

A Standardized Comparative Benchmark of Fully Convolutional and Hybrid Deep Networks for Multi-Anatomical Structure Segmentation in Ultrasound Imaging

Md. Rashed^a, Md. Riad Hassan^{b,e,*}, Md. Imran Hossain^{a,*}, Md Rakibul Islam^{c,d}, Feroza Naznin^b and Md Zahidul Islam^c

^aDepartment of Information and Communication Engineering, Pabna University of Science and Technology, Pabna, 6600, Bangladesh

^bDepartment of Computer Science and Engineering, Green University of Bangladesh, Dhaka, Bangladesh

^cDepartment of Information and Communication Technology, Islamic University, Kushtia, Bangladesh

^dDept of Computer Science and Engineering, Bangladesh University of Business and Technology, Dhaka, Bangladesh

^eIICT, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh

ARTICLE INFO

Keywords:

Ultrasound Image Segmentation
Multi-anatomical Structure Segmentation
Deep Learning
Fully Convolutional Networks
Kidney and Spleen Ultrasound
Medical Imaging

ABSTRACT

Ultrasound image segmentation remains challenging because of speckle noise, low contrast, shadowing artifacts, and poorly defined anatomical boundaries. These difficulties are amplified when models are required to generalize across heterogeneous anatomical structures and datasets. This study presents a standardized comparative benchmark of fully convolutional, attention-based, and hybrid deep learning architectures for multi-anatomical ultrasound segmentation. A harmonized dataset was constructed by integrating the Open Kidney Ultrasound and Spleenex datasets, yielding 984 B-mode images with pixel-level annotations for kidney tissue structures and spleen regions. Nine representative architectures—UNet, UNet++, Attention UNet, ResUNet++, Residual UNet, CMUNeXt, UCTransNet, MISSFormer, and EMCAD—were trained and evaluated under identical preprocessing, augmentation, optimization, and loss settings. Experimental results showed that CMUNeXt achieved the best overall performance, with a Dice score of 0.7791 and IoU of 0.6323, while requiring only 3.14M trainable parameters. Per-class and area-wise analyses revealed that large structures such as the spleen were segmented reliably, whereas small and low-contrast kidney tissue regions remained difficult. These findings suggest that efficient multi-scale architectures offer a favorable accuracy–complexity trade-off for ultrasound segmentation, while robust segmentation of small anatomical structures remains an open challenge.

1. Introduction

Ultrasound (US) radiology remains a highly clinician-dependent imaging modality, relying predominantly on the radiologist's visual interpretation and diagnostic expertise to accurately recognize organ structures, identify lesions and tissue types, assess bone structures, and perform precise anatomical measurements Klivanov and Hossack (2015). This manual approach offers the advantage of leveraging clinical experience and contextual understanding, allowing radiologists to adapt their assessments based on patient-specific variations MedScience (2025). However, indistinct tissue margins and the presence of speckle noise in US images make manual contouring difficult, often resulting in considerable intra- and inter-observer variation, thereby underscoring the necessity for automated segmentation Jiang, Guo, Bi, Huang, Yu and Wang (2023). Automatic tissue awareness in US imaging can significantly assist radiologists by providing consistent and objective identification of anatomical structures, thereby reducing observer variability and enhancing segmentation reproducibility Yang, Chen,

Liang, Bai, Wang, Zhao, Ma et al. (2023). Ultrasound-specific artifacts, including speckle noise and boundary ambiguity, require advanced segmentation algorithms that can effectively model both fine-grained local features and broader contextual information. Xiao, Zhang, Shao, Liu, Shi, He and Kong (2025), Duque, Marquardt, Velikova, Lacourpaille, Nordez, Crouzier, Lee, Mateus and Navab (2024).

Multi-organ segmentation in US imaging has major therapeutic benefits since it allows for automated localization of organs such as the kidney, liver, and spleen, which aids in illness diagnosis, treatment planning, and longitudinal monitoring while reducing operator reliance. It also allows for quantitative assessment of organ size, texture, and pathology, which improves diagnostic accuracy and workflow efficiency. However, research in this subject still confronts significant shortcomings. US images contain speckle noise and poor contrast, making border recognition difficult; Organ shape variability and overlapping anatomical structures introduce additional complexity to the segmentation task. Existing models, which are frequently trained on restricted and single-center datasets, struggle to generalize across patient groups, diseases, and imaging technologies. Furthermore, there is a scarcity of large-scale, consistent, and well-annotated multi-organ US datasets, making technique benchmarking and comparison difficult.

*Corresponding author

 rashedulislam.ice.pust@gmail.com (Md. Rashed);

riad@cse.green.edu.bd (Md.R. Hassan); imran05ice@pust.ac.bd (Md.I. Hossain); rakibuliuict@gmail.com (M.R. Islam); feroza@cse.green.edu.bd (F. Naznin); zahidimage@gmail.com (M.Z. Islam)

ORCID(s):

While convolutional neural networks (CNNs), including U-Net and its variants, have demonstrated strong performance in segmenting individual organs, their effectiveness in multi-organ ultrasound segmentation is limited due to organ shape variability, overlapping structures, and heterogeneous tissue appearance. These challenges highlight the need for architectures capable of capturing both local details and contextual relationships across multiple anatomical regions. Litjens, Kooi, Bejnordi, Setio, Ciompi, Ghafoorian, Laak, Ginneken and Sánchez (2017), Chen, Li, Dai, Zhang and Yap (2022a).

Interpreting a full-body ultrasound scan is a highly challenging task, requiring simultaneous recognition of multiple organs, tissue types, and anatomical structures, which can vary widely across patients. Manual assessments are time-consuming, subjective, and prone to inter- and intra-observer variability, even for experienced radiologists. Recent advances in deep learning have shown promise in automating organ and tissue segmentation by capturing both global anatomical context and fine-grained tissue details. Motivated by these developments, this study constructs a harmonized kidney–spleen ultrasound benchmark and evaluates nine representative architectures using the same preprocessing, augmentation, loss function, optimizer, and evaluation metrics Wang, Wu and Liu (2025), Chen, Cai, Wang, Chen, Zhang, Han, Guo, Ding and Zhang (2024a), Liu, Qu, Xie, Shi and Song (2024).

This paper addresses the following research questions:

RQ1) How do conventional CNN, attention-based, and hybrid architectures compare under a standardized ultrasound segmentation protocol?

RQ2) Which anatomical structures are most difficult to segment, and how does performance vary across tissue classes?

RQ3) How does anatomical region size affect segmentation performance?

RQ4) Which models provide the best accuracy–complexity trade-off for potential resource-constrained ultrasound applications?

The objectives of this study are: (i) *To conduct a comprehensive comparative analysis* of representative fully convolutional and hybrid deep learning architectures for multi-anatomical structure segmentation in ultrasound (US) images. (ii) *To curate and unify a standardized ultrasound dataset* by integrating the Open Kidney US and Spleenex datasets, resulting in 984 B-mode images with harmonized pixel-level annotations for five tissue types to ensure fair and reproducible benchmarking. (iii) *To apply consistent preprocessing, label unification, and data augmentation techniques* in order to establish a standardized experimental setting for objective model comparison. (iv) *To evaluate nine representative segmentation models* (UNet, UNet++, Att-UNet, ResUNet++, Residual UNet, CMUNeXt, UCTransNet, MISSFormer, and EMCAD) under identical training configurations using a combined loss function. (v) *To analyze the strengths, limitations, and practical implications* of these architectures and to provide clear future research directions for advancing multi-anatomical ultrasound segmentation.

The main contributions of this study are summarized as follows:

1. **Harmonized Benchmark Construction:** We integrate the OKUS and Spleenex datasets into a unified kidney–spleen ultrasound segmentation benchmark through standardized preprocessing and consistent label mapping.
2. **Fair multi-model comparison:** We evaluate nine representative architectures under identical data splits, augmentation, loss function, optimizer, and training schedule.
3. **Accuracy–complexity analysis:** We compare segmentation performance jointly with parameter count to identify efficient models suitable for resource-limited settings.
4. **Class-wise and size-wise evaluation:** We analyze per-tissue Dice scores and area-wise segmentation behavior to identify failure modes, especially for small anatomical structures.
5. **Qualitative feature analysis:** We use visual predictions and t-SNE embeddings to examine model behavior beyond aggregate metrics.

This study offers a thorough evaluation of deep learning models for multi-anatomical US image segmentation, providing valuable insights into their performance, limitations, and future research directions, particularly emphasizing the need for improved handling of small structures and robust noise filtering.

The remainder of this paper is organized as follows. Section 2 reviews related work on ultrasound segmentation and attention/efficiency-aware architectures. Section 3 presents the materials and methods, including study design, dataset description and preprocessing, training and evaluation protocol, and comparative baseline. Section 4 details the experimental results, both quantitative and qualitative. Section 5 discusses the findings, conclusion, and future directions.

2. Related Works

Recent advances in deep learning have substantially improved ultrasound image segmentation. Several studies have explored some innovative approaches to improve the US medical image segmentation.

2.1. U-Net Variants

The MSCA-UNet model was proposed by Chen et al. (2024) Chen, Zhu, Liu and Gao (2024b) for the segmentation of medical ultrasound images. In order to modify the channel weights using the channel attention method, the model has a multi-scale module that creates feature maps at different scales. The information at the edges is preserved using the global feature module. On several datasets, this network enhanced the IoU and dice score. These studies recognize the challenges posed by noisy ultrasound images. It is quite difficult to remove the borders of the lesions because of their lack of contrast and inconsistent grayscale. Additionally, there was a very tiny dataset of ultrasound images used.

The Attention and Unet models are specifically made for kidney segmentation MRI data with good boundary delineation. This architecture's features are extremely pertinent to the US medical picture segmentation challenge. However, it is limited in its ability to capture the organ boundary's low contrast and thick edge. Additionally, in US imaging, segmentation accuracy varies according to organ size Amiri, Brooks, Behboodi and Rivaz (2020).

A study Arora, Kumar and Soni (2024) on kidney segmentation in MRI, relevant due to similar boundary delineation challenges, evaluated U-Net and Attention U-Net with standard and focal cross-entropy losses, achieving a Dice coefficient of 0.966. Its emphasis on precise border segmentation applies to multi-organ US imaging, where overlapping organ structures create challenges.

In cardiovascular US imaging, an Attention-UNet model for carotid plaque segmentation achieved a correlation coefficient of 0.96 and an AUC of 0.97, surpassing U-Net, UNet++, and UNet3P models. Its ability to segment bright and fuzzy plaques highlights its potential for multi-organ US imaging, where variable tissue textures complicate segmentation Jain, Dubey, Saba, Khanna, Laird, Nicolaides and Sharma (2022).

Debesh Jha et al. Jha, Smedsrud, Riegler, Johansen, De Lange, Halvorsen and Johansen (2019) suggest ResUNet++ is an improved version of ResUNet designed for accurate polyp segmentation in colonoscopic images. It outperforms U-Net and ResUNet, achieving high Dice and mIoU scores on Kvasir-SEG and CVC-612 datasets. Its strong performance in handling complex tissue boundaries makes it promising for multi-organ US segmentation.

2.2. Hybrid Models (Convolutional and Transformer-based)

EMCAD Rahman, Munir and Marculescu (2024) architecture introduces an efficient attention decoder with various attention mechanisms for reducing the model complexity to segment the multi-organ on ultrasound images. It shows the segmentation accuracy of 83.63%, which can perform better than other state-of-the-art methods for multi-organ medical US image segmentation. But this study does not show performance on real-world clinical validation. This research also does not explain how the decoder can overcome the common clinical data challenges.

The CMUNeXt Tang, Ding, Quan, Wang, Ning and Zhou (2024) model is proposed to extract the global context by using the inductive bias. This architecture leverages a large kernel and an inverted bottleneck to mix distant spatial and local information thoroughly, which extracts the global context information efficiently. The CMUNeXt demonstrates that the outperforms state-of-the-art heavy-weight and lightweight networks of medical images. This architecture is utilized by the small size of medical data.

The study Chen, Yin, Dai, Zhang, Yin and Cui (2022b) fully convolutional neural network (FCNN), originally for CT scans, was successfully adapted to T2-weighted MRI for multi-organ segmentation, showing high Dice scores and

outperforming traditional methods. Its strong performance without modality-specific retraining highlights its robustness and potential for multi-organ US segmentation despite imaging challenges.

Another study Wu, Liu, Sun, Liu, Hou, Jiang and Zhang (2024) proposed a novel multi-scale, deep-supervised CNN for kidney segmentation in US images achieved high accuracy (Dice: 95.86%) by combining multi-scale inputs, pyramid pooling, and refined prediction through deep supervision. Its strong performance and robustness to imaging challenges make it well-suited for multi-organ US segmentation. BACANet Huang, Deng, Li and Yuan (2021) is a lightweight deep learning model designed for real-time liver US segmentation, combining efficiency with high accuracy. It uses convolutional attention, large kernel modules, and boundary supervision to enhance feature extraction and edge detection.

Tianyu Zhao et al. Wang, Zhang and Liu (2022) proposed A U-Net-based method for breast tumor US image segmentation integrates residual blocks and a fusion attention mechanism to enhance accuracy and stability. The residual blocks prevent gradient vanishing, while the attention mechanism improves feature extraction by focusing on spatial and channel information.

The MISSFormer Zhao and Dai (2022) and ConvTransNet Zhou, Hou, Lu, Zhao, Dang and Dong (2022) methods leverages the transformer-based architecture with CNN technologies to capture the global and local contexts effectively. These architectures potentially handled complex anatomical structures in US imaging. These architectures can not address the problem of feature discrimination limitation.

3. Materials and Methods

3.1. Study Design

This study is designed as a systematic comparative evaluation of deep learning architectures for multi-anatomical structure US segmentation. Nine representative models—UNet, UNet++, Att-UNet, ResUNet++, Residual UNet, CMUNeXt, UCTransNet, MISSFormer, and EMCAD—were selected to benchmark their performance under consistent experimental conditions. To ensure fair comparison, all models were trained and evaluated using identical protocols, including the same dataset splits, preprocessing steps, and optimization strategy.

A multi-anatomical US dataset was curated by integrating the Open Kidney Ultrasound (OKUS) and Spleenex datasets, comprising a total of 984 B-mode ultrasound images annotated for five tissue types: capsule, cortex, medulla, central echo complex, and spleen. The dataset was standardized through intensity normalization and harmonized annotations to enable reproducible benchmarking. Data augmentation techniques, including random rotation, horizontal and vertical flips, and intensity scaling, were applied during training to increase robustness against ultrasound-specific variations and improve generalization.

Table 1

Summary of ultrasound segmentation datasets used in this study.

Dataset	Organ	Images	Annotation Classes	Image Format
OKUS	Kidney	534	Capsule, cortex, medulla, Central Echo Complex	PNG/B-mode
Spleenex	Spleen	450	Spleen	PNG/B-mode
Combined	Kidney + spleen	984	5 foreground classes + background	256×256 input

All models were trained using a Dice–Focal composite loss function to address class imbalance and optimize segmentation accuracy. The Adam optimizer was used with an initial learning rate of 0.001. Training was performed for 400 epochs with a batch size of 8, and early stopping was applied based on validation loss to prevent overfitting.

All experiments were conducted on a high-performance workstation equipped with an NVIDIA GeForce RTX 4090 GPU with 24 GB of memory, ensuring efficient training and reproducibility. This study design provides a controlled framework to systematically evaluate model performance, identify strengths and limitations, and guide future research in multi-organ ultrasound segmentation.

3.2. Dataset Description and Preparation

The Open Kidney Ultrasound (OKUS) Singla, Ringstrom, Hu, Lessoway, Reid, Ngan and Rohling (2023) Dataset, a comprehensive, high-quality collection of kidney B-mode ultrasound images. Compiled over five years, the dataset includes scans from over 500 adult patients with common conditions such as diabetes mellitus, immunoglobulin A nephropathy, and hypertension. It encompasses both native and transplanted kidneys and features detailed multi-class semantic segmentation annotations generated by two expert sonographers. The cohort has an average age of 53.2 ± 14.7 years, with 63% male and 37% female participants. The majority of images (91.2%) are of fair or good quality. The Spleenex Dataset Karimi, Seraj, Mirzadeh Sarcheshmeh, Fazli, Seraj, Eslami, Khanmohamadi, Sajjadian Moosavi, Ghattan Kashani, Sajjadian Moosavi and Shariat Panahi (2025) comprises 450 anonymized US images aimed at advancing spleen segmentation for medical diagnostics, particularly for conditions like splenomegaly. Collected using Philips Affinity 70 and Supersonic Aixplorer devices, the images are converted from DICOM to 512×512 PNG format, with intensity normalized to a $[0, 1]$ range for consistency. Ground truth segmentation Masks were meticulously crafted using 3DSlicer software through a collaborative effort, involving two radiologists and four computer engineering students, with radiologists' verification ensuring accuracy. The dataset strictly adheres to ethical standards, with patient anonymity preserved and informed consent obtained.

The combined dataset was constructed as a harmonized multi-anatomical benchmark rather than a same-frame multi-organ dataset. Kidney images contain capsule, cortex, medulla, and central echo complex annotations, whereas spleen images contain spleen annotations. The unified label

space enables evaluation of model robustness across heterogeneous ultrasound anatomical targets, although not all anatomical classes are present in each image. Table 1 represents the summary of ultrasound segmentation datasets.

Images were resized to 256×256 pixels using bilinear interpolation, while segmentation masks were resized using nearest-neighbor interpolation to preserve discrete labels. Pixel intensities were normalized to $[0, 1]$ using dataset-specific mean and standard deviation values to ensure consistent intensity ranges and stable training. Since some architectures expected three-channel input, greyscale images were replicated across three channels. All ground-truth labels were converted to one-hot encoded masks to support multi-class learning consistency. To enhance model robustness and mitigate overfitting, synchronized data augmentation was applied exclusively to the training subset, including random rotations up to $\pm 15^\circ$, horizontal flips, and vertical flips with a 50% probability each, while preserving perfect spatial correspondence between anatomical structures and their labels; validation and test images were left unaugmented. The dataset comprises K anatomical classes including background, however, the background class was deliberately excluded from all quantitative metric computations so that reported performance scores reflect segmentation accuracy solely over the $K - 1$ foreground anatomical structures of clinical interest. This standardized preprocessing pipeline ensures dataset uniformity, increases anatomical diversity, and establishes a reproducible foundation for comparative evaluation of deep learning models in multi-anatomical structure ultrasound segmentation.

Figure 1 illustrates the distribution of the multi-anatomical US dataset. The outer ring shows the contribution of the OKUS and Spleenex datasets, while the inner ring represents tissue types: Capsule (green), Central Echo Complex (blue), Medulla (yellow), Cortex (red), and Spleen (magenta). Percentages indicate relative proportions.

3.3. Training and Evaluation Protocol

The training of the model was carried out using the Adam optimizer with an initial learning rate of 0.001. A batch size of 8 was used, and training was conducted for 400 epochs. We also employed early stopping based on the validation loss to mitigate overfitting. The dataset was partitioned into training, validation, and test subsets comprising 654, 165, and 165 images, respectively. The validation set was used for model selection and early stopping. Final performance was reported only on the held-out test set.

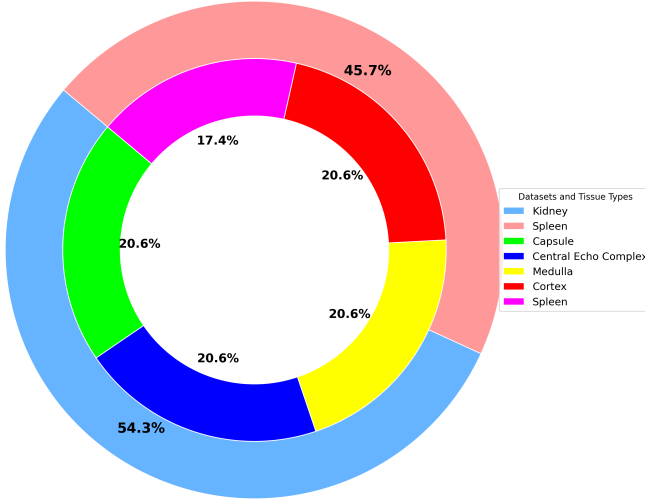


Figure 1: Dataset and tissue type distribution in the Multi-Anatomical US Dataset.

Table 2 represents the hyperparameter settings used for the comparative evaluation.

3.3.1. 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

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

where ϵ is a small constant for numerical stability.

The focal loss is given by

$$\mathcal{L}_{\text{Focal}} = -\frac{1}{HW} \sum_{i,j} \sum_{k=1}^K Y_{i,j,k} (1 - \hat{Y}_{i,j,k})^\gamma \log(\hat{Y}_{i,j,k}), \quad (2)$$

where γ controls the focusing strength.

The final training loss is computed as

$$\mathcal{L} = \alpha \mathcal{L}_{\text{Dice}} + (1 - \alpha) \mathcal{L}_{\text{Focal}}, \quad (3)$$

with $\alpha = 0.7$ in all experiments.

Dice loss improves overlap optimization under class imbalance, whereas focal loss emphasizes difficult pixels and minority classes

3.3.2. Comparative Baseline

To evaluate the performance of deep learning architectures on multi-anatomical structure ultrasound segmentation, nine representative models were selected as comparative baselines. These models span both fully convolutional and hybrid designs, incorporating attention mechanisms and multi-scale feature representations.

- **UNet:** A classical encoder-decoder architecture with skip connections that preserves spatial information for precise segmentation. It serves as a widely used baseline for biomedical image segmentation.
- **UNet++:** An extension of UNet that introduces nested and dense skip connections to improve feature propagation and model accuracy, particularly for fine-grained structures.
- **Att-UNet:** Integrates attention gates within the UNet framework to focus on relevant regions, enhancing the segmentation of small or complex anatomical structures.
- **ResUNet++:** Combines residual connections with UNet++'s nested skip connections, improving gradient flow and learning efficiency for deeper networks.
- **Residual U-Net:** Incorporates residual blocks into the UNet architecture to facilitate training of deeper networks and improve feature learning.
- **CMUNeXt:** A lightweight, multi-scale convolutional architecture designed for efficient segmentation, balancing high accuracy with low computational cost.
- **UCTransNet:** A hybrid convolution-transformer architecture that leverages both local convolutional features and global context modeling via transformer blocks.
- **MISSFormer:** A transformer-based architecture emphasizing multi-scale self-attention, capable of capturing long-range dependencies and anatomical context.
- **EMCAD:** An ensemble-based multi-scale convolutional approach integrating channel and spatial attention, designed for robust segmentation of heterogeneous structures.

These models were selected to provide a comprehensive benchmark, covering classical CNNs, attention-enhanced networks, and transformer-based architectures, enabling systematic analysis of their strengths, limitations, and suitability for multi-anatomical structure segmentation in US imaging.

4. Results

This section presents the quantitative and qualitative performance of the selected deep learning models for multi-anatomical structure segmentation in ultrasound (US) images, following the experimental protocol described in the Materials and Methods. All performance was reported only on the held-out test set, using identical preprocessing, training configurations, and the Dice–Focal composite loss function, ensuring a fair and consistent comparison across all models.

Table 2: Training and hyperparameter settings used for the comparative evaluation of deep learning models in multi-anatomical structure US segmentation.

Setting	Value
Input size	256×256
Batch size	8
Epochs	400
Optimizer	Adam
Initial learning rate	0.001
Loss	0.7 Dice + 0.3 Focal
Focal gamma	2.0
Early stopping	Validation loss
Random seed	2222
Augmentation	Rotation, flips, intensity scaling

4.1. Evaluation Metrics

Segmentation performance was evaluated using standard pixel-wise and region-based measures to thoroughly assess classification accuracy and anatomical segmentation standard. Below is the list of utilized metrics list:

- **Dice Similarity Coefficient (DSC):** A region-based overlap metric measuring segmentation quality per class. Higher values indicate better agreement between predicted and ground-truth masks.
- **Intersection over Union (IoU):** A stricter region-based overlap metric defined as the ratio of the intersection to the union of predicted and ground-truth masks. IoU penalises both over- and under-segmentation more severely than DSC and serves as a complementary accuracy indicator.
- **Loss:** The Dice–Focal composite loss (Eq. 3) used during training to optimise segmentation accuracy and address class imbalance. Lower values indicate greater training stability and reduced prediction error.
- **Standard Deviation (SD):** Reported alongside each mean metric as $\text{mean} \pm \text{SD}$. A lower SD indicates more consistent per-sample performance, which is clinically important for reliable deployment across diverse anatomical presentations.
- **95% Confidence Interval (CI):** Computed as $\text{mean} \pm 1.96 \times \frac{\text{SD}}{\sqrt{N}}$ for $N = 165$ test samples. The CI quantifies uncertainty in the estimated mean and enables direct comparison of performance ranges across models.
- **Statistical Significance (p -value):** Pairwise differences relative to the best-performing model (CMUNeXt) were assessed using a paired t -test on per-sample DSC scores. Significance thresholds are denoted as $*p < 0.05$, $**p < 0.01$, and $***p < 0.001$.
- **Number of Parameters (M):** The total number of trainable parameters in millions, reflecting model

complexity and computational cost for practical deployment.

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

4.2. Overall Quantitative Results

Table 3 presents a comprehensive quantitative evaluation of nine deep-learning segmentation architectures on the harmonised multi-anatomy ultrasound benchmark, comprising kidney and spleen segmentation tasks across $N = 165$ test samples. Model performance is assessed along four complementary axes: Dice Similarity Coefficient (DSC), Intersection over Union (IoU), composite segmentation loss, and the number of trainable parameters (Params). DSC and IoU measure the degree of spatial overlap between predicted and ground-truth masks—higher values indicate better segmentation fidelity—whereas lower loss reflects greater training stability and reduced prediction error. Parameter count serves as a proxy for computational complexity and practical deployability, with smaller models being preferable for resource-constrained clinical environments. Statistical significance of pairwise differences relative to the top-performing model was assessed via paired t -test, with p -values corrected for directional interpretation ($*p < 0.05$, $**p < 0.01$, $***p < 0.001$).

Top-performing model. Among all evaluated architectures, **CMUNeXt** achieves the best performance across every primary metric, recording a mean DSC of 0.7792 ± 0.1897 (95% CI: [0.7503, 0.8081]), a mean IoU of 0.6323 ± 0.1539 (95% CI: [0.6088, 0.6558]), and a mean loss of 0.0814 ± 0.0782 , while maintaining a lightweight architecture of only 3.15 M parameters. These results collectively demonstrate an outstanding accuracy–efficiency trade-off: CMUNeXt delivers state-of-the-art segmentation quality at a parameter budget more than an order of magnitude smaller than several competing models. The narrow 95% confidence intervals further indicate that its performance is consistent across test cases, an important property for reliable clinical deployment.

High-performing architectures. **UCTransNet** and **ResUNet++** constitute the second and third tiers of performance, with mean DSC values of 0.7648 ± 0.1898 ($p = 0.004$, $**$) and 0.7618 ± 0.1900 ($p = 0.014$, $*$), respectively, both statistically significantly lower than CMUNeXt. UCTransNet achieves competitive segmentation accuracy, but at a substantially higher computational cost of 59.75 M parameters—approximately 19× the size of CMUNeXt—raising concerns about practical scalability and inference efficiency. In contrast, ResUNet++ requires only 4.06 M parameters, making it the most parameter-efficient alternative to CMUNeXt, albeit with a slightly higher mean loss (0.0971 ± 0.1052) suggesting somewhat greater prediction variability, particularly in ambiguous boundary regions.

Competitive mid-tier architectures. **UNet++**, **Att-UNet**, and **UNet** occupy a performance cluster with Dice scores in the range of 0.7518–0.7540, IoU values of 0.6066–0.6219, and p -values that do not reach the $p < 0.05$ threshold for UNet++ ($p = 0.267$) or Att-UNet ($p = 0.165$), indicating that their mean performance is not statistically distinguishable from CMUNeXt under the current test set size. UNet, however, does reach statistical significance ($p = 0.026$, *) despite a nearly identical mean Dice of 0.7518 ± 0.2009 , a result explained by its marginally lower standard deviation compared to Att-UNet (0.2009 vs 0.2030), which shifts the paired t -statistic past the significance threshold. This SD-driven divergence in p -values between models of near-identical means is an important methodological nuance: it reflects differences in per-sample consistency rather than absolute accuracy. Notably, UNet++ exhibits a Dice–IoU rank inversion relative to UNet—higher Dice (0.7540 vs 0.7518) but lower IoU (0.6178 vs 0.6219)—suggesting that UNet++ achieves broad regional overlap but with less precise boundary delineation, a distinction relevant in anatomical structures with sharp margins such as the kidney. Furthermore, UNet++ records a disproportionately high loss standard deviation (0.2743 vs mean 0.1590, coefficient of variation = 1.72), indicating marked instability in per-sample loss, which warrants further investigation into its behaviour on challenging ultrasound frames. Att-UNet (34.88 M) and UNet (13.40 M) also carry considerably higher parameter burdens than CMUNeXt for only marginal gains, further underscoring the efficiency advantage of the top-performing model.

Lower-performing architectures. **MISSFormer**, **Residual U-Net**, and **EMCAD** exhibit substantially reduced performance across all metrics, with mean DSC values of 0.7029 ± 0.2261 , 0.6412 ± 0.2220 , and 0.6170 ± 0.2463 , respectively, all significantly inferior to CMUNeXt ($p < 0.001$, ***). Their correspondingly higher loss values—0.1106, 0.1940, and 0.1420—and elevated standard deviations suggest limited robustness to ultrasound-specific acquisition artefacts such as speckle noise, acoustic shadowing, and low tissue contrast, which are inherent challenges of the imaging modality under study. Of particular concern is **EMCAD**, whose mean IoU 0.4975 ± 0.2000 falls below 0.50—the only model in the benchmark to do so—implying that its predictions overlap with the ground-truth mask for less than half the union area on average. Despite having the smallest parameter count (2.58 M), EMCAD delivers the worst segmentation quality, demonstrating that model compactness alone is insufficient without appropriate inductive biases for medical image segmentation. Residual U-Net similarly struggles, with the highest mean loss (0.1940 ± 0.2413) and a coefficient of variation exceeding 1.24, reflecting high instability across test samples. MISSFormer, despite its transformer-based design and a substantial parameter footprint (30.06 M), fails to generalise effectively to the ultrasound domain, suggesting that the attention mechanisms it employs may require domain-specific adaptation to handle

the lower signal-to-noise ratio of ultrasound compared to the CT or MRI modalities on which such architectures were typically developed.

Summary. The results collectively establish three key findings. First, model size is not a reliable predictor of segmentation performance: CMUNeXt (3.15 M) outperforms UC-TransNet (59.75 M), Att-UNet (34.88 M), and MISSFormer (30.06 M) by statistically significant margins. Second, per-sample consistency—as reflected by standard deviation and confidence interval width—is as diagnostically important as mean accuracy, since high-variance models (UNet++, EMCAD, Residual U-Net) are inherently less reliable in clinical workflows where predictable performance across diverse anatomical presentations is required. Third, architectures that combine multi-scale feature extraction with efficient structural design, exemplified by CMUNeXt and ResUNet++, offer the most favourable balance of accuracy, reliability, and computational feasibility for multi-organ ultrasound segmentation, supporting their prioritisation in future clinical translation efforts.

4.3. Per-Tissue Dice Score Analysis

Table 4 presents a per-tissue Dice score comparison of different segmentation models across five anatomical structures in ultrasound imaging, namely Capsule, Central Echo Complex, Medulla, Cortex, and Spleen, along with the average Dice score across all classes. Among all evaluated methods, CMUNeXt achieves the highest average Dice score (0.7791), demonstrating consistently strong performance across all tissue types, particularly excelling in the segmentation of the Central Echo Complex and Spleen, with Dice scores of 0.8359 and 0.9834, respectively. Attention-based and hybrid architectures, including UCTransNet, ResUNet++, Att-UNet, UNet++, and UNet, also exhibit robust segmentation capability, with average Dice scores ranging from 0.7517 to 0.7648, indicating reliable multi-tissue delineation. These models show notably high accuracy in well-defined regions such as the Central Echo Complex (Dice > 0.82) and Spleen (Dice > 0.93), while performance in the Cortex and Medulla remains comparatively lower, reflecting the difficulty of accurately segmenting subtle boundaries and low-contrast regions inherent to ultrasound data. Transformer-based models such as MISSFormer achieve moderate performance with an average Dice score of 0.7028, but show reduced accuracy in more complex tissue regions, suggesting limitations in capturing fine-grained spatial details. In contrast, Residual U-Net and EMCAD yield lower average Dice scores (0.6412 and 0.6170, respectively), indicating challenges in maintaining consistent anatomical delineation across tissue types. Overall, the results confirm that architectures integrating attention mechanisms and multi-scale convolutional features are more effective in handling the heterogeneous and noise-prone characteristics of ultrasound images, leading to improved and more stable multi-anatomical segmentation performance.

Table 3

Performance comparison of segmentation models on the harmonized kidney–spleen benchmark ($N = 165$ test samples). **Bold** values indicate the best performance in each column. Statistical significance was assessed against CMUNeXt using a paired t -test ($*p < 0.05$, $**p < 0.01$, $***p < 0.001$). 95% confidence intervals (CIs) were computed as $\text{mean} \pm 1.96 \times \text{SD} / \sqrt{N}$, where SD denotes standard deviation and N is the number of test samples. Params(M) denotes the number of trainable parameters in millions. $\uparrow$ higher is better; $\downarrow$ lower is better.

Method	Dice mean $\pm$ SD $\uparrow$	Dice 95% CI $\uparrow$	IoU mean $\pm$ SD $\uparrow$	IoU 95% CI $\uparrow$	Loss mean $\pm$ SD $\downarrow$	Params (M)	p -value
CMUNeXt	0.7792 $\pm$ 0.1897	[0.7503, 0.8081]	0.6323 $\pm$ 0.1539	[0.6088, 0.6558]	0.0814 $\pm$ 0.0782	3.15	—
UCTransNet	0.7648 $\pm$ 0.1898	[0.7358, 0.7938]	0.6203 $\pm$ 0.1541	[0.5968, 0.6438]	0.0836 $\pm$ 0.0854	59.75	0.004**
ResUNet++	0.7618 $\pm$ 0.1900	[0.7328, 0.7908]	0.6138 $\pm$ 0.1542	[0.5903, 0.6373]	0.0971 $\pm$ 0.1052	4.06	0.014*
UNet++	0.7540 $\pm$ 0.2117	[0.7217, 0.7863]	0.6178 $\pm$ 0.1720	[0.5916, 0.6440]	0.1590 $\pm$ 0.2743	9.16	0.267
Att-UNet	0.7519 $\pm$ 0.2030	[0.7209, 0.7829]	0.6066 $\pm$ 0.1649	[0.5814, 0.6318]	0.0880 $\pm$ 0.0938	34.88	0.165
UNet	0.7518 $\pm$ 0.2009	[0.7211, 0.7825]	0.6219 $\pm$ 0.1632	[0.5970, 0.6468]	0.1394 $\pm$ 0.1697	13.40	0.026*
MISSFormer	0.7029 $\pm$ 0.2261	[0.6684, 0.7374]	0.5717 $\pm$ 0.1837	[0.5437, 0.5997]	0.1106 $\pm$ 0.0932	30.06	< 0.001***
Residual U-Net	0.6412 $\pm$ 0.2220	[0.6073, 0.6751]	0.5497 $\pm$ 0.1803	[0.5222, 0.5772]	0.1940 $\pm$ 0.2413	13.04	< 0.001***
EMCAD	0.6170 $\pm$ 0.2463	[0.5794, 0.6546]	0.4975 $\pm$ 0.2000	[0.4670, 0.5280]	0.1420 $\pm$ 0.1032	2.58	< 0.001***

Table 4

Per-Tissue Dice Score Comparison of Different Segmentation Methods for Multi-Anatomical Structures in US Imaging.

Method	Capsule	Central Echo Complex	Medulla	Cortex	Spleen	Average
CMUNeXt	0.7207	0.8359	0.6832	0.6724	0.9834	0.7791
UCTransNet	0.7129	0.8433	0.6463	0.6413	0.9801	0.7648
ResUNet++	0.7054	0.8299	0.6661	0.6301	0.9771	0.7617
UNet++	0.7123	0.8197	0.6427	0.6244	0.9709	0.7540
Att-UNet	0.7208	0.8390	0.6399	0.6248	0.9346	0.7518
UNet	0.6946	0.8566	0.6454	0.6296	0.9325	0.7517
MISSFormer	0.6522	0.7653	0.5773	0.5936	0.9258	0.7028
Residual U-Net	0.5517	0.7605	0.6028	0.5892	0.7016	0.6412
EMCAD	0.5711	0.7125	0.5376	0.4907	0.7729	0.6170

4.4. Qualitative Results

Figure 2 illustrates the qualitative segmentation results on the kidney US dataset using baseline and attention-enhanced convolutional architectures. Attention-based models, such as Att-UNet and CMUNeXt, demonstrate strong visual agreement with the ground truth by accurately delineating kidney anatomical boundaries and producing coherent, well-localized tissue regions with minimal leakage. In comparison, conventional architectures such as UNet and UNet++ exhibit slightly irregular boundaries and mild under-segmentation in regions with complex textures and low contrast, indicating limitations in preserving fine structural details in challenging kidney US images.

Figure 3 further evaluates the kidney US dataset, focusing on advanced and hybrid architectures, including UCTransNet, ResUNet++, MISSFormer, Residual UNet, and EMCAD. Transformer-based and hybrid models show improved spatial consistency, enhanced contextual modeling, and refined boundary delineation compared to traditional CNN-based methods. Although occasional boundary distortions and minor misclassifications are observed in highly heterogeneous regions, these models generally achieve superior structural continuity and spatial accuracy, highlighting

the benefits of attention mechanisms and advanced feature fusion strategies for kidney tissue-type segmentation.

Figure 4 presents a qualitative visual comparison of tissue-type segmentation results on the spleen US dataset using additional fully convolutional network (FCN)-based models. While these architectures are capable of capturing the overall spleen anatomy, they often struggle with precise boundary delineation due to speckle noise, weak contrast, and heterogeneous tissue appearance inherent in spleen US images. Fragmented predictions, boundary smoothing, and region leakage are particularly noticeable along organ boundaries. Compared to the attention-enhanced and transformer-based models evaluated on the kidney dataset (Figs. 11 and 12), standard FCN-based approaches demonstrate reduced sensitivity to fine structural details and global contextual information.

Overall, the qualitative comparisons across Figures 2–4 consistently demonstrate that models incorporating attention mechanisms, transformer components, and advanced feature fusion strategies provide more accurate and robust tissue-type segmentation in US imaging. These visual observations strongly complement and support the quantitative results reported in the study for both kidney and spleen datasets.

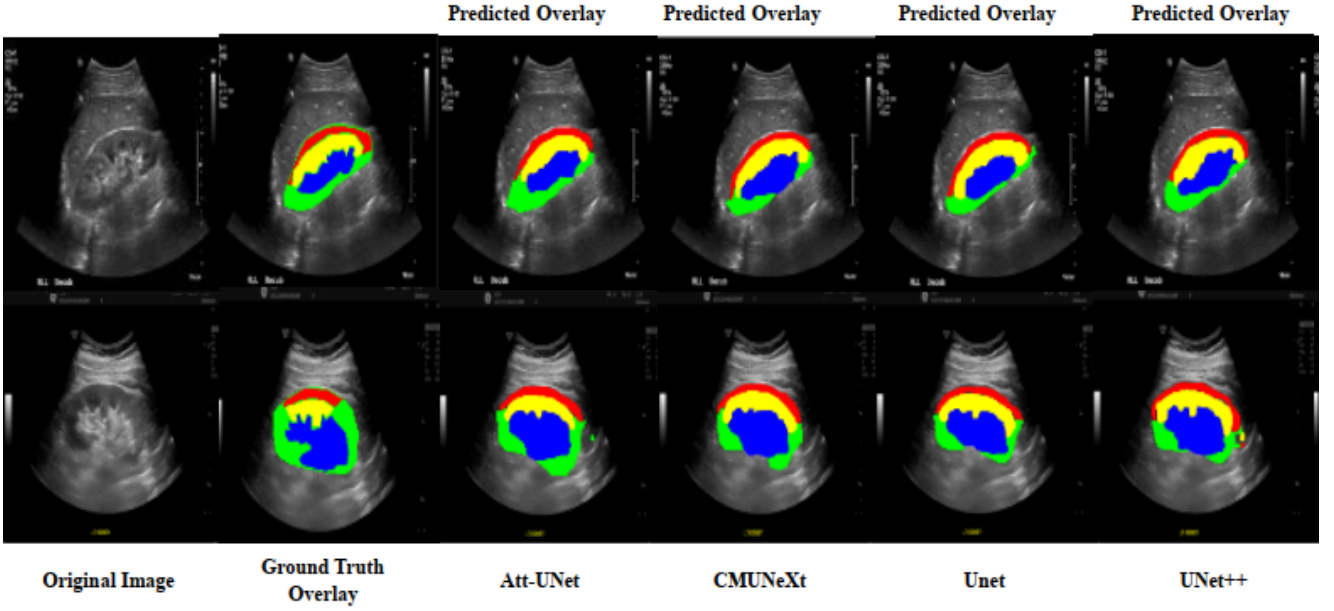


Figure 2: Visual comparison of tissue-type segmentation on kidney scans using Att-UNet, CMUNeXt, UNet, and UNet++. Capsule (green), Central Echo Complex (blue), Medulla (yellow), and Cortex (red) are shown to highlight segmentation differences across models, especially in small and challenging anatomical regions.

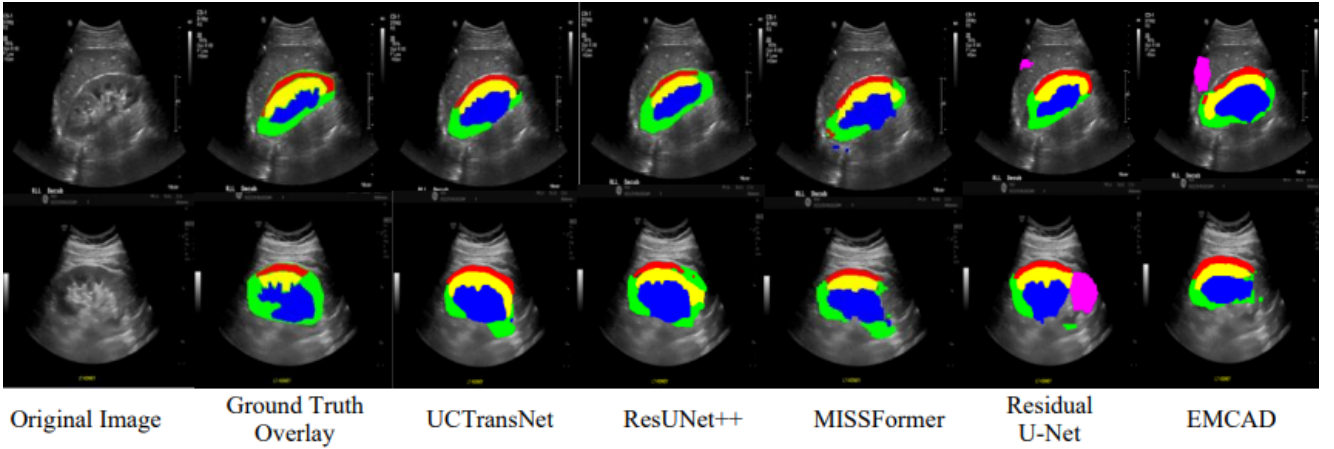


Figure 3: Visual comparison of tissue-type segmentation on kidney scans using UTransNet, ResUNet++, MISSFormer, Residual UNet, EMCAD. Capsule (green), Central Echo Complex (blue), Medulla (yellow), and Cortex (red) are shown to highlight segmentation differences across models, especially in small and challenging anatomical regions.

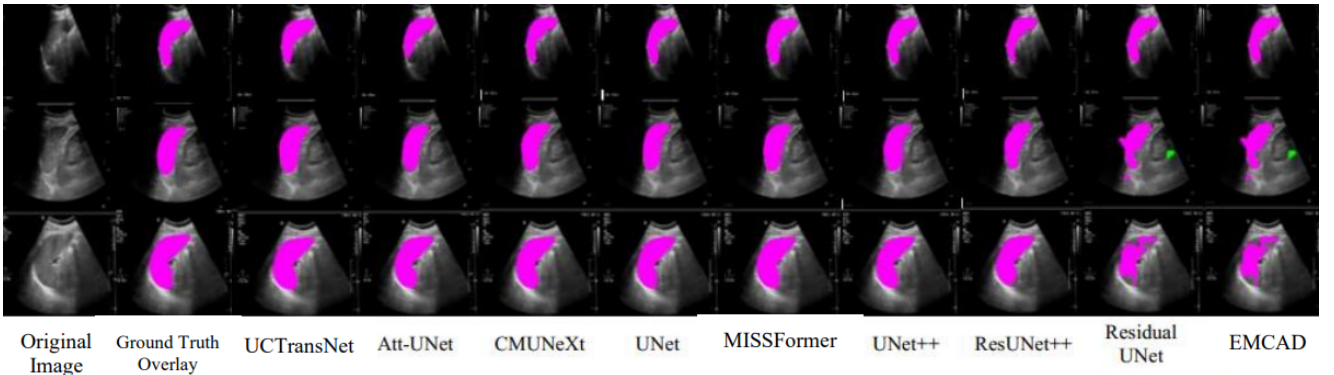


Figure 4: Visual comparison of tissue-type segmentation on spleen scans using fully convolutional networks. Spleen is shown in magenta color to highlight segmentation performance, emphasizing the strengths and limitations of each architecture.

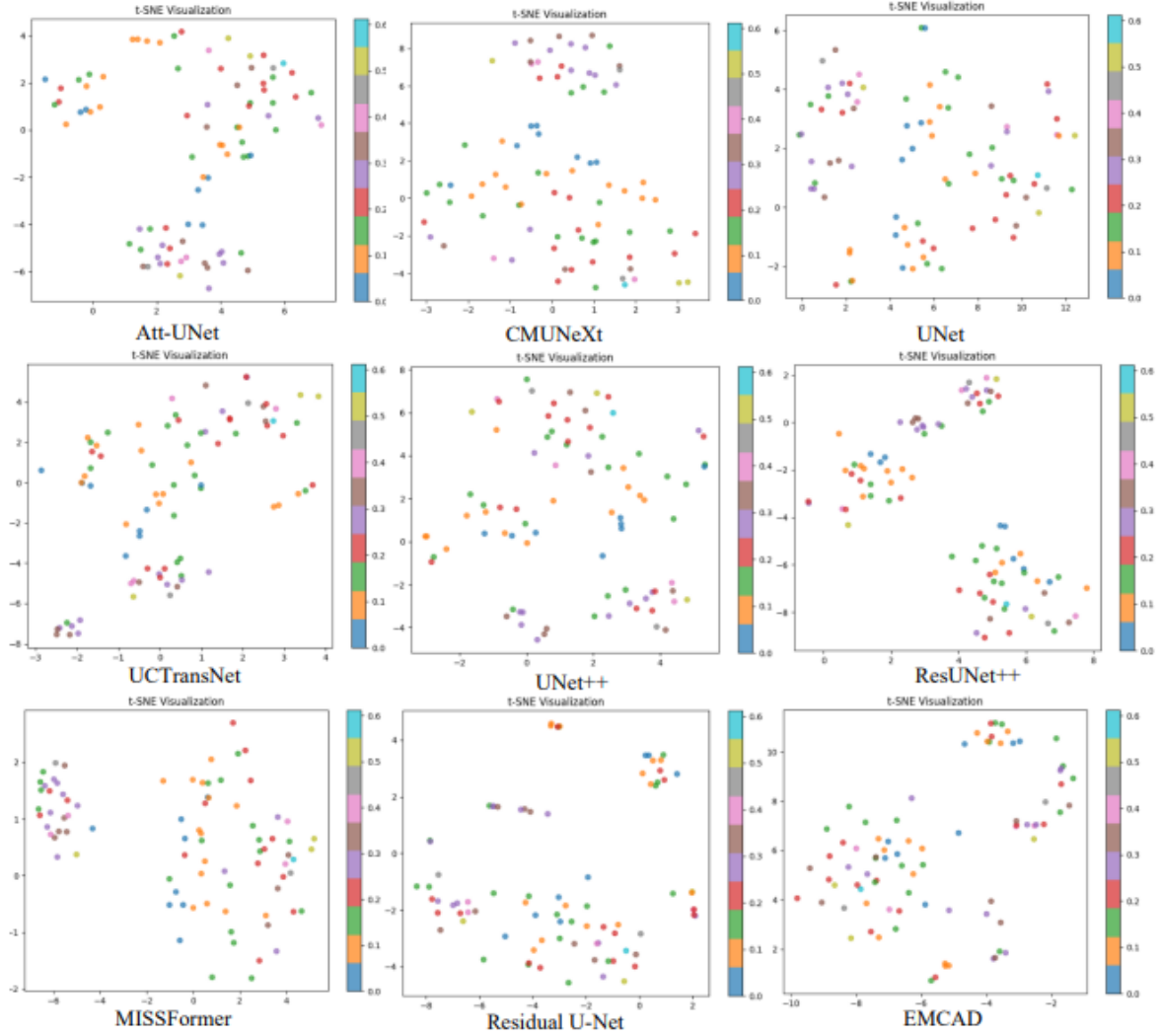


Figure 5: t-SNE visualization of deep feature embeddings learned by different segmentation models.

5. Discussion

The quantitative results presented in Table 3 and the qualitative analyses in Figures 5 and 6 collectively reveal substantive differences in how modern segmentation architectures represent and partition multi-anatomical ultrasound (US) images. Rather than restating these findings, this discussion examines the architectural mechanisms that explain the observed performance hierarchy, the structural properties of US data that make certain image categories persistently difficult, the fundamental trade-off between local texture modelling and global context aggregation, and the practical implications of these results for clinical deployment. Honest limitations are also acknowledged to contextualise the scope of the conclusions drawn.

5.1. Why CMUNeXt Outperforms Competing Architectures

CMUNeXt achieved the highest mean Dice Similarity Coefficient (DSC) of 0.7792 ± 0.1897 , the highest mean IoU of 0.6323 ± 0.1539 , and the lowest mean loss of $0.0814 \pm$

0.0782 , while requiring only 3.15 M trainable parameters—a budget more than 18× smaller than UCTransNet (59.75 M) and nearly 10× smaller than Att-UNet (34.88 M). Its superiority is not attributable to scale alone but to several synergistic architectural properties.

Large-kernel convolution. CMUNeXt employs large-kernel depthwise convolutions, which expand the effective receptive field without the quadratic computational cost of self-attention. In the context of US imaging, where anatomical structures such as the kidney and spleen span substantial fractions of the image plane, a wider receptive field at early feature extraction stages allows the network to capture organ-level shape and boundary context before pooling, reducing the dependence on deep-layer abstraction for spatial reasoning.

Multi-scale feature representation. Through hierarchical feature fusion across multiple resolution stages, CMUNeXt preserves both fine-grained boundary details and coarse structural context simultaneously. This is critical in US segmentation, where the same anatomical region may present

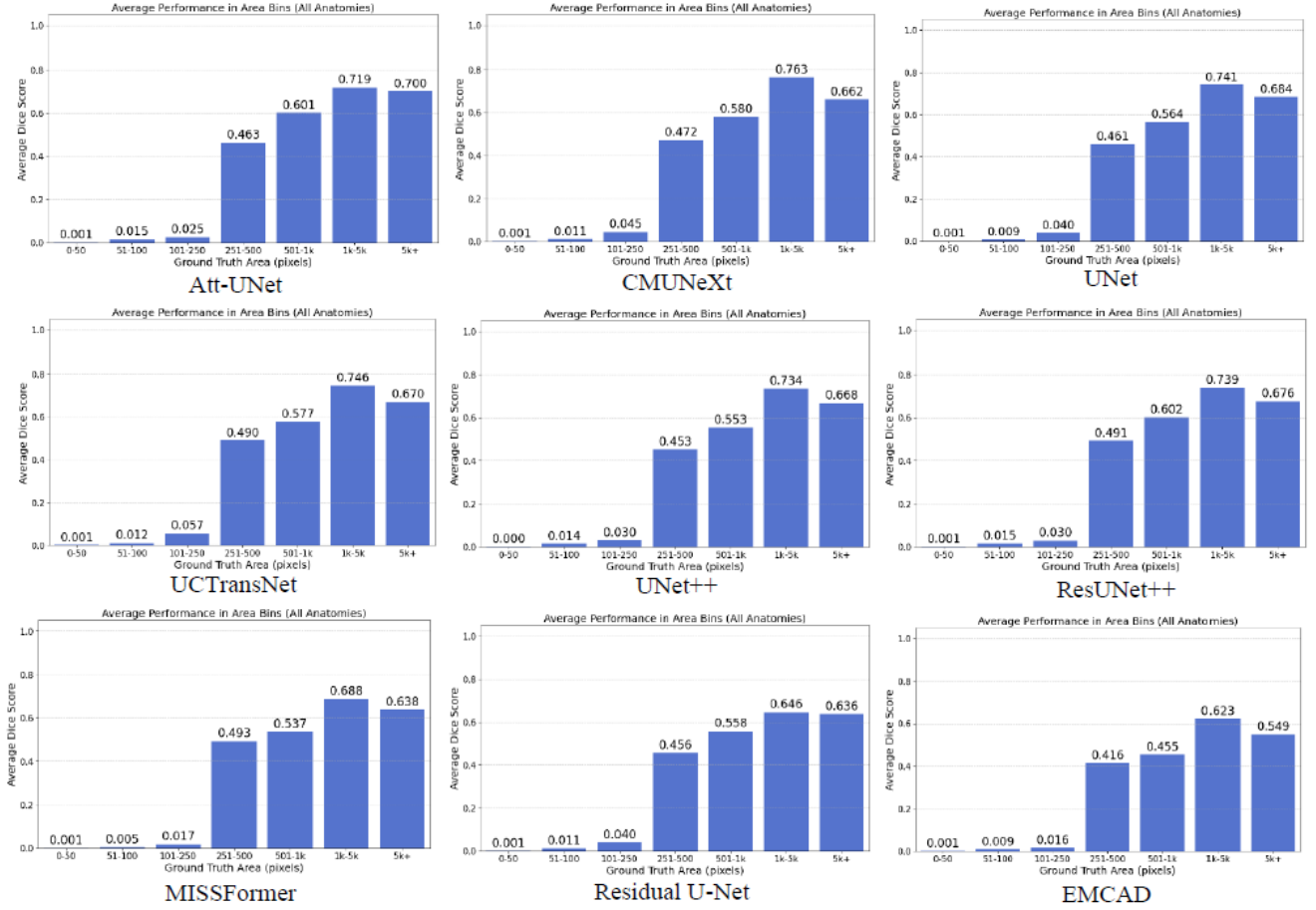


Figure 6: Area-wise segmentation performance of different Fully Convolutional Networks on the multi-anatomical ultrasound dataset.

with varying apparent scale depending on probe position, depth setting, and patient body habitus. The t-SNE visualisations in Figure 5 corroborate this: CMUNeXt embeddings exhibit clear inter-class separation with comparatively tight intra-class clusters, reflecting that its multi-scale representations are semantically consistent across structurally diverse test instances.

Lightweight design and regularisation effect. Compact architectures impose an implicit regularisation that is advantageous when training data are limited or heterogeneous, as is characteristic of multi-anatomy medical imaging benchmarks. The narrow 95% confidence interval of CMUNeXt ([0.7503, 0.8081] for DSC) relative to several larger models corroborates that its performance is stable across test cases, not merely averaged over a bimodal distribution of successes and failures.

Balanced local and contextual modelling. Unlike pure transformer architectures that tokenise the image at coarse resolution and may lose sub-pixel boundary detail, CMUNeXt integrates local convolutional texture features with broader spatial context through its inverted residual design. This hybrid inductive bias is well-matched to the demands of US segmentation, where organs are embedded in a heterogeneous acoustic field and accurate boundary delineation

requires simultaneous awareness of local echo patterns and global organ geometry.

5.2. Why Small Anatomical Structures Remain Difficult

Figure 6 reveals a consistent and architecturally universal failure mode: for regions containing fewer than 250 pixels, all nine evaluated models achieve Dice scores approaching zero, regardless of their overall benchmark ranking. This observation is not incidental but reflects several compounding properties of US data and segmentation task formulation.

Low contrast and speckle noise. Ultrasound images are formed by coherent interference of acoustic backscatter, producing characteristic speckle patterns whose spatial statistics can overwhelm the intensity differences that discriminate small anatomical structures from surrounding tissue. Unlike CT or MRI, where organ boundaries are defined by reliable attenuation or relaxation contrasts, US boundaries in small structures are often only marginally distinguishable from noise at standard clinical resolution.

Insufficient spatial context. Convolutional and attention-based feature extractors rely on neighbourhood context to disambiguate foreground from background. For structures occupying fewer than 250 pixels ($\leq 0.1\%$ of a 512×512

image), the local receptive field contains mostly background, and the feature representation becomes dominated by the surrounding tissue statistics rather than the target structure itself. This manifests as the near-zero Dice plateau observed in the 0–250 pixel size bins.

Class imbalance. Small structures contribute negligibly to the aggregate pixel count, creating severe foreground–background imbalance. Standard composite losses (cross-entropy + Dice loss) partially mitigate this through Dice-based weighting, but cannot fully compensate when the foreground region is so small that even a near-perfect prediction yields a Dice denominator dominated by noise fluctuations. Architectures with high loss standard deviations—notably EMCAD (0.1420 ± 0.1032) and Residual U-Net (0.1940 ± 0.2413)—are particularly susceptible to this instability.

Boundary ambiguity. At small scales, the annotation boundary itself may span only a few pixels, making the ground-truth mask sensitive to inter-annotator variability. Predictions that are spatially close but not exactly coincident with such boundaries receive disproportionate penalty under pixel-level Dice computation, further suppressing reported scores. Addressing this challenge will likely require scale-adaptive loss functions, dedicated small-object augmentation strategies, or auxiliary detection heads that explicitly localise small structures prior to segmentation.

5.3. CNN versus Transformer Behaviour in Ultrasound Segmentation

The performance ordering in Table 3 provides an informative perspective on the relative merits of convolutional and transformer-based inductive biases for US image segmentation.

CNNs capture local texture effectively. Standard convolutional architectures (UNet, Att-UNet, ResUNet++) achieve competitive DSC scores (0.7518–0.7618) precisely because US segmentation is substantially a texture-discrimination problem at the local scale. Echo patterns, shadow artefacts, and reverberation noise are spatially correlated over distances of tens of pixels, a range well-matched to the receptive fields of shallow-to-mid-depth convolutional layers. The t-SNE embeddings of UNet, however, exhibit notable inter-class scatter, suggesting that while local features are captured, they lack the semantic consistency needed for robust boundary delineation across structurally diverse instances.

Transformers capture global context but at a cost. MISSFormer, a transformer-based architecture with 30.06 M parameters, achieves a mean DSC of only 0.7029 ± 0.2261 —significantly below CMUNeXt ($p < 0.001$)—despite its capacity for long-range dependency modelling. This result suggests that global self-attention, while powerful in natural image recognition and in CT/MRI segmentation where training sets are large and domain statistics are stable, may not generalise effectively to small-to-moderate US datasets. Transformer models require substantially more data to learn meaningful attention patterns; when training data are insufficient to populate the attention weight space reliably,

global context modelling degrades to approximate memorisation of training-set distributions, manifesting as the high standard deviation (0.2261) and widened confidence interval ([0.6684, 0.7374]) observed for MISSFormer.

Ultrasound segmentation benefits from both local and global cues. The highest-performing models in this benchmark—CMUNeXt, UCTransNet, and ResUNet++—are all hybrid architectures that combine local convolutional feature extraction with some form of broader spatial aggregation, whether through large-kernel convolutions, cross-attention channel fusion, or residual multi-scale pathways. This pattern is consistent with the physical nature of US image formation: accurate organ segmentation requires awareness of both the local echo texture (to detect boundaries) and the global organ shape and position (to resolve ambiguous local regions). Neither purely local nor purely global processing is sufficient in isolation.

Heavy transformer models risk over-fitting on small datasets. UCTransNet, with 59.75 M parameters, achieves the second-highest DSC (0.7648), but its advantage over ResUNet++ (0.7618, 4.06 M) is marginal ($\Delta\text{DSC} = 0.0030$, $p = 0.014$). The disproportionate parameter cost for this marginal gain—a ratio of approximately $14.7\times$ more parameters for 0.4% more Dice—raises concerns about over-parameterisation and the associated risks of overfitting, sensitivity to hyperparameter choices, and prohibitive inference cost in resource-constrained settings.

5.4. Clinical and Computational Implications

The results carry several practical implications for the translation of automated US segmentation into clinical workflows.

Lightweight models suit point-of-care and portable ultrasound. The growing adoption of handheld and point-of-care US devices in emergency medicine, primary care, and low-resource settings imposes strict constraints on inference latency and memory footprint. CMUNeXt (3.15 M parameters) and ResUNet++ (4.06 M parameters) are strong candidates for deployment in such environments, offering near-equivalent accuracy to much larger models at a fraction of the computational cost. In contrast, UCTransNet (59.75 M) and Att-UNet (34.88 M) incur substantial memory and latency overheads that may be prohibitive on embedded inference hardware.

Segmentation as a foundation for clinical measurement. Automated segmentation of kidney and spleen boundaries enables downstream clinical tasks including volumetric measurement (relevant for renal function monitoring and splenomegaly assessment), anatomical localisation for biopsy guidance, and longitudinal tracking of organ size changes under treatment. The Dice and IoU scores reported here, while strong at the benchmark level, must be contextualised against clinically meaningful error thresholds: a mean DSC of 0.7792 with a standard deviation of 0.1897 implies that a non-trivial proportion of test predictions fall below 0.60, which may be insufficient for precise volumetric measurements without clinician review.

External validation is required before clinical adoption. All results presented in this study are derived from a single harmonised benchmark dataset. While the harmonisation procedure accounts for multi-source heterogeneity, the dataset does not capture the full spectrum of acquisition variability encountered in clinical practice, including differences in US machine vendors, probe frequencies, operator experience, and patient pathology. External validation on prospective, multi-site data is a prerequisite for any claim of clinical readiness.

Confidence intervals should inform clinical decision thresholds. The wide confidence intervals observed across all models—particularly for architectures with high standard deviation (EMCAD: [0.5794, 0.6546]; MISSFormer: [0.6684, 0.7374])—highlight the importance of reporting uncertainty alongside point estimates. In clinical contexts, a model that performs well on average but exhibits high per-case variance poses a patient safety risk; confidence intervals and per-case reliability metrics should be integrated into model selection criteria alongside mean performance.

5.5. Limitations

This study has several limitations that should be acknowledged to properly scope its conclusions.

Single benchmark evaluation. The reported results reflect performance on a single harmonised multi-anatomy US dataset comprising kidney and spleen segmentation tasks. Generalisation to other anatomical targets (e.g., liver, thyroid, or cardiac structures) or to pathological cases (e.g., hydronephrosis, splenomegaly, or renal cysts) cannot be assumed without dedicated experimental validation.

Moderate test set size. With $N = 165$ test samples, the statistical power to detect small but clinically meaningful performance differences is limited. The non-significant p -values for UNet++ ($p = 0.267$) and Att-UNet ($p = 0.165$) relative to CMUNeXt should not be interpreted as evidence of equivalence; rather, they reflect insufficient power to resolve differences of the observed magnitude ($\Delta\text{DSC} < 0.030$) at this sample size. Larger prospective test sets would sharpen these comparisons.

Summary-statistic p -values. Statistical comparisons were performed using a paired t -test on per-sample Dice scores. While appropriate for continuous, approximately normally distributed outcomes, the high standard deviations observed across all models (range 0.1897–0.2463) suggest that the Dice distribution may be left-skewed, in which case non-parametric alternatives such as the Wilcoxon signed-rank test may provide more conservative and appropriate inference.

Absence of post-processing. No morphological post-processing (e.g., connected-component filtering, conditional random field refinement) was applied to model predictions. It is possible that such steps would disproportionately benefit models with fragmented predictions (e.g., Residual U-Net, EMCAD), narrowing the performance gap with CMUNeXt. Future work should evaluate models under standardised post-processing pipelines to ensure fair comparison.

Single-fold evaluation and reproducibility. Results were obtained from a single train–test split. Cross-validation or repeated holdout evaluation would provide more reliable estimates of expected performance and variance, and reduce the risk that the reported ordering reflects a split-specific artefact.

Loss standard deviation anomaly. UNet++ records a loss standard deviation of 0.2743 against a mean of 0.1590 (coefficient of variation = 1.72), which is substantially higher than all other models and warrants further investigation. This may reflect bimodal behaviour—accurate predictions on most samples and catastrophic failures on a small subset—which would be clinically significant but is not captured by mean performance metrics alone.

6. Conclusion and Future Work

6.1. Conclusion

This study presented a standardised comparative benchmark of nine representative deep learning architectures—spanning classical fully convolutional networks, attention-enhanced encoders, and hybrid convolution–transformer designs—for multi-anatomical structure segmentation in B-mode ultrasound imaging. A harmonised benchmark dataset was constructed by integrating the Open Kidney Ultrasound (OKUS) and Spleenex datasets, yielding 984 annotated images covering five foreground tissue classes: capsule, cortex, medulla, central echo complex, and spleen. All models were trained and evaluated under rigorously identical conditions—shared preprocessing, augmentation, optimiser, loss function (0.7 Dice + 0.3 Focal), and an 80:10:10 data split—to ensure that observed performance differences are attributable to architectural choices rather than experimental artefacts. Evaluation was conducted on a held-out test set of $N = 165$ samples, with statistical significance assessed by paired t -test against the top-performing model.

Four principal conclusions emerge from the experimental evidence.

First: efficient multi-scale hybrid architectures offer the best accuracy–complexity trade-off. CMUNeXt achieved the highest mean Dice Similarity Coefficient (DSC) of 0.7792 ± 0.1897 (95% CI: [0.7503, 0.8081]), the highest mean Intersection over Union (IoU) of 0.6323 ± 0.1539 , and the lowest mean loss of 0.0814 ± 0.0782 , while requiring only 3.15 M trainable parameters. Its superiority over models with substantially larger parameter budgets—UCTransNet (59.75 M, DSC = 0.7648, $p = 0.004$), Att-UNet (34.88 M, DSC = 0.7519, $p = 0.165$), and MISSFormer (30.06 M, DSC = 0.7029, $p < 0.001$)—demonstrates that parameter count is not a reliable predictor of segmentation quality. Instead, the combination of large-kernel depthwise convolutions, inverted residual multi-scale feature fusion, and a compact bottleneck design enables CMUNeXt to capture both fine-grained local echo texture and broader organ-level spatial context simultaneously, which is precisely the dual requirement imposed by heterogeneous multi-anatomy US

imaging. Per-tissue analysis further confirmed the consistency of this advantage: CMUNeXt recorded the highest Dice for Spleen (0.9834), Central Echo Complex (0.8359), Capsule (0.7207), Medulla (0.6832), and Cortex (0.6724), making it the only model to lead across all five tissue classes concurrently. ResUNet++ (4.06 M, DSC = 0.7618) emerged as the most parameter-efficient alternative, achieving near-equivalent accuracy to CMUNeXt at a comparable parameter cost, and represents a strong second choice for resource-constrained deployment scenarios.

Second: robust segmentation of small anatomical structures remains an open and architecturally universal challenge. Across all nine evaluated models, Dice scores approach zero for ground-truth regions smaller than 250 pixels, irrespective of architecture type, parameter count, or attention mechanism. This failure is driven by the compounding effects of low acoustic contrast, coherent speckle noise, severely imbalanced foreground-to-background pixel ratios, and sub-pixel annotation uncertainty at small scales—factors that are intrinsic to the ultrasound imaging modality and cannot be resolved by architectural scaling alone. The Dice–Focal composite loss employed in this study provides partial mitigation through Dice-based weighting and hard-example emphasis, but is insufficient when the foreground region occupies fewer than 0.1% of the image plane. This finding identifies small-structure segmentation as the primary unsolved problem in multi-anatomical US segmentation, with direct clinical implications for the delineation of cortical and capsular kidney structures whose accurate measurement is diagnostically important in nephrology.

Third: transformer-based architectures do not transfer effectively to small-to-moderate ultrasound datasets without domain-specific adaptation. MISSFormer, despite its capacity for long-range dependency modelling and its 30.06 M parameter investment, achieved a mean DSC of only 0.7029 ± 0.2261 —significantly below CMUNeXt ($p < 0.001$) and below several purely convolutional models with far fewer parameters. Similarly, UCTransNet, the second-best performing model overall, required 59.75 M parameters for a statistically significant but practically marginal $\Delta\text{DSC} = 0.0144$ advantage over CMUNeXt. These results suggest that global self-attention mechanisms, developed and validated on large-scale CT and MRI corpora, are prone to approximate memorisation of training-set statistics when applied to small-to-moderate US datasets, manifesting as high standard deviation and wide confidence intervals rather than consistent per-case performance. The t-SNE feature embeddings corroborate this interpretation: transformer models show moderate inter-class separation but residual cluster dispersion in structurally ambiguous boundary regions, confirming that long-range attention alone does not resolve the local texture discrimination problem central to US segmentation.

Fourth: per-sample consistency is as clinically important as mean accuracy, and must be reported alongside point estimates. The high standard deviations observed across all models (range 0.1897–0.2463 for DSC) imply

that mean performance metrics conceal substantial per-case variability. The UNet++ loss standard deviation anomaly— $\text{CV} = 1.72$, $\text{SD} = 0.2743$ against mean = 0.1590—is a particularly salient example: a model that achieves the third-highest mean DSC (0.7540) while exhibiting this degree of loss instability poses a safety risk in clinical workflows where prediction reliability across diverse anatomical presentations is required. The non-significant p -values for UNet++ ($p = 0.267$) and Att-UNet ($p = 0.165$) further illustrate that statistical indistinguishability from the best model does not imply clinical equivalence when SD differs. Confidence intervals and per-case reliability distributions should therefore be regarded as primary—not supplementary—reporting requirements in future comparative studies.

Taken together, these findings establish that efficient multi-scale hybrid architectures, particularly CMUNeXt and ResUNet++, currently offer the most favourable balance of segmentation accuracy, consistency, and computational feasibility for multi-organ ultrasound segmentation. They further identify small-structure delineation and domain-specific transformer adaptation as the two principal open problems whose resolution would most meaningfully advance the clinical utility of automated US segmentation.

6.2. Future Work

Building on the findings and acknowledged limitations of this study, we identify five concrete directions for future research.

Scale-adaptive architectures and loss functions for small structures. The universal failure below 250 pixels motivates the development of dedicated small-structure segmentation modules. Promising directions include: (i) scale-adaptive loss formulations that dynamically reweight foreground contributions based on region size at the sample level, rather than applying a fixed $\alpha = 0.7$ Dice weighting globally; (ii) auxiliary detection heads that first localise candidate small-structure bounding boxes before applying a fine-grained segmentation decoder, decoupling the detection and delineation sub-problems; and (iii) multi-resolution input pipelines that provide the network with zoomed local crops of candidate small-structure regions alongside the full-resolution image, supplying the contextual signal that standard single-scale encoders cannot extract from structures occupying fewer than 0.1% of the image plane.

Domain-adapted transformer architectures for ultrasound. The underperformance of MISSFormer and the parameter inefficiency of UCTransNet suggest that transformer architectures require structural adaptation to the US domain rather than direct transfer from CT/MRI settings. Specific adaptation strategies include: (i) locality-biased attention mechanisms, such as shifted-window or deformable attention, that prioritise spatially proximate feature interactions consistent with the short-range correlation structure of US speckle fields; (ii) noise-aware pre-training objectives, such as masked speckle prediction or contrastive anatomy alignment across probe orientations, to provide transformers with domain-relevant self-supervised initialisation; and

(iii) knowledge distillation from large-capacity transformer teachers to compact student architectures of the CMUNeXt class, preserving global context modelling at a fraction of the deployment cost.

Large-scale, multi-site, and pathology-inclusive dataset construction. All results in this study are derived from a single harmonised benchmark of 984 images from two sources, limiting the generalisability of conclusions to specific acquisition devices, patient demographics, and normal anatomy. Future work should prioritise the construction of a large-scale, prospective multi-anatomy US corpus that encompasses: multiple ultrasound vendors and probe frequencies; diverse patient demographics including paediatric, geriatric, and pathological populations (e.g., hydronephrosis, splenomegaly, renal cysts, and parenchymal disease); and expert annotations with inter-rater reliability metrics to quantify labelling uncertainty. Such a dataset would enable cross-site generalisation studies, domain adaptation benchmarks, and statistically powered subgroup analyses that the current benchmark cannot support.

Standardised post-processing and uncertainty quantification pipelines. The absence of morphological post-processing in this benchmark, while necessary for controlled architectural comparison, leaves open the question of how much of the performance gap between models is reducible by connected-component filtering, conditional random field smoothing, or test-time augmentation. Future benchmarks should evaluate models under both raw-prediction and standardised post-processing conditions. In parallel, the integration of uncertainty quantification frameworks—such as Monte Carlo dropout, deep ensembles, or conformal prediction—would enable models to output calibrated confidence maps alongside segmentation masks, supporting the clinical requirement for explicit per-case reliability estimates that mean DSC alone cannot provide.

Prospective clinical validation and real-time deployment evaluation. Demonstrating benchmark performance is a necessary but insufficient condition for clinical translation. Future work should evaluate CMUNeXt and ResUNet++ in prospective clinical settings, assessing: inference latency and memory footprint on embedded and point-of-care hardware platforms (e.g., ARM-based devices and portable US systems); robustness under acquisition variability introduced by different operators, gel coupling quality, and patient positioning; and agreement with clinical reference standards for downstream tasks such as renal volumetry, splenomegaly grading, and biopsy guidance. Cross-institutional validation studies, ideally using the proposed benchmark protocol as a reproducibility template, would provide the evidence base required to support regulatory review and clinical adoption of automated multi-anatomy US segmentation.

Declarations

All declarations are provided below:

CRedit authorship contribution statement

Md. Rashed: Conceptualization, Data curation, Methodology, Visualization, Project administration, and Writing – original draft. **Riad Hassan:** Writing – review and editing, Formal analysis, Resources, Supervision, and Validation. **Md. Imran Hossain:** Supervision, Formal analysis, Investigation, and Validation. **Md Rakibul Islam:** Writing – review and editing, Formal analysis, Resources, Supervision, and Validation. **Feroza Naznin:** Formal analysis, Validation, Investigation. **Md Zahidul Islam:** Supervision, Writing – review and editing, Formal analysis, Investigation.

Funding Statement

The authors did not receive any financial support for the research.

Data Availability Statement

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 standardized six-class labels and preprocessing scripts are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest related to this work.

AI Assistance Statement

AI tools (Claude and Grammarly) were used only to improve grammar, spelling, and language clarity. No part of the research design, analysis, results, or scientific content was generated by these tools. All technical and intellectual contributions in this manuscript were made by the authors.

References

- Amiri, M., Brooks, R., Behboodi, B., Rivaz, H., 2020. Two-stage ultrasound image segmentation using u-net and test time augmentation. *International Journal of Computer Assisted Radiology and Surgery* 15, 981–988.
- Arora, R.K., Kumar, A., Soni, A., 2024. Deep learning approaches for enhanced kidney segmentation: Evaluating u-net and attention u-net with cross-entropy and focal loss functions, in: 2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC), IEEE. pp. 854–860.
- Chen, G., Li, L., Dai, Y., Zhang, J., Yap, M.H., 2022a. Aau-net: an adaptive attention u-net for breast lesions segmentation in ultrasound images. *IEEE Transactions on Medical Imaging* 42, 1289–1300. doi:10.1109/TMI.2022.3145637.
- Chen, G., Yin, J., Dai, Y., Zhang, J., Yin, X., Cui, L., 2022b. A novel convolutional neural network for kidney ultrasound images segmentation. *Computer Methods and Programs in Biomedicine* 218, 106712.
- Chen, H., Cai, Y., Wang, C., Chen, L., Zhang, B., Han, H., Guo, Y., Ding, H., Zhang, Q., 2024a. Multi-organ foundation model for universal ultrasound image segmentation with task prompt and anatomical prior. *IEEE Transactions on Medical Imaging* 44, 1005–1018.
- Chen, Z., Zhu, H., Liu, Y., Gao, X., 2024b. Msca-unet: multi-scale channel attention-based unet for segmentation of medical ultrasound images. *Cluster Computing* 27, 6787–6804.
- Duque, V., Marquardt, A., Velikova, Y., Lacourpaille, L., Nordez, A., Crouzier, M., Lee, H., Mateus, D., Navab, N., 2024. Ultrasound segmentation analysis via distinct and completed anatomical borders. *International Journal of Computer Assisted Radiology and Surgery* 19, 1419–1427. doi:10.1007/s11548-024-03170-7.

- Huang, X., Deng, Z., Li, D., Yuan, X., 2021. Missformer: An effective medical image segmentation transformer. *arXiv preprint arXiv:2109.07162*.
- Jain, P.K., Dubey, A., Saba, L., Khanna, N.N., Laird, J.R., Nicolaides, A., Sharma, N., 2022. Attention-based unet deep learning model for plaque segmentation in carotid ultrasound for stroke risk stratification: an artificial intelligence paradigm. *Journal of Cardiovascular Development and Disease* 9, 326.
- Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D., 2019. Resunet++: An advanced architecture for medical image segmentation, in: 2019 IEEE international symposium on multimedia (ISM), IEEE, pp. 225–2255.
- Jiang, J., Guo, Y., Bi, Z., Huang, Z., Yu, G., Wang, J., 2023. Segmentation of prostate ultrasound images: The state of the art and the future directions of segmentation algorithms. *Artificial Intelligence Review* 56, 615–651. doi:10.1007/s10462-022-10179-4.
- Karimi, A., Seraj, J., Mirzadeh Sarcheshmeh, F., Fazli, K., Seraj, A., Eslami, P., Khanmohamadi, M., Sajjadian Moosavi, H., Ghattan Kashani, H., Sajjadian Moosavi, A., Shariat Panahi, M., 2025. Improving spleen segmentation in ultrasound images using a hybrid deep learning framework. *Scientific Reports* 15, 1670.
- Klibanov, A.L., Hossack, J.A., 2015. Ultrasound in radiology: from anatomic, functional, molecular imaging to drug delivery and image-guided therapy. *Investigative Radiology* 50, 657–670. doi:10.1097/RLI.000000000000188.
- Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Laak, J.A.V.D., Ginneken, B.V., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis* 42, 60–88. doi:10.1016/j.media.2017.07.005.
- Liu, X., Qu, L., Xie, Z., Shi, Y., Song, Z., 2024. Deep mutual learning among partially labeled datasets for multi-organ segmentation. *arXiv:2407.12611*.
- MedScience, O., 2025. The role of ultrasound in medical imaging. URL: https://openmedscience.com/the-role-of-ultrasound-in-medical-imaging/?utm_source=chatgpt.com. accessed: 2025-10-27.
- Rahman, M.M., Munir, M., Marculescu, R., 2024. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11769–11779.
- Singla, R., Ringstrom, C., Hu, G., Lessoway, V., Reid, J., Ngan, C., Rohling, R., 2023. The open kidney ultrasound data set, in: International Workshop on Advances in Simplifying Medical Ultrasound, Springer, pp. 155–164.
- Tang, F., Ding, J., Quan, Q., Wang, L., Ning, C., Zhou, S.K., 2024. Cmunext: An efficient medical image segmentation network based on large kernel and skip fusion, in: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), IEEE, pp. 1–5.
- Wang, C., Zhang, J., Liu, S., 2022. Medical ultrasound image segmentation with deep learning models. *IEEE Access* 11, 10158–10168.
- Wang, H., Wu, G., Liu, Y., 2025. Efficient generative-adversarial u-net for multi-organ medical image segmentation. *Journal of Imaging* 11, 19. doi:10.3390/jimaging11010019.
- Wu, J., Liu, F., Sun, W., Liu, Z., Hou, H., Jiang, R., Zhang, X., 2024. Boundary-aware convolutional attention network for liver segmentation in ultrasound images. *Scientific Reports* 14, 21529.
- Xiao, X., Zhang, J., Shao, Y., Liu, J., Shi, K., He, C., Kong, D., 2025. DeepLearning-Based Medical Ultrasound Image and Video Segmentation Methods: Overview, Frontiers, and Challenges. *Sensors* 25, 2361. doi:10.3390/s25082361.
- Yang, Y., Chen, F., Liang, H., Bai, Y., Wang, Z., Zhao, L., Ma, S., et al., 2023. CNN-based automatic segmentations and radiomics feature reliability on contrast-enhanced ultrasound images for renal tumors. *Frontiers in Oncology* 13, 1166988. doi:10.3389/fonc.2023.1166988.
- Zhao, T., Dai, H., 2022. Breast tumor ultrasound image segmentation method based on improved residual u-net network. *Computational Intelligence and Neuroscience* 2022, 3905998.
- Zhou, T., Hou, S., Lu, H., Zhao, Y., Dang, P., Dong, Y., 2022. Exploring and analyzing the improvement mechanism of u-net and its application in medical image segmentation. *Sheng wu yi xue Gong Cheng xue za zhi = Journal of Biomedical Engineering = Shengwu Yixue Gongchengxue Zazhi* 39, 806–825.