MASP-UNet: Acquisition-Aware Lightweight Segmentation of Renal Structures in Ultrasound

Anurag Aryal, Mohammed Yusuf Ansari, Eduardo Feo-Flushing
Carnegie Mellon University in Qatar, Doha, Qatar
{aaryal, ma1, efeoflus}@andrew.cmu.edu

***Abstract*—Renal substructure segmentation in ultrasound requires models that are both accurate and computationally efficient for point-of-care deployment. While high-capacity networks achieve strong segmentation performance, their computational demands limit practical use, whereas lightweight alternatives lack sufficient representational capacity. Existing methods also overlook acquisition metadata, despite ultrasound image characteristics varying systematically with imaging hardware. We present MASP-UNet, a lightweight acquisition-aware encoder–decoder that conditions its bottleneck representation on ultrasound acquisition metadata while integrating Spatial Pyramid Pooling (SPP), Multi-Context Aggregation (MCA), and attention-gated skip connections to capture multi-scale anatomical context within a compact architecture. On the Open Kidney Ultrasound Dataset, MASP-UNet achieves a mean Dice coefficient of 0.65 across cortex, medulla, and central echogenic complex segmentation, matching the best-performing baseline within 0.003 Dice while requiring only 5.7M parameters, 4.7 GFLOPs, and 53 ms CPU inference latency. These results indicate that acquisition-aware, multi-scale bottleneck conditioning improves the efficiency–accuracy trade-off for lightweight renal ultrasound segmentation, making the approach well-suited for resource-constrained clinical deployment.**

I. Introduction

Ultrasound is the primary imaging modality for renal assessment owing to its non-ionizing nature, low cost, and widespread clinical availability [1], [2]. Accurate segmentation of the renal cortex, medulla, and central echogenic complex (CEC) enables quantitative biomarkers such as cortical thickness and kidney volume for clinical diagnosis and longitudinal monitoring [3], [4]. However, manual delineation is labor-intensive and subject to considerable inter-operator variability, motivating automated segmentation methods [5].

Recent deep learning methods have improved renal ultrasound segmentation but remain challenging to deploy at the point of care. Large encoder–decoder architectures such as PSPNet achieve high accuracy at substantial computational cost [1], whereas lightweight models like UNeXt and EMCAD reduce computation at the expense of anatomical fidelity [6], [7]. Moreover, existing lightweight segmentation models do not explicitly exploit acquisition metadata, despite image appearance varying systematically with acquisition hardware, and this information being routinely available in ultrasound systems.

In this work, we investigate acquisition metadata as an inexpensive conditioning signal for a compact segmentation network that preserves point-of-care efficiency. To this end, we propose MASP-UNet, an acquisition-aware lightweight architecture that conditions its bottleneck representation on ultrasound acquisition parameters while jointly integrating multi-scale context through Spatial Pyramid Pooling (SPP) and Multi-Context Aggregation (MCA). This bottleneck-centric design enhances feature representation without increasing encoder or decoder complexity. Experimental results show that MASP-UNet achieves segmentation accuracy comparable to substantially larger architectures while remaining suitable for real-time CPU inference.

Our contributions are summarized as follows:

- To our knowledge, MASP-UNet is the first lightweight renal ultrasound segmentation neural network to explicitly leverage acquisition metadata during segmentation.
- We propose an acquisition-aware bottleneck that jointly integrates ultrasound acquisition metadata, Spatial Pyramid Pooling (SPP), and Multi-Context Aggregation (MCA). Through module and locality-based ablation studies, we demonstrate that concentrating these components at the bottleneck is an effective engineering optimization that improves representation while preserving encoder–decoder efficiency.
- MASP-UNet achieves a mean Dice coefficient of 0.65, within 0.003 Dice of PSPNet, which achieved the highest test-set mean Dice, using only 5.7M parameters, 4.7 GFLOPs, and 53 ms CPU latency. We validate our results through five-fold cross-validation, ablation studies, CPU benchmarking on commodity hardware, and Grad-CAM analysis.

II. Methodology: Network Architecture

MASP-UNet follows a U-Net backbone [8] in which the encoder progressively downsamples ultrasound images through convolutional blocks and max pooling, preserving spatial resolution at each scale. At the bottleneck, the feature map passes through Spatial Pyramid Pooling (SPP) [9], which captures multi-scale structural features across varying receptive fields. Acquisition metadata encoding transducer type, operating frequency, and spatial resolution is projected as a learned embedding and concatenated into the bottleneck feature space. Multi-Context Aggregation (MCA) then consolidates these multi-scale and hardware-conditioned features into a unified representation. The decoder reconstructs the segmentation map through successive upsampling blocks, with

This research was funded by Carnegie Mellon University in Qatar under the Qatar Student Initiated Undergraduate Research Program.

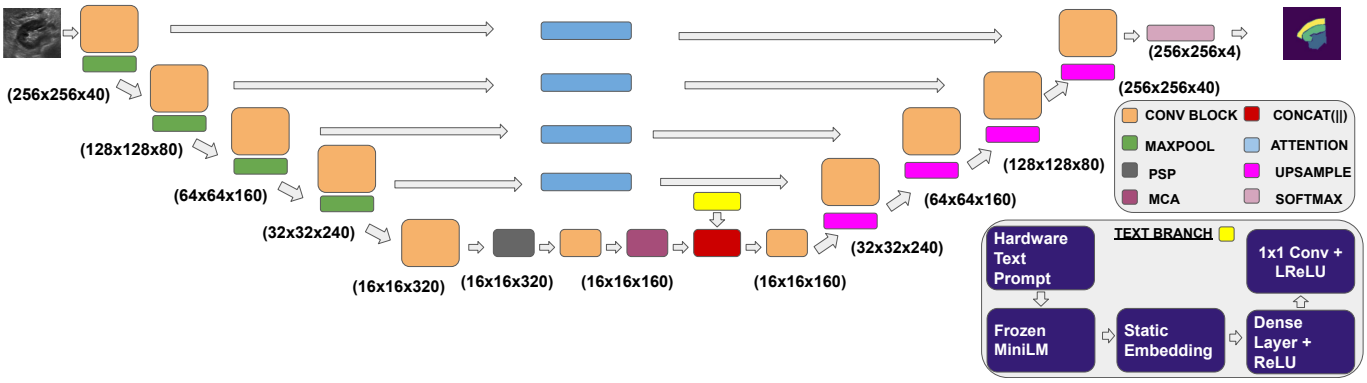


Fig. 1. MASP-UNet architecture & overview of Acquisition Metadata Embedding module.

attention gates applied on each skip connection to selectively retain anatomically relevant features from the encoder.

1) Spatial Pyramid Pooling: Renal structures exhibit significant scale variation, from the large cortical mantle to fine medullary pyramids. An SPP module is placed exclusively at the bottleneck to capture multi-scale spatial context without expanding decoder complexity. Given the bottleneck feature map $F \in \mathbb{R}^{H\times W\times C}$, SPP applies k pooling operations at increasing window sizes and concatenates their outputs:

$$F_{\text{SPP}} = \text{Concat}\left[P_1(F), P_2(F), \ldots, P_k(F)\right] \tag{1}$$

where P_i denotes average pooling at window size i. The concatenated output is projected back to $\mathbb{R}^{H\times W\times C}$ via a 1×1 convolution followed by batch normalization, yielding F_{spp}.

2) Acquisition Metadata Embedding: Ultrasound image statistics vary with acquisition hardware [10]. We therefore build a hardware text prompt T from six DICOM acquisition attributes provided with the dataset: manufacturer, model name, transducer identifier, transducer frequency, and pixel spacing (ΔX, ΔY). Only hardware fields are used, and the same fixed template is applied to every image:

```
Ultrasound acquired with <Manufacturer>
<ManufacturerModelName>. Transducer:
<TransducerData>. Frequency:
<TransducerFrequency>. Physical
resolution: deltaX=<Physical Delta
X>, deltaY=<Physical Delta Y>.
```

Numeric fields are inserted as stored (e.g., `deltaX=0.020001`). A missing frequency is written as `unspecified`, and the 18 of 472 images without a metadata record receive the fallback prompt `Manufacturer unknown. Transducer unknown. Frequency unknown.` The prompt is encoded by the frozen sentence transformer ϕ (`all-MiniLM-L6-v2`) with mean pooling over the final hidden states, giving $e = \phi(T) \in \mathbb{R}^{384}$. Since acquisition settings are fixed within a scanning session, e is precomputed and cached, adding no inference overhead. It is projected into the bottleneck through a two-layer MLP and a 1×1 convolution with learned parameters W_1, W_2, b_1, b_2:

$$h = \text{ReLU}(W_1 e + b_1), \quad h \in \mathbb{R}^{128} \tag{2}$$

$$Z = \text{reshape}(\text{ReLU}(W_2 h + b_2)) \in \mathbb{R}^{H\times W\times 1} \tag{3}$$

$$F_{\text{meta}} = \text{LeakyReLU}(\text{BN}(\text{Conv}_{1\times1}(Z))) \in \mathbb{R}^{H\times W\times C} \tag{4}$$

F_{meta} is concatenated channel-wise with the SPP output and compressed to the native bottleneck depth:

$$F_{\text{fused}} = \text{LeakyReLU}(\text{BN}(\text{Conv}_{1\times1}([F_{\text{spp}} \parallel F_{\text{meta}}]))) \tag{5}$$

3) Multi-Context Aggregation: Acoustic ambiguity in ultrasound produces indistinct boundaries between the renal cortex and medulla. The MCA block is applied immediately after metadata fusion to refine F_{fused} using parallel dilated convolutions at rates $r \in R$, capturing local parenchymal texture and global context within a fixed parameter budget:

$$F_{\text{MCA}} = \sum_{r\in R} \text{Conv}_r(F_{\text{fused}}) \tag{6}$$

where each Conv_r denotes a dilated convolution with rate r.

4) Attention Gates: Renal structures occupy a small spatial fraction of the ultrasound field of view, with the surrounding background dominated by acoustic speckle. Attention gates are applied on each skip connection to weight encoder features x using the decoder gating signal g:

$$\alpha = \sigma(W_\psi(W_x x + W_g g + b)), \quad F_{\text{att}} = \alpha \cdot x \tag{7}$$

where W_x and W_g are learned linear transformations, W_ψ projects the combined signal to a scalar attention coefficient, σ denotes the sigmoid activation, and $\alpha \in [0, 1]$ is the resulting spatial attention map.

III. Experimental Setup

1) Dataset and Preprocessing: Experiments were conducted on the Open Kidney Ultrasound Dataset [11], comprising 534 B-mode images each annotated independently by two sonographers across four classes: background (94.0%), cortex, medulla, and CEC. One of the two expert annotations is selected randomly per image during training to model inter-observer variability. Images are resized to 256×256, normalized to $[0, 1]$, and one-hot encoded. Images containing

only background are excluded to mitigate class imbalance [12]. A patient-level split reserves 10% of the data as a held-out test set to prevent data leakage; the remaining 90% is used for five-fold cross-validation.

2) Training: All models are trained under identical conditions using a composite loss combining class-weighted Dice, IoU, and Focal Tversky terms [13]–[15]. Class weights are set to 1.0, 8.0, 12.0, and 15.0 for background, cortex, medulla, and CEC respectively. Training uses Adam with AMSGrad and a staged learning rate schedule [16]. Augmentation includes random rotation up to 20°, horizontal flipping, and brightness and contrast jitter. All seeds and hyperparameters are fixed across runs to ensure reproducibility.

3) Evaluation Metrics: We report per-class and mean Dice coefficient (DC), parameter count, model size, inference time, GFLOPs, memory usage, and power consumption, where $\mathrm{DC} = 2|A \cap B|/(|A|+|B|)$ for predicted mask A and ground-truth mask B [17], [18].

4) Baseline Models: MASP-UNet is benchmarked against seven architectures trained under identical conditions, spanning self-configuring, boundary-aware, and lightweight transformer and MLP-based designs: nnU-Net [19], SegNet [20], EfficientNet-B0 [21], PSPNet (ResNet-50) [9], BDRPCN [22], UNeXt [7], and EMCAD [6].

IV. Results and Discussion

1) Component-Wise Architecture Ablation: Starting from a baseline U-Net, convolutional channels were progressively reduced as SPP, attention gates, MCA, and the metadata embedding were added, keeping every variant within the ∼5M parameter budget. Fig. 2 reports per-class Dice coefficients for Sonographer 1 across five-fold cross-validation.

Each module contributes a measurable, largely non-overlapping gain. SPP captures multi-scale spatial context absent in the baseline, improving cortex and medulla Dice. MCA delivers the largest single gain in the medulla (0.40 to 0.48) by aggregating dilated multi-receptive-field context for fine, low-contrast structures, while attention gates produce the sharpest CEC improvement (0.30 to 0.35) by suppressing background speckle in a class that occupies a minor image fraction. The metadata embedding adds a further consistent gain across all foreground classes, most notably in the CEC (0.37 to 0.40), although this in-distribution gain alone does not establish robustness to unseen hardware.

Overall, the fully integrated model improves over the baseline by 0.11 (cortex), 0.09 (medulla), and 0.13 (CEC) Dice, with the largest relative gains concentrated in the smallest and most acoustically ambiguous structures – a pattern consistent across Sonographer 1, Sonographer 2, and the randomized split.

2) Locality-Based Positioning Ablation: To evaluate optimal module placement, components were benchmarked at alternative network locations while dynamically scaling the backbone to maintain the ∼5M parameter budget. Table I reports five-fold cross-validation results for Sonographer 1, consistent with trends observed across all annotation splits.

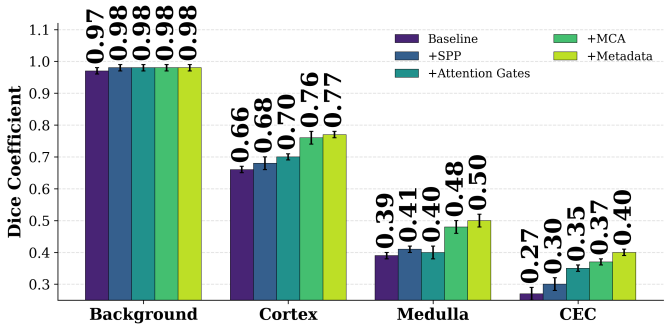


Fig. 2. Component-wise ablation results (Sonographer 1, 5-fold CV).

TABLE I
Effect of Module Placement Zone on DCE and Latency (Sonographer 1, 5-fold CV)

Placement Zone	Spatial Res.	Inf. Time (ms)	Dice Coefficient			
			Background	Cortex	Medulla	CEC
Encoder	64 × 64	69 ± 3	0.95 ± .01	0.65 ± .03	0.40 ± .03	0.30 ± .02
Decoder	64 × 64	64 ± 4	0.97 ± .01	0.69 ± .03	0.42 ± .01	0.33 ± .02
Bottleneck	16 × 16	**53 ± 3**	**0.98 ± .01**	**0.77 ± .01**	**0.50 ± .02**	**0.40 ± .01**

Encoder placement inserts the modules before the final encoder block, and decoder placement inserts them after the first upsampling stage; both operate on 64×64 feature maps, four times larger than the 16×16 bottleneck. Processing multi-scale and Acquisition Metadata Embedding at this higher resolution increases compute cost, raising inference latency to 69 ms and 64 ms respectively, while diverting parameter budget away from the convolutional layers responsible for fine texture extraction in the encoder and spatial reconstruction in the decoder. This trade-off is reflected in reduced Dice scores at both locations, most notably in the cortex and medulla. Confining the modules to the bottleneck avoids this cost entirely, achieving the highest Dice across all foreground classes (0.77 cortex, 0.50 medulla, 0.40 CEC) at the lowest inference latency (53 ms).

3) Cross-Validation vs SOTAs: We benchmark MASP-UNet against seven leading architectures using five-fold cross-validation. Class-wise Dice coefficients for Sonographer 1 are shown in Fig. 3, with error bars representing standard deviation across folds.

MASP-UNet's error bars remain narrow across most classes, indicating stable training. PSPNet edges ahead marginally on cortex (0.78 vs. 0.77 Dice) but requires ∼42M parameters against our ∼5M. This pattern holds across Sonographer 2 (∼0.71/0.51/0.42) and the randomized split (∼0.70/0.51/0.42) for cortex, medulla, and CEC respectively, indicating that MASP-UNet's performance is stable across annotator variability.

4) Independent Test Set Benchmark: To evaluate deployment under point-of-care constraints, we benchmark all models on the held-out 10% test set using 40 images at 256×256 resolution, batch size 1, FP32 precision, on a CPU-only MacBook Air M2 following a 10-iteration warm-up. Latency and power draw are sampled at 100 Hz via `powermetrics` and reflect the forward pass mean and standard deviation.

TABLE II
STRUCTURAL PARAMETERS, MODEL SIZE, AND DCES ON THE INDEPENDENT TEST SET (STANDARD DEVIATIONS (±) ARE COMPUTED ACROSS ALL THREE ANNOTATION CONFIGURATIONS)

Network	Background	Cortex	Medulla	CEC	Mean DCE	Trainable Params	Non-Train Params	Size (MB)
MASP-UNet	0.97 ± 0.01	0.73 ± 0.02	$\mathbf{0.50 \pm 0.02}$	0.40 ± 0.01	0.650 ± 0.01	5,765,130	6,718	22.02
Effnet-B0	0.95 ± 0.02	0.59 ± 0.02	0.40 ± 0.03	0.34 ± 0.03	0.570 ± 0.02	10,065,472	43,943	38.56
SegNet	0.96 ± 0.01	0.63 ± 0.03	0.40 ± 0.03	0.33 ± 0.02	0.580 ± 0.02	29,443,276	15,880	112.38
nnU-Net	$\mathbf{0.98 \pm 0.01}$	0.55 ± 0.02	0.34 ± 0.03	0.27 ± 0.02	0.535 ± 0.02	31,106,628	NA	118.66
PSPnet	0.97 ± 0.02	$\mathbf{0.76 \pm 0.01}$	0.47 ± 0.03	$\mathbf{0.41 \pm 0.01}$	$\mathbf{0.653 \pm 0.02}$	42,147,396	38,208	160.93
BDR-PCN	0.96 ± 0.01	0.51 ± 0.02	0.39 ± 0.02	0.32 ± 0.01	0.545 ± 0.02	62,624,201	25,088	238.99
UNeXT	0.95 ± 0.03	0.56 ± 0.04	0.41 ± 0.02	0.30 ± 0.02	0.555 ± 0.03	1,394,357	11,248	6.38
EMCAD	0.97 ± 0.01	0.49 ± 0.02	0.34 ± 0.01	0.27 ± 0.03	0.518 ± 0.02	1,910,381	NA	7.68

TABLE III
HARDWARE INFERENCE EFFICIENCY METRICS (MACBOOK AIR M2)

Network	Inf. Time (ms)	Peak Memory (MB)	Avg Power (W)	GFLOPs
MASP-UNet	53 ± 3	35–50	7.5 ± 0.2	4.7
Effnet-B0	83 ± 7	40–65	8.0 ± 0.7	7.1
SegNet	703 ± 6	70–110	11.4 ± 0.7	55.3
nnU-Net	552 ± 4	90–140	12.3 ± 0.6	29.9
PSPnet	964 ± 5	150–210	14.8 ± 0.4	73.6
BDR-PCN	1173 ± 3	180–260	16.7 ± 0.3	87.7
UNeXT	10 ± 2	20–30	4.5 ± 0.4	0.56
EMCAD	8 ± 3	15–25	4.2 ± 0.3	0.387

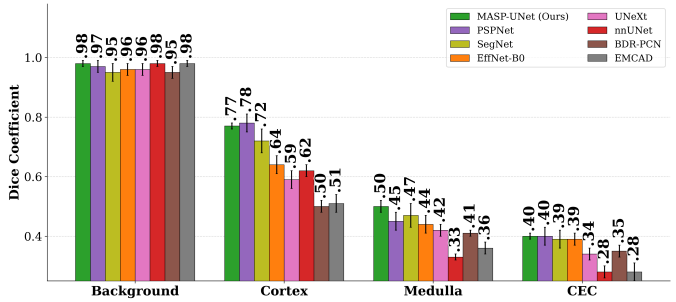


Fig. 3. MASP-UNet vs SOTAs (Sonographer 1, 5-fold CV).

Memory tracks peak Resident Set Size net of framework overhead. Results are in Tables II and III.

Heavy networks such as PSPNet achieve a marginally higher mean Dice (0.653 vs. 0.650) but at prohibitive cost, requiring 964 ms inference, 73.6 GFLOPs, and 14.8 W. Ultra-lightweight backbones such as UNeXt and EMCAD reach real-time latencies (10 ms, 8 ms) but degrade substantially on internal structures, with mean Dice of 0.555 and 0.518 respectively. MASP-UNet closes this gap, trailing PSPNet by only 0.003 mean Dice while reducing GFLOPs by $15.7\times$ and latency to 53 ms, and achieves the highest Dice of all models on the medulla (0.50), indicating that bottleneck-confined multi-scale aggregation can better capture fine renal structure without the parameter cost of full-network scaling.

V. GAPS AND FUTURE WORK

Although MASP-UNet achieves an $18.2\times$ lower latency than PSPNet, its GFLOP-to-latency scaling vs. ultralight baselines like UNeXt and EMCAD is suboptimal due to MCA dilated convolutions reducing CPU cache efficiency, motivating more hardware-friendly designs in future work. Moreover, while caching the frozen language model removes runtime latency, it still adds static initialization overhead, so we plan to replace the NLP pipeline with lightweight tabular MLP encoding in subsequent studies. DCE performance remains limited for medulla and CEC; we will therefore focus on improved fine-grained contrast modeling in future work. Grad-CAM (Fig. 4) further shows tightly localized activations in low-performing classes, indicating boundary-precision rather than localization failure, so our further research will focus on boundary-aware regression instead of larger receptive fields. Additionally, all experiments, including the metadata-conditioning ablation, were conducted on a single dataset with a 40-image test set; the metadata branch is evaluated for its in-distribution effect rather than against held-out acquisition hardware, so cross-device generalization and sensitivity to the prompt wording remain untested.

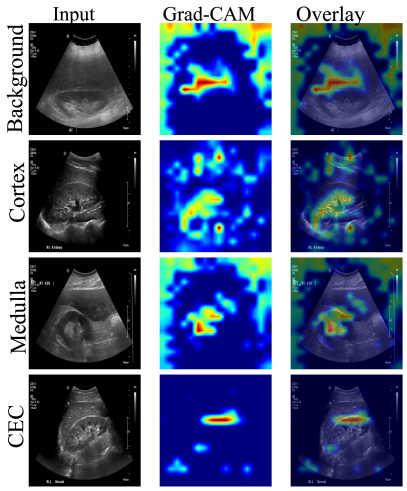


Fig. 4. Grad-CAM: best representative image per class, showing input, heatmap, and overlay for Background, Cortex, Medulla, and CEC.

VI. CONCLUSION

This paper introduces MASP-UNet, a lightweight network for point-of-care kidney ultrasound that confines multi-scale spatial processing & hardware metadata conditioning to a compact bottleneck, limiting deployment overhead while preserving global context. On a consumer-grade laptop (MacBook Air M2), it matches heavy SOTA models within a 0.003 mean Dice deficit at $15.7\times$ lower GFLOPs and 53 ms latency – suggesting that lightweight bottleneck-confined conditioning is a promising direction for preserving anatomical localization.

REFERENCES

[1] M. Y. Ansari, Y. Yang, P. K. Meher, and S. P. Dakua, "Dense-psp-unet: A neural network for fast inference liver ultrasound segmentation," *Computers in biology and medicine*, vol. 153, p. 106478, 2023.

[2] E. M. Remer, N. Papanicolaou, D. D. Casalino, J. T. Bishoff, M. D. Blaufox, C. A. Coursey, M. Dighe, S. C. Eberhardt, S. Goldfarb, and H. J. Harvin, "Acr appropriateness criteria® on renal failure," *The American Journal of Medicine*, vol. 127, no. 11, pp. 1041–1048, 2014.

[3] E. Widjaja, J. W. Oxtoby, T. L. Hale, P. W. Jones, P. N. Harden, and I. W. McCall, "Ultrasound measured renal length versus low dose ct volume in predicting single kidney glomerular filtration rate," *The British Journal of Radiology*, vol. 77, no. 921, pp. 759–764, 2004.

[4] M. Korkmaz, B. Aras, S. Güneşli, and M. Yılmaz, "Clinical significance of renal cortical thickness in patients with chronic kidney disease," *Ultrasonography*, vol. 37, no. 1, p. 50, 2018.

[5] R. Singla, M. Kadatz, R. Rohling, and C. Nguan, "Kidney ultrasound for nephrologists: A review," *Kidney Medicine*, vol. 4, no. 6, p. 100464, 2022.

[6] M. M. Rahman, M. Munir, and R. Marculescu, "Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 11 769–11 779.

[7] J. M. J. Valanarasu and V. M. Patel, "Unext: Mlp-based rapid medical image segmentation network," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds. Cham: Springer Nature Switzerland, 2022, pp. 23–33.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *MICCAI*, pp. 234–241, 2015.

[9] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE/CVF CVPR*, 2017, pp. 6230–6239.

[10] R. Vega, M. Dehghan, A. Nagdev, B. Buchanan, J. Kapur, J. L. Jaremko, and D. Zonoobi, "Overcoming barriers in the use of artificial intelligence in point of care ultrasound," *npj Digital Medicine*, vol. 8, no. 1, p. 213, 2025.

[11] R. Singla, C. Ringstrom, G. Hu, V. Lessoway, J. Reid, C. Nguan, and R. Rohling, "The open kidney ultrasound data set," in *Int. Workshop Adv. Simplifying Med. Ultrasound (ASMUS)*, 2023, pp. 155–164.

[12] D. Philippi, K. Rothaus, and M. Castelli, "A vision transformer architecture for the automated segmentation of retinal lesions in spectral domain optical coherence tomography images," *Sci. Rep.*, vol. 13, p. 517, 2023.

[13] T. Sugino, Y. Saga, T. Sato, H. Abe, K. Onodera, Y. Matsuo, A. Takeuchi, T. Ishida, S. Ogawa, Y. Matsuo, A. Kurosaki, K. Matsushima, S. Maeda, Y. Ota, T. Aoki, T. Yamada, S. Kikuchi, M. Ito, Y. Kurisu, R. Okazawa, K. Fukumoto, Y. Ogasawara, H. Tamura, M. Hattori, and N. Nakamichi, "Loss weightings for improving imbalanced brain structure segmentation on mr images," *Frontiers in Neuroscience*, vol. 15, p. 676180, 2021.

[14] S. A. Taghanaki, K. Abhishek, J. A. Cohen, S. Cohen, M. G. Myronenko, B. w. Lin, S. M. Taghanaki, I. Ghafoorian, and G. Hamarneh, "Combo loss: Handling input and output imbalance in multi-organ segmentation," *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.*, vol. 9, no. 1, pp. 49–57, 2021.

[15] N. Abraham and N. M. Khan, "A novel focal tversky loss function with improved attention u-net for lesion segmentation," *Proc. IEEE Int. Symp. Biomed. Imaging (ISBI)*, pp. 683–687, 2018.

[16] A. Mortazi, B. Abidi, Z. J. Wang, and M. E. A. Abdolali, "Selecting the best optimizers for deep learning–based medical image segmentation," *Journal of Medical Imaging*, vol. 10, no. 5, p. 054502, 2023.

[17] M. M. Rahman and R. Marculescu, "Ultralightunet: Rethinking u-shaped networks with multi-kernel lightweight convolutions for medical image segmentation," OpenReview preprint, 2025, withdrawn ICLR 2025 submission. [Online]. Available: https://openreview.net/forum?id=BefqqrgdZ1

[18] S. Ouyang, J. Zhang, Y. Xiang, and X. Tang, "Swind-net: A lightweight segmentation network for laparoscopic liver segmentation," *Computer Assisted Surgery*, vol. 29, no. 1, p. 2329675, 2024.

[19] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Köhler, T. Norajitra, S. Wirkert, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, pp. 203–211, 2021.

[20] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[21] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 6105–6114.

[22] S. Yin, Q. Peng, H. Li, Z. Zhang, X. You, K. Fischer, S. L. Furth, G. E. Tasian, and Y. Fan, "Automatic kidney segmentation in ultrasound images using subsequent boundary distance regression and pixelwise classification networks," *Medical Image Analysis*, vol. 60, p. 101619, 2020.