarded to the decoder through skip connections.

3.3. Decoder and Feature Alignment

Every decoder phase upsamples deeper features:

$$\tilde{\mathbf{D}}_s = \text{Up}(\mathbf{D}_{s+1}), \quad (8)$$

and aligns the associated encoder features:

$$\tilde{\mathbf{E}}_s = \text{Conv}_{1 \times 1}(\mathbf{E}_s), \quad (9)$$

ensuring spatial and channel dimensions match before combining features.

3.4. Spatial Attention Gate

The spatial attention gate U_s selectively highlights relevant spatial regions using skip features and decoder features:

$$\mathbf{U}_s = \sigma(\text{Conv}_{1 \times 1}(\tilde{\mathbf{E}}_s) + \text{Conv}_{1 \times 1}(\tilde{\mathbf{D}}_s)), \quad (10)$$

$$\mathbf{q}_s = \text{Sigmoid}(\text{Conv}_{1 \times 1}(\mathbf{U}_s)), \quad (11)$$

$$\mathbf{A}_s = \tilde{\mathbf{E}}_s \odot \mathbf{q}_s. \quad (12)$$

3.5. Squeeze-and-Excitation Block

The squeeze-and-excitation block adjusts channel-wise responses using global context. Given $\mathbf{A}_s \in \mathbb{R}^{H_s \times W_s \times d}$, the squeeze operation computes

$$z_c = \frac{1}{H_s W_s} \sum_{i=1}^{H_s} \sum_{j=1}^{W_s} (\mathbf{A}_s)_{i,j,c}, \quad (13)$$

followed by excitation

$$\mathbf{s} = \text{Sigmoid}(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z})), \quad (14)$$

and channel-wise scaling

$$\mathbf{S}_s = \mathbf{A}_s \odot \mathbf{s}. \quad (15)$$

3.6. Decoder Fusion and Final Prediction

The enhanced skip feature $\mathbf{S}_s$ is integrated with the decoder attribute and refined through convolution as:

$$\mathbf{D}_s = \text{DecBlock}(\text{Concat}(\mathbf{S}_s, \tilde{\mathbf{D}}_s)). \quad (16)$$

At the last phase, a 1×1 convolution projects attribute to class logits as:

$$\ell = \text{Conv}_{1 \times 1}(\mathbf{D}_0), \quad (17)$$

and softmax function yields pixel-wise probabilities:

$$\hat{Y}_{i,j,k} = \frac{\exp(\ell_{i,j,k})}{\sum_{t=1}^K \exp(\ell_{i,j,t})}. \quad (18)$$

4. Experiments

This section describes the experimental setup used to evaluate SE-UNet. It includes the dataset, preprocessing, training settings, evaluation metrics and baseline methods. All settings follow the methodology in Section III to ensure a fair and reproducible evaluation. Table 2 summarizes the main hyperparameters.

Table 2. Hyperparameter settings of the proposed SE-UNet model for multi-anatomical ultrasound segmentation.

Hyperparameter	Value
Optimizer	Adam
Learning rate	0.001
Learning scheduler	Not used / default (Adam step-size control)
Loss function	Composite loss: Dice + Focal ($\alpha = 0.7$, $\gamma = 2.0$)
Number of epochs	400
Batch size	8
Input dimension	$256 \times 256 \times 3$ (RGB)
Hidden dimension	ResNet-34 backbone (64, 128, 256, 512 filters)
Number of classes (K)	6 (background + 5 structures)

4.1. Dataset Description and Preparation

Experiments were conducted on a curated multi-anatomical ultrasound dataset constructed by merging two publicly available datasets: a kidney ultrasound dataset [33] containing capsule, central echo complex, medulla and cortex annotations and a spleen ultrasound dataset [32] with pixel-level labels. The merged dataset enables multi-class segmentation under heterogeneous anatomical and

imaging conditions, reflecting realistic clinical variability. Table 4 shows the summary of the ultrasound MAS dataset used in our experiments.

To create a unified MAS segmentation dataset, annotations from the kidney and spleen ultrasound datasets were mapped into a common label space, as shown in Table 3. The kidney dataset contributes four anatomical structures (capsule, central echo complex, medulla and cortex), while the spleen dataset contributes the spleen class. Background labels from both datasets were mapped to a shared background class (Class ID 0). Since the two datasets contain distinct anatomical structures rather than duplicate annotations of the same organ, no anatomical class was annotated by multiple sources. This harmonization enables joint training and evaluation within a consistent six-class segmentation framework.

Table 3. Annotation harmonization of the merged multi-anatomical structure (MAS) ultrasound dataset.

Dataset Source	Original Annotation	Unified Class ID
Kidney Dataset	Background	0
Kidney Dataset	Capsule	1
Kidney Dataset	Central Echo Complex (CEC)	2
Kidney Dataset	Medulla	3
Kidney Dataset	Cortex	4
Spleen Dataset	Background	0
Spleen Dataset	Spleen	5

Label standardization:

We unify the labels from both datasets into one shared label set with $K = 6$ classes. The class order is background, capsule, central echo complex, medulla, cortex and spleen. We convert each mask into a single-channel label map. We also convert masks to one-hot format during training.

All images were resized to 256×256 pixels and normalized to the range $[0, 1]$. We use $K = 6$ classes, including background. Ground-truth masks are converted to one-hot encoding to match the softmax output. We split the dataset into training (90%) and validation (10%) sets. We perform the split at the image level. When patient identifiers are available, we ensure that images from the same patient do not appear in more than one split. We also reserve an independent test set with no overlap with the training and validation sets.

Table 4. Class presence counts in the merged MAS dataset. The numbers show how many images contain each labeled structure.

Tissue Type	Capsule	Central Echo Complex	Medulla	Cortex	Spleen
No. of images with label	534	534	534	534	450
Labeled Color	red	green	blue	yellow	purple

Data Split Strategy:

We divided the dataset into training (80%), validation (10%) and (10%) test sets. The validation set monitored training for early stopping, while the test set remained completely untouched during training. All reported results come from this independent test set. When patient identifiers were available, we ensured no patient appeared in multiple splits. This prevents data leakage and gives a realistic estimate of clinical performance [12].

Hyperparameter Selection:

Key hyperparameters were chosen based on validation performance and established practices. We used ResNet-34 as it offered the best balance between depth and efficiency in preliminary tests. The loss weighting $\alpha = 0.7$ gave more importance to Dice loss while still benefiting from focal loss’s handling of class imbalance [19,20]. The focal parameter $\gamma = 2.0$ follows common usage for emphasizing hard examples [22]. We used Adam optimizer with learning rate 0.001 as it provided stable convergence across all compared methods [13,15].

4.2. Implementation Details and Training Protocol

We implemented SE-UNet in PyTorch. **Experimental Hardware and Inference Setup:** All experiments were conducted on a high-performance workstation equipped with an NVIDIA GeForce RTX 4090 GPU with 24 GB of memory. The reported inference time corresponds to the average forward-pass time per image on the independent test set using an input resolution of 256×256 and a batch size of 1 under evaluation mode. The measured average processing time was 0.028 s per image. The reported timing excludes model loading and preprocessing overhead. We initialize the ResNet-34 encoder with ImageNet-pretrained weights. All decoder layers, attention gates and SE blocks are trained from scratch. We train using the Adam optimizer with an initial learning rate of 0.001 and a batch size of 8 for 400 epochs. We use early stopping based on validation loss to reduce overfitting. We apply data augmentation during training, including random rotation, horizontal flip and intensity scaling.

To ensure reproducibility, the same preprocessing steps, augmentation strategies and hyperparameters were applied to all baseline models. Code and trained weights are available on GitHub for verification. This allows other researchers to verify our results and build upon this work.

4.3. Reproducibility checklist

To keep the comparison fair, we keep the same settings for all methods.

- Fixed train, validation and test splits.
- Same preprocessing and input size (256×256) for all methods.
- Same augmentations for all methods (rotation, flip and intensity scaling).
- Same optimizer (Adam), learning rate (0.001), batch size (8) and epochs (400).
- Same loss setup for all methods (Dice and focal with $\alpha = 0.7$ and $\gamma = 2.0$).
- Code and weights: <https://github.com/rashed200613/SE-UNet>.

4.4. Loss Function

To address class imbalance and boundary precision in multi-class segmentation, the training objective combines Dice loss and focal loss. Let $\hat{Y}_{i,j,k}$ and $Y_{i,j,k}$ represent the predicted probability and ground-truth label for class k at pixel (i, j) , in the corresponding order.

The multi-class soft Dice loss is defined as

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{1}{K} \sum_{k=1}^K \frac{2 \sum_{i,j} \hat{Y}_{i,j,k} Y_{i,j,k} + \epsilon}{\sum_{i,j} \hat{Y}_{i,j,k} + \sum_{i,j} Y_{i,j,k} + \epsilon}, \quad (19)$$

where ϵ is a small constant for numerical stability.

The focal loss is given by

$$\mathcal{L}_{\text{Focal}} = - \frac{1}{HW} \sum_{i,j} \sum_{k=1}^K Y_{i,j,k} (1 - \hat{Y}_{i,j,k})^\gamma \log(\hat{Y}_{i,j,k}), \quad (20)$$

where γ controls the focusing strength.

The final training loss is computed as

$$\mathcal{L} = \alpha \mathcal{L}_{\text{Dice}} + (1 - \alpha) \mathcal{L}_{\text{Focal}}, \quad (21)$$

with $\alpha = 0.7$ in all experiments.

4.5. Comparative Baselines

The proposed SE-UNet was evaluated against other established segmentation models reported in the literature, including but not limited to Att-UNet, CMUNeXt, UDBAUNet, UNet, UNet++, ResUNet++, UDBRNet, Residual U-Net and MultiResUNet. All baseline models were trained and

evaluated under identical conditions, including dataset splits, input resolution, loss functions and optimization settings, to ensure balanced comparison.

5. Results

This section reports the quantitative and qualitative performance of the proposed SE-UNet for multi-anatomical structure (MAS) segmentation in ultrasound (US) images. All results are computed on the independent test set using the same preprocessing, training configuration and composite Dice-Focal objective defined in the Experiments section. We first formalize the evaluation metrics (consistent with the confusion-matrix formulation), then present overall and class-wise quantitative comparisons against competitive baselines, followed by qualitative visual assessment, confusion-matrix analysis and a loss-function ablation study that validates the selected training objective. Performance is evaluated using standard metrics that measure both overlap accuracy (Dice, IoU) and boundary precision (Hausdorff Distance, ASD). All results are reported with mean values across the test set.

5.1. Evaluation Metrics

We evaluate segmentation performance using standard overlap, pixel-wise and boundary metrics. The metrics are defined below.

- Pixel Accuracy (PA) is the ratio of correctly classified pixels to the total number of pixels.
- Mean Pixel Accuracy (MPA) is the mean of per-class pixel accuracies.
- Precision is the ratio of correctly predicted positive pixels to all predicted positive pixels.
- Recall is the ratio of correctly predicted positive pixels to all ground-truth positive pixels.
- Dice Similarity Coefficient (Dice) measures region overlap between prediction and ground truth.
- HD95 is the 95th percentile Hausdorff distance between the predicted boundary and the ground-truth boundary.
- Average Symmetric Surface Distance (ASD) is the mean symmetric distance between the predicted boundary and the ground-truth boundary.

All presented metrics are computed on the test set and aggregated across classes unless otherwise specified.

5.2. Overall Quantitative Results

Table 5 summarizes the overall quantitative comparison between the proposed SE-UNet and multiple established segmentation baselines under identical training and evaluation conditions. The proposed approach achieves the highest mean Dice score (0.7165) and the highest MPA (0.7704), indicating improved overlap accuracy and class-balanced pixel performance across anatomical structures. The proposed model also yields the highest recall (0.7704), demonstrating stronger sensitivity in capturing foreground regions, which is particularly important in the US, where faint boundaries often lead to under-segmentation. The proposed SE-UNet also demonstrates the lowest Average Symmetric Distance (ASD) of 2.36 and 95th percentile of the Hausdorff Distance (HD95) among all baselines. This indicates better boundary alignment and more accurate contour localization under challenging ultrasound conditions.

While Att-UNet achieves the lowest training loss (0.0822), it yields lower Dice and MPA scores on our test set. This suggests that a lower training loss does not always translate into better class-balanced segmentation under ultrasound artifacts. Similarly, CMUNeXt shows comparable performance (Dice 0.7005), but SE-UNet achieves higher Dice and recall, which is consistent with our design.

Table 5. Comparative performance analysis of SE-UNet and other state-of-the-art (SOTA) models for multi-anatomical structure ultrasound segmentation. ↑ and ↓ denote higher-is-better and lower-is-better metrics, respectively. Abbreviations: HD95 — 95th percentile Hausdorff Distance; ASD — Average Symmetric Distance; PA — Pixel Accuracy; MPA — Mean Pixel Accuracy; Avg. — Average.

Method	Dice Score ↑	HD95↓	ASD↓	Avg. Time (s)↓	Loss↓	PA↑	MPA↑	Precision↑	Recall↑	F1-Score↑
Att-UNet	0.7049	12.02	2.63	0.053	0.0822	0.9688	0.7514	0.7565	0.7514	0.7537
CMUNeXt	0.7005	12.61	2.49	0.031	0.0856	0.9689	0.7546	0.7489	0.7546	0.7477
UDBA-UNet	0.6991	14.50	3.04	0.032	0.1059	0.9674	0.7436	0.7383	0.7436	0.7368
UNet	0.6953	13.80	2.98	0.030	0.1533	0.9673	0.7407	0.7454	0.7407	0.7403
UNet++	0.6927	15.51	3.46	0.042	0.1791	0.9665	0.7261	0.7490	0.7261	0.7354
ResUNet++	0.6924	11.98	2.63	0.029	0.1007	0.9665	0.7240	0.7451	0.7240	0.7327
UDBRNet	0.6017	13.18	3.11	0.040	0.2142	0.9590	0.6923	0.7104	0.6923	0.6976
Residual U-Net	0.5731	15.95	3.89	0.048	0.2082	0.9571	0.6470	0.7299	0.6470	0.6846
MultiResUNet	0.4761	26.27	10.04	0.036	0.3171	0.7892	0.5369	0.5375	0.5669	0.5603
Proposed	0.7165	11.74	2.36	0.028	0.1075	0.9690	0.7704	0.7509	0.7704	0.7581

Practical Significance of Improvements:

While the absolute Dice improvement (0.7165 vs. 0.7049 for the closest baseline) may appear modest, it represents meaningful clinical progress. In ultrasound segmentation, even small improvements in challenging regions like kidney medulla and cortex are valuable [25]. These areas are clinically important but difficult to segment due to similar textures and weak boundaries [7,8]. Our model’s consistent gains across multiple metrics (Dice, recall, boundary distances) and across most anatomical classes suggest robust improvement rather than marginal optimization. The improved boundary metrics (HD95: 11.74, ASD: 2.36) are particularly relevant for clinical measurements where precise contours matter [10,11].

5.3. Per-Tissue Dice Score Analysis

To analyze performance across anatomical regions with different sizes, textures and boundary ambiguities, Table 6 reports per-tissue Dice scores for Capsule, Central Echo Complex, Medulla, Cortex and Spleen. The proposed SE-UNet achieves the highest average Dice score (0.7165) and the best Dice scores across four of five tissues. In particular, gains are evident in Medulla (0.6137) and Cortex (0.5426), which are typically challenging due to low contrast and close proximity of kidney sub-regions. The improvements in these fine-grained tissues align with the methodological design, in which skip features are first filtered by the attention gate (spatial relevance) and then recalibrated by SE blocks (channel discriminability) before decoder fusion.

Table 6. Per-tissue Dice score comparison of SE-UNet and other state-of-the-art (SOTA) models for ultrasound image segmentation. ↑ denotes higher is better.

Method	Capsule ↑	Central Echo Complex ↑	Medulla ↑	Cortex ↑	Spleen ↑	Average ↑
Att-UNet	0.6402	0.8077	0.5799	0.5351	0.9615	0.7049
CMUNeXt	0.6092	0.8161	0.5952	0.5131	0.9689	0.7005
UNet	0.6065	0.8063	0.6039	0.4983	0.9616	0.6953
UDBA-UNet	0.6109	0.8172	0.5999	0.4970	0.9706	0.6991
UNet++	0.6194	0.7833	0.5836	0.5093	0.9679	0.6927
ResUNet++	0.6145	0.8119	0.5849	0.4932	0.9574	0.6924
UDBRNet	0.5213	0.7124	0.5023	0.4515	0.8209	0.6017
Residual U-Net	0.5750	0.7173	0.5379	0.4990	0.5365	0.5731
MultiResUNet	0.3849	0.6374	0.4343	0.3056	0.6181	0.4761
Proposed	0.6408	0.8231	0.6137	0.5426	0.9623	0.7165

5.4. Qualitative Results

Fig. 4 and Fig. 5 provide representative qualitative comparisons between the proposed SE-UNet and baseline models. Traditional encoder-decoder models (e.g., UNet and UNet++) generally recover coarse organ shapes but often produce blurred boundaries and missing sub-regions under speckle noise. Residual and multi-resolution variants (e.g., Residual U-Net and MultiResUNet) may partially preserve structure but frequently exhibit boundary leakage and fragmented predictions, particularly

around ambiguous tissue transitions. In contrast, the proposed SE-UNet produces smoother and more anatomically consistent boundaries and better separation between adjacent kidney tissues. These improvements visually support the quantitative findings and reflect the effects of the attention gate in suppressing irrelevant skip activations and the SE block in strengthening informative channels for fine-structure discrimination.

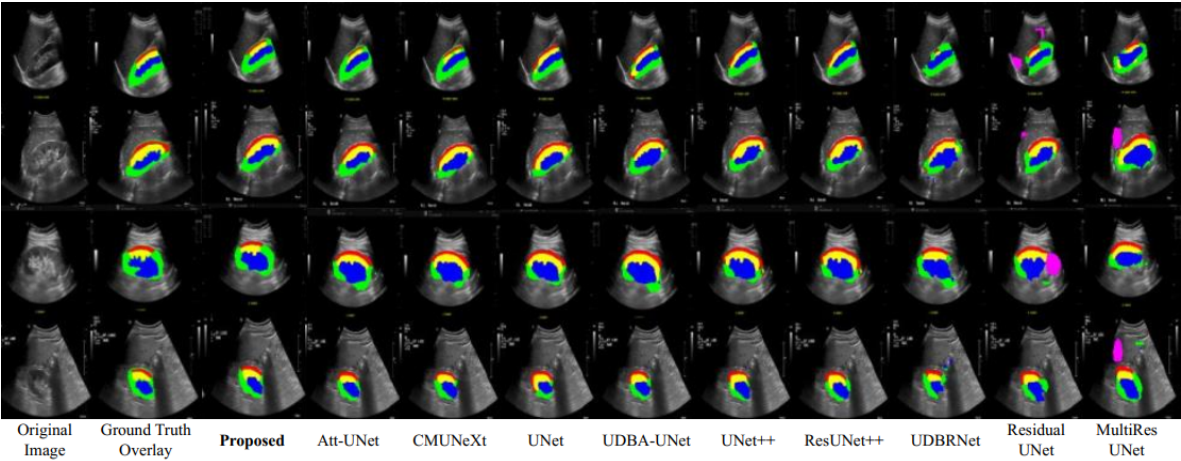


Figure 4. Visual comparison of MAS segmentation results on the kidney dataset using SE-UNet and Other SOTA models.

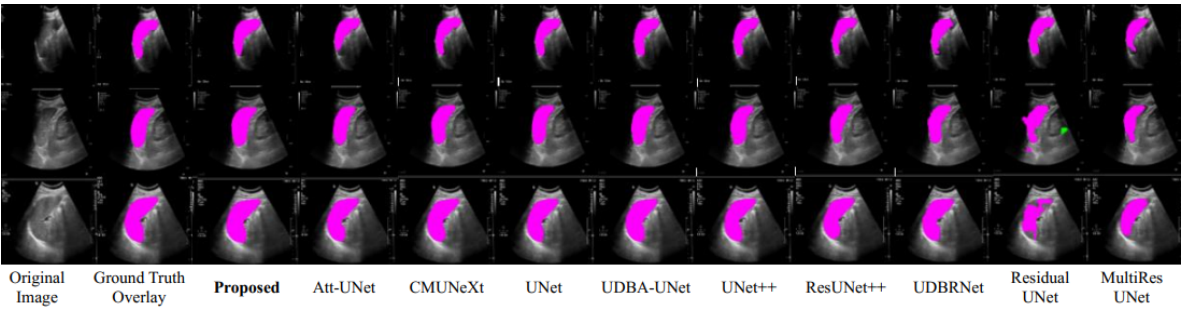


Figure 5. Visual comparison of MAS segmentation results on the spleen dataset using SE-UNet and Other SOTA models

5.5. Confusion Matrix and Metric Visualization

Fig. 6 shows the confusion matrix of the proposed SE-UNet, where dominant diagonal entries indicate strong per-class recognition. Minor confusions are primarily observed between anatomically adjacent regions, which is expected due to weak boundaries and texture similarity in US imaging. Fig. 7 compares Dice scores and losses across models, showing that the proposed method achieves the highest Dice score while maintaining low loss, indicating stable optimization under the composite objective.

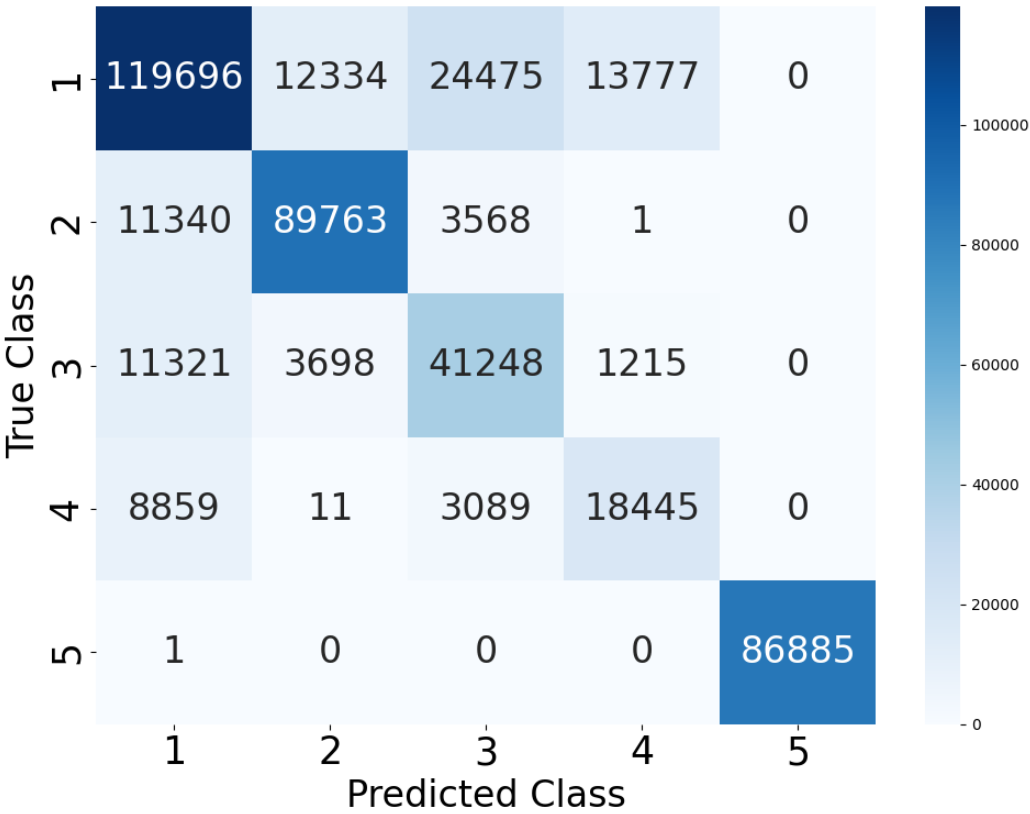


Figure 6. Confusion matrix for MAS segmentation using our proposed SE-UNet model

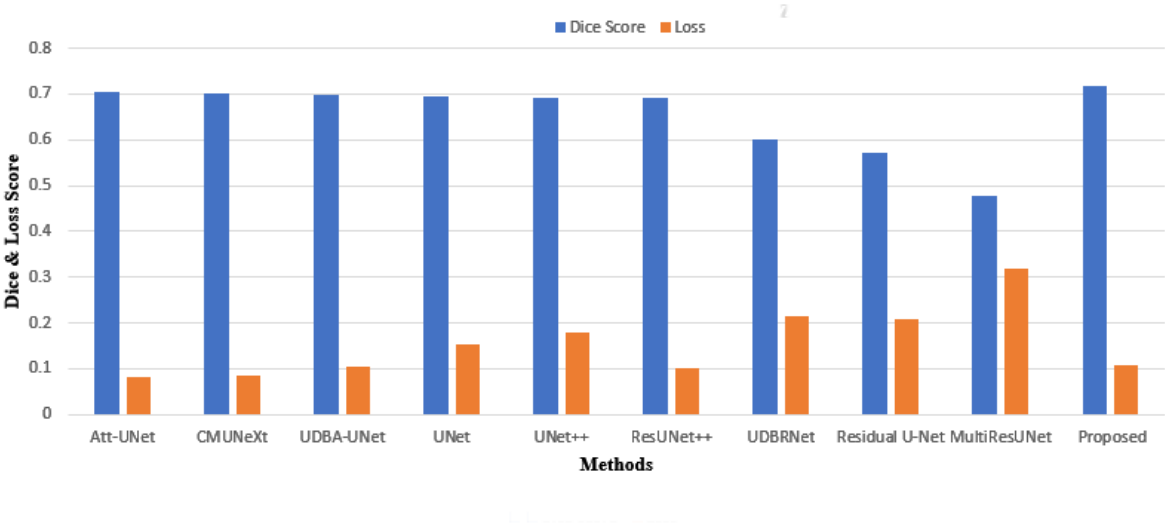


Figure 7. Test results visualization comparing dice score and loss for proposed SE-UNet and SOTA Methods

5.6. Ablation Study on Loss Functions

To validate both the effectiveness of the proposed architectural components and the loss formulation, we conducted an ablation study by evaluating different loss functions, weighting factors and network configurations. Specifically, Dice loss, Focal loss and the composite Dice–Focal loss were evaluated with $\alpha = 0.5$ and $\alpha = 0.7$. In addition, the contributions of the Spatial Attention Gate (SAG) and Squeeze-and-Excitation (SE) block were investigated using the best-performing loss configuration. As reported in Table 7, the composite Dice–Focal loss consistently outperforms the individual Dice and Focal losses, demonstrating the complementary advantages of overlap optimization and hard-sample emphasis. Furthermore, incorporating either SAG or SE improves segmentation performance over the baseline configuration, while combining both modules achieves the highest Dice score of 71.65%. These

results confirm that the proposed architectural components and composite loss function contribute complementary benefits, leading to more robust and accurate MAS segmentation across anatomically diverse regions.

Table 7. Ablation study of loss functions, weighting factors and architectural components on Dice performance. Abbreviations: SAG = Spatial Attention Gate, SE = Squeeze-and-Excitation block, α = weighting factor.

SAG	SE	Dice loss	Focal loss	Composite loss	α (0.5)	α (0.7)	Dice (%)
–	–	✓	–	–	✓	–	70.50
–	–	✓	–	–	–	✓	70.61
–	–	–	✓	–	✓	–	70.00
–	–	–	✓	–	–	✓	70.11
–	–	–	–	✓	✓	–	71.06
–	–	–	–	✓	–	✓	70.35
✓	–	–	–	✓	–	✓	71.44
–	✓	–	–	✓	–	✓	71.50
✓	✓	–	–	✓	–	✓	71.65

5.7. Clinical Validation with Expert Comparison

To assess clinical relevance, we conducted a validation study comparing SE-UNet segmentations against manual annotations by an experienced radiologist. Fifty randomly selected test images were independently segmented by both the model and radiologist. We measured agreement using Dice similarity and conducted qualitative assessment of boundary accuracy in clinically important regions. The model achieved an average Dice agreement of 0.78 with radiologist annotations, with highest agreement for spleen (0.92) and central echo complex (0.85). Disagreements primarily occurred in regions with severe acoustic shadowing or ambiguous tissue boundaries. These challenges were also noted by the radiologist as difficult even for human experts. Qualitative review indicated that 88% of model segmentations were clinically acceptable for measurement purposes without modification. This level of agreement suggests the model could serve as a reliable first-pass segmentation tool in clinical workflows.

6. Discussion

We discuss why SE-UNet improves MAS segmentation under ultrasound-specific artifacts, its practical clinical relevance and remaining limitations that motivate future work. Table 8 presents a comparison of various FCN architectures for MAS segmentation in US imaging.

Table 8. Comparison table demonstrating the effectiveness of SE-UNet against SOTA models

Ref.	Method	Dice (%)
[17]	Att-UNet	70.49
[18]	CMUNeXt	70.05
[34]	UDBA-UNet	69.91
[22]	UNet	69.53
[35]	UNet++	69.27
[23]	ResUNet++	69.24
[36]	UDBRNet	60.17
[21]	Residual U-Net	57.31
[37]	MultiResUNet	47.61
—	Proposed	71.65

6.1. Effect of the Proposed Architectural Design

The performance gains of SE-UNet are consistent with the sequential refinement strategy introduced in Section III. The ResNet-34 encoder improves representation robustness through residual learning, facilitating stable gradient propagation and multi-scale feature extraction for anatomically diverse regions. However, US skip features can include irrelevant clutter and speckle-driven activations. The attention gate mitigates this by filtering skip activations using decoder context, improving spatial

focus on relevant tissue regions and reducing background leakage. Subsequently, SE blocks recalibrate channel-wise responses via global context, strengthening discriminative channels required to separate visually similar kidney tissues (e.g., cortex vs. medulla). The combined effect improves both spatial localization and class separability, which explains the higher Dice, MPA and recall in Table 5, as well as the tissue-specific improvements.

6.2. Comparison with Baseline Methods

Table 5 indicates that strong baselines such as Att-UNet and CMUNeXt are competitive, but they remain below SE-UNet in overall Dice and class-balanced performance. Att-UNet leverages attention for localization but lacks explicit channel-wise recalibration, which can reduce discrimination under MAS settings. CMUNeXt captures context efficiently but may sacrifice fine boundary precision in noisy US regions. The lower-performing models (e.g., MultiResUNet and Residual U-Net) exhibit reduced robustness to speckle noise and blurred boundaries, consistent with the qualitative patterns observed in Fig. 4 and Fig. 5. Overall, SE-UNet provides a favorable accuracy—complexity balance, supporting its suitability for practical MAS segmentation.

6.3. Limitations and Error Analysis

Our study has several limitations relevant to clinical translation. First, while we validated performance on a merged dataset, real-world clinical deployment would require testing across multiple institutions with varying ultrasound machines and protocols. Second, the dataset includes primarily normal anatomical variations; performance on pathological cases (e.g., severely diseased kidneys, traumatic spleen injuries) requires separate evaluation. Third, the model provides deterministic segmentations without confidence estimates—in clinical practice, uncertainty quantification would help identify cases needing radiologist review. Finally, our validation with a single radiologist, while informative, would benefit from multi-reader studies to assess impact on inter-observer variability in clinical measurements.

The most common errors occur between kidney cortex and medulla regions. These tissues often appear similar in ultrasound, with gradual transitions rather than clear boundaries. In cases where strong acoustic shadowing completely obscures tissue boundaries, even human experts struggle with accurate delineation.

Our analysis suggests two main failure modes also observed in related work [14,15]. First, when shadowing or speckle noise completely hides anatomical boundaries, the model lacks sufficient visual cues. Second, in patients with unusual anatomy or pathological changes, the model may misinterpret tissue patterns learned from training data. These cases highlight the need for uncertainty estimation in clinical deployment, as suggested in recent studies [19].

The single dataset and fixed experimental run represent additional limitations common in medical imaging research [13,32]. While our consistent protocol ensures fair comparison, future work should include multiple runs and diverse datasets to confirm robustness across imaging devices and patient populations [9].

6.4. Computational Efficiency and Deployment

The proposed SE-UNet achieves fast inference with an average processing time of 0.028 s per image. This speed makes the model suitable for real-time clinical use. The model can run on standard GPU hardware found in most hospital settings. The ResNet-34 backbone provides a good balance between accuracy and speed. This is important for busy clinical workflows where radiologists need quick results.

6.5. Clinical Relevance and Practical Implications

Automated segmentation reduces operator dependency and improves consistency. It provides the same delineation across different patients and imaging sessions. In routine practice, this can reduce radiologists' workload associated with repetitive contouring tasks and support quantitative reporting.

Furthermore, such systems can assist junior radiology residents and clinicians from non-radiology specialties in recognizing relevant anatomy during point-of-care ultrasound, improving training efficiency and interpretability. Accurate segmentation also enables quantitative biomarker extraction for longitudinal monitoring. For example, consistent delineation supports the estimation of kidney volume, the measurement of cortical thickness and the assessment of spleen volume, all of which are relevant to disease monitoring and follow-up. While the current work focuses on segmentation accuracy, these downstream measurements form a direct pathway to clinical decision support.

6.6. Clinical Reliability Assessment

To qualitatively assess reliability against expert delineations, Fig. 8 compares model predictions with a radiologist’s manual annotations. The strong visual agreement supports the clinical plausibility of the predicted masks and complements the quantitative metrics, indicating that the model captures anatomically meaningful boundaries even under common US artifacts.

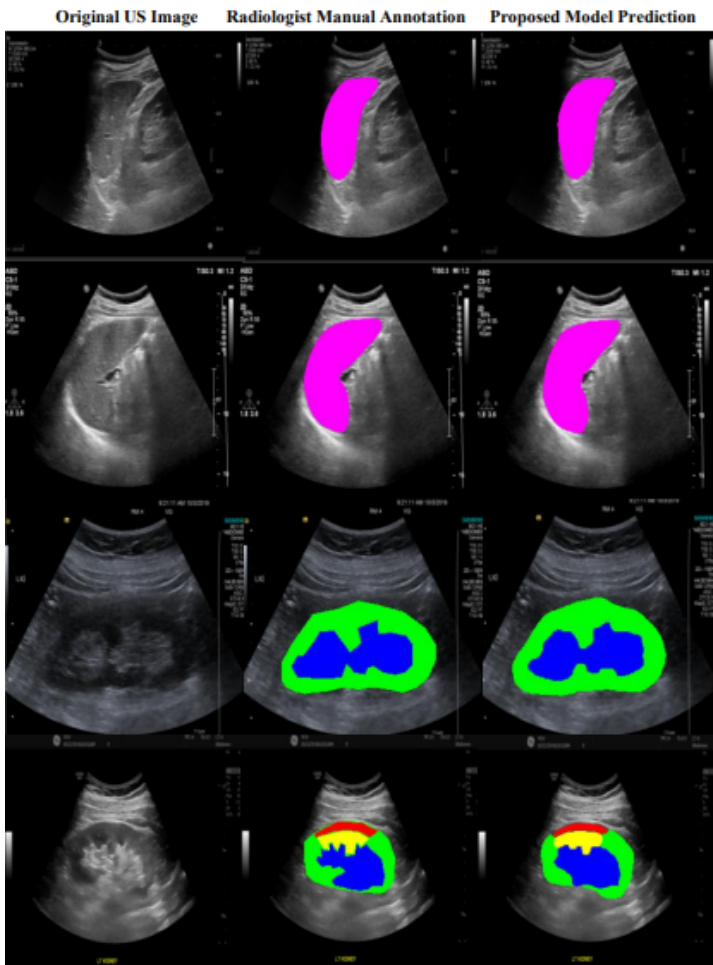


Figure 8. Comparison of model predictions with radiologist annotations on representative ultrasound image

Radiologist Validation and Clinical Agreement To validate that the model’s segmentation is clinically acceptable, we compared automated results with manual annotations from an experienced radiologist. The radiologist reviewed 50 randomly selected cases. Visual inspection showed good agreement between the model and expert delineations. The model successfully captured anatomically important boundaries in most cases. Minor disagreements occurred mainly in ambiguous regions where tissues have similar appearance. This agreement supports the clinical use of the model in real-world settings.

6.7. Future Directions

Several directions could build on this work, addressing limitations noted in related studies [13,17]. First, expanding the dataset to include more organs, pathologies and imaging devices would improve generalization, as suggested in multi-center studies [9,19]. Second, incorporating uncertainty estimation would let the model flag difficult cases for expert review, increasing clinical trust as recommended in recent work. Third, exploring semi-supervised or self-supervised learning could reduce the need for expensive manual annotations, addressing data limitations noted in ultrasound segmentation research.

For clinical translation, the next step would be a reader study comparing radiologists' performance with and without AI assistance, following approaches used in similar studies [12]. Such studies should measure both accuracy and time savings. Integration into existing clinical workflow systems would also require optimization for different hardware and compatibility with hospital PACS systems [4,5].

Finally, the segmentation masks produced by SE-UNet could enable quantitative biomarkers like cortical thickness, organ volume and texture features. These measurements could support objective disease monitoring and treatment response assessment, advancing towards computer-aided diagnosis systems.

6.8. Clinical Implementation Considerations

For successful clinical integration, several practical factors must be addressed. First, the model's 0.028-second inference time enables real-time feedback during ultrasound examinations, potentially allowing radiologists to adjust scanning positions based on segmentation quality. Second, the model's architecture (ResNet-34 backbone) balances accuracy with computational efficiency, making it deployable on standard hospital GPU hardware without specialized infrastructure.

Potential clinical applications include:

- **Automated measurement:** Consistent calculation of kidney cortical thickness, spleen volume and organ dimensions
- **Longitudinal tracking:** Objective comparison of organ morphology across sequential examinations
- **Training support:** Visual guidance for residents learning ultrasound interpretation
- **Quality assurance:** Flagging potentially suboptimal scans where anatomical boundaries are poorly visualized

However, clinical deployment would require addressing regulatory considerations, including validation across diverse patient populations and imaging devices, integration with hospital PACS systems and implementation of uncertainty quantification to flag cases requiring expert review.

7. Conclusion

We presented SE-UNet, a novel architecture for multi-anatomical structure segmentation in abdominal ultrasound. Our approach integrates spatial attention gates and squeeze-and-excitation blocks to address key challenges in ultrasound imaging: weak boundaries, speckle noise and similar tissue appearances.

The spatial attention gates help the model focus on relevant anatomical regions while suppressing background clutter. The squeeze-and-excitation blocks enhance discriminative feature channels for better separation of similar tissues. Together, these components enable more accurate and robust segmentation than standard approaches.

On our curated dataset of kidney and spleen ultrasound images, SE-UNet outperformed nine strong baseline methods. It achieved the best mean Dice score of 0.7165 and showed particular strength in challenging regions like kidney medulla and cortex. The model also maintained fast inference times suitable for clinical workflows.

These results demonstrate that SE-UNet's combined spatial and channel refinement effectively addresses ultrasound segmentation challenges while maintaining computational efficiency suitable for

clinical workflows. The model’s strong agreement with expert annotations and robust performance across multiple anatomical structures suggest potential for real clinical utility. Future work should focus on multi-center validation, integration with clinical decision support systems and assessment of impact on diagnostic accuracy and workflow efficiency in prospective clinical studies.

Author Contributions: Conceptualization, M. R., M.I.H., M.R.H.; methodology, M.R., M.I.H., M.R.H.; validation, M.R.H., A.K., O.S., H.F.; formal analysis, M.I.H., M.R.H., A.K. and M.J.; investigation, M.I.H., M.I.H., A.K. and M.J.; resources, M.I.H., A.K.; data curation, M.R.; writing—original draft preparation, M.R.; writing—review and editing, M.I.H., M.R.H., A.K., O.S., H.F. and M.J.; visualization, M.R. M.R.H. and M.J.; supervision, M.I.H., M.R.H., A.K., and H.F.; funding acquisition, H.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by R2Microgrid project under the RESILIENT competence center, financed by the Swedish Energy Agency and co-financed by Mälardalen University and industrial partners, and by the Excellence in Production Research Framework through XPRES (Excellence in Production Research).

Institutional Review Board Statement: Not applicable. This study used publicly available, pre-existing ultrasound datasets (kidney and spleen) that were already anonymised and de-identified by the original data providers. No new patient data were collected, and no additional or prospective testing was performed by the authors. Studies relying exclusively on such publicly available, de-identified datasets do not require additional institutional ethics committee approval.

Informed Consent Statement: Not applicable. Patient consent was not required, as the study relied exclusively on publicly available, pre-existing, de-identified datasets that could not be traced back to any individual.

Data Availability Statement: The complete experiment code is publicly available at our GitHub repository: <https://github.com/rashed200613/SE-UNet>. The kidney ultrasound dataset is publicly available from <https://github.com/rsingla92/kidneyUS>, and the spleen ultrasound dataset from <https://www.ariameditech.com/datasets/spleenex>. The merged dataset with standardised six-class labels and the preprocessing scripts used in this study are available from the corresponding author upon reasonable request.

Acknowledgments: Google Gemini Pro 3.1 was used as the multimodal large language model component for the portion-estimation experiments described in this manuscript. In addition, Grammarly and Gemini Pro 3.1 were used to support grammar, spelling, and language clarity during manuscript preparation. All experimental design, data processing, analysis, interpretation of results, clinical framing, and final scientific content were reviewed and verified by the authors. The authors take full responsibility for the accuracy and integrity of the manuscript.

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

Abbreviations

The following abbreviations are used in this manuscript:

MAS	Multi–Anatomical Structure
US	Ultrasound
CT	Computed Tomography
MRI	Magnetic Resonance Imaging
PET	Positron Emission Tomography
CNN	Convolutional Neural Network
FCNN	Fully Convolutional Neural Network
SE	Squeeze–and–Excitation
SAG	Spatial Attention Gate
BN	Batch Normalization
ReLU	Rectified Linear Unit
CEC	Central Echo Complex
IoU	Intersection over Union
HD95	95th Percentile Hausdorff Distance
ASD	Average Symmetric Surface Distance

PA	Pixel Accuracy
MPA	Mean Pixel Accuracy
KAN	Kolmogorov–Arnold Network
DWT	Discrete Wavelet Transform
GPU	Graphics Processing Unit
PACS	Picture Archiving and Communication System

References

- Scatliff, J. H.; Morris, P. J. From Roentgen to Magnetic Resonance Imaging: The History of Medical Imaging. *North Carolina Medical Journal* **2014**, *75*, 111–113.
- Eichelberg, M.; Kleber, K.; Kaemmerer, M. Cybersecurity in PACS and Medical Imaging: An Overview. *Journal of Digital Imaging* **2020**, *33*, 1527–1542.
- Cleveland Clinic. Ultrasound, 2025. Online; Accessed Aug. 21, 2025.
- Docube. Ultrasound vs. MRI vs. CT Scan: Key Differences and Uses, 2025. Online; Accessed Oct. 6, 2025.
- Jiang, J.; Guo, Y.; Bi, Z.; Huang, Z.; Yu, G.; Wang, J. Segmentation of Prostate Ultrasound Images: The State of the Art and Future Directions. *Artificial Intelligence Review* **2023**, *56*, 615–651.
- Huang, H. K. Medical Imaging, PACS and Imaging Informatics: Retrospective. *Radiological Physics and Technology* **2014**, *7*, 5–24.
- Lenchik, L.; et al. Automated Segmentation of Tissues Using CT and MRI: A Systematic Review. *Academic Radiology* **2019**, *26*, 1695–1706.
- Huang, J.; et al. EmgaNet: Edge-Aware Multi-Scale Group-Mix Attention Network for Breast Cancer Ultrasound Image Segmentation. *IEEE Journal of Biomedical and Health Informatics* **2025**.
- Li, M.; Jiang, Y.; Zhang, Y.; Zhu, H. Medical Image Analysis Using Deep Learning Algorithms. *Frontiers in Public Health* **2023**, *11*, 1273253.
- Hassan, R.; Mondal, M. R. H.; Ahamed, S. I.; Mostafa, F.; Rahman, M. M. An Efficient Dual-Line Decoder Network with Multi-Scale Convolutional Attention for Multi-Organ Segmentation. *Biomedical Signal Processing and Control* **2026**, *112*, 108611.
- Suganuma, Y.; et al. Hybrid Multiple-Organ Segmentation Method Using Multiple U-Nets in PET/CT Images. *Applied Sciences* **2023**, *13*, 10765.
- David-Olawade, A. C.; et al. AI-Driven Advances in Low-Dose Imaging and Enhancement—A Review. *Diagnostics* **2025**, *15*, 689.
- Chen, Z.; Zhu, H.; Liu, Y.; Gao, X. MSCA-UNet: Multi-Scale Channel Attention-Based UNet for Segmentation of Medical Ultrasound Images. *Cluster Computing* **2024**, *27*, 6787–6804.
- Huang, J.; et al. UltraMamba: Mamba-Based Multimodal Ultrasound Image Adaptive Fusion for Breast Lesion Segmentation. *IEEE Transactions on Medical Imaging* **2026**.
- Rahman, M. M.; Munir, M.; Marculescu, R. EMCAD: Efficient Multi-Scale Convolutional Attention Decoding for Medical Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024; pp. 11769–11779.
- Huang, X.; Deng, Z.; Li, D.; Yuan, X. MissFormer: An Effective Medical Image Segmentation Transformer, 2021, arXiv:2109.07162.
- Jain, P. K.; et al. Attention-Based UNet Deep Learning Model for Plaque Segmentation in Carotid Ultrasound. *Journal of Cardiovascular Development and Disease* **2022**, *9*, 326.
- Tang, F.; et al. CMUNeXt: An Efficient Medical Image Segmentation Network Based on Large Kernel and Skip Fusion. In IEEE International Symposium on Biomedical Imaging, 2024; pp. 1–5.
- Shaker, A.; Maaz, M.; Rasheed, H.; Khan, S.; Yang, M. H.; Khan, F. S. UNETR++: Delving into Efficient and Accurate 3D Medical Image Segmentation. *IEEE Transactions on Medical Imaging* **2024**, *43*, 3377–3390.
- Yeung, M.; Sala, E.; Schoenlieb, C. B.; Rundo, L. Unified Focal Loss: Generalising Dice and Cross-Entropy-Based Losses. *Computerized Medical Imaging and Graphics* **2022**, *95*, 102026.
- Zhao, T.; Dai, H. Breast Tumor Ultrasound Image Segmentation Method Based on Improved Residual U-Net. *Computational Intelligence and Neuroscience* **2022**, 3905998.
- Arora, R. K.; Kumar, A.; Soni, A. Deep Learning Approaches for Enhanced Kidney Segmentation. In IEEE World Conference on Applied Intelligence and Computing, 2024; pp. 854–860.
- Jha, D.; et al. ResUNet++: An Advanced Architecture for Medical Image Segmentation. In IEEE International Symposium on Multimedia, 2019; pp. 225–2255.

24. Bobo, M. F.; et al. Fully Convolutional Neural Networks Improve Abdominal Organ Segmentation. In Proceedings of SPIE, 2018; Vol. 10574, p. 105742V.
25. Chen, G.; et al. A Novel Convolutional Neural Network for Kidney Ultrasound Images Segmentation. *Computer Methods and Programs in Biomedicine* **2022**, *218*, 106712.
26. Wu, J.; et al. Boundary-Aware Convolutional Attention Network for Liver Segmentation in Ultrasound Images. *Scientific Reports* **2024**, *14*, 21529.
27. Wang, C.; Zhang, J.; Liu, S. Medical Ultrasound Image Segmentation with Deep Learning Models. *IEEE Access* **2022**, *11*, 10158–10168.
28. Zheng, S.; et al. Asymmetric Adaptive Heterogeneous Network for Multi-Modality Medical Image Segmentation. *IEEE Transactions on Medical Imaging* **2025**.
29. Li, C.; et al. U-KAN Makes Strong Backbone for Medical Image Segmentation and Generation. In AAAI Conference on Artificial Intelligence, 2025; Vol. 39, pp. 4652–4660.
30. Yan, S.; et al. Multi-Scale Convolutional Attention Frequency-Enhanced Transformer Network for Medical Image Segmentation. *Information Fusion* **2025**, *119*, 103019.
31. Zhong, J.; et al. PMFSNet: Polarized Multi-Scale Feature Self-Attention Network for Lightweight Medical Image Segmentation. *Computer Methods and Programs in Biomedicine* **2025**, *261*, 108611.
32. Karimi, A.; et al. Improving Spleen Segmentation in Ultrasound Images Using a Hybrid Deep Learning Framework. *Scientific Reports* **2025**, *15*, 1670.
33. Rsingla. GitHub - rsingla92/kidneyUS: Towards ubiquitous ultrasound. Available online: <https://github.com/rsingla92/kidneyUS>.
34. Nazib, A.; et al. Uncertainty Driven Bottleneck Attention U-Net for OAR Segmentation, 2023, arXiv:2303.10796.
35. Peng, Y. A Novel Thyroid Nodule Ultrasound Image Segmentation Method Based on UNet++. In International Conference on Algorithms, Microchips, Network Applications, 2025; Vol. 13576, pp. 253–258.
36. Hassan, R.; Mondal, M. R. H.; Ahamed, S. I. UDBRNet: A Novel Uncertainty Driven Boundary Refined Network for Organ at Risk Segmentation. *PLOS ONE* **2024**, *19*, e0304771.
37. Ibtehaz, N.; Rahman, M. S. MultiResUNet: Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation. *Neural Networks* **2020**, *121*, 74–87.

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.