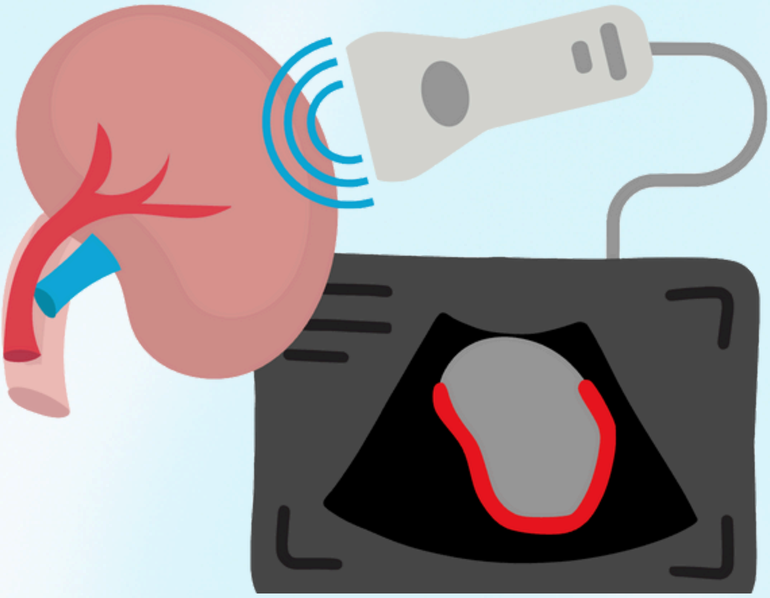
MASTER THESIS

Towards Enhanced Surgical Guidance:

Feasibility of Deep Learning–Based Segmentation for Intra-operative Kidney Ultrasound Registration



Julia Ouwerkerk
September 2025

Delft University of Technology
Master of Science BioMedical Engineering
Track Medical Physics

Master Thesis

Towards Enhanced Surgical Guidance:

Feasibility of Deep Learning–Based Segmentation for Intra-operative Kidney Ultrasound Registration

Julia, J. Ouwerkerk
Student number: 5053528
17 September 2025

Thesis in partial fulfilment of the requirements of

Master of Science in BioMedical Engineering
Track: Medical Physics

at Delft University of Technology
to be defended publicly on Wednesday, September 17th, 2025 at 10:00

Thesis committee:

First Assessor and Chair:	Prof.dr. J.J. van den Dobbelsteen,	TU Delft
Second Assessor:	Prof.dr. B.H.W. Hendriks,	TU Delft
Daily Supervisor:	MSc M. A. J. Hiep,	NKI AVL

Abstract

This study investigated the feasibility of deep learning-based segmentation in intra-abdominal kidney ultrasound registration, to enable image-guided robotic-assisted partial nephrectomy (RAPN). Two state-of-the-art models, DeepLabV3+ and SAMUS, were trained and evaluated using a novel intra-abdominal kidney ultrasound (IAKUS) dataset of 2,265 images from 15 RAPN patients. Moreover, a transfer-learning approach was adopted using the publicly available open kidney ultrasound (OKUS) dataset for pre-training. Results showed that SAMUS consistently outperformed DeepLabV3+ across all metrics, achieving an average Dice score of $88.0 \pm 2.0\%$ and Hausdorff distance of 13.7 ± 3.8 mm, consistent with literature. SAMUS was pre-trained on $\sim$ 30k ultrasound images which enabled a zero-shot test, outperforming trained DeepLabV3+ configurations. Furthermore, no measurable difference was seen between OKUS and IAKUS training. Both findings suggest ultrasound specific features may be more important than organ specific features for training, and data diversity may be more important than strict anatomical similarity. The SAMUS model obtained a registration accuracy of 4.3 ± 2.8 mm and inference time of 4.35 fps, in line with literature reported clinical feasibility. The target registration was even improved by an average of 2.6 ± 4.2 mm in 11/13 patients compared with manual-based registration. This proves that deep learning-based registration is not only feasible in a clinical setting but exceeds manual-based registration. **Key words: ultrasound, robotic-assisted partial nephrectomy, surgical navigation, deep learning, segmentation, registration**

Contents

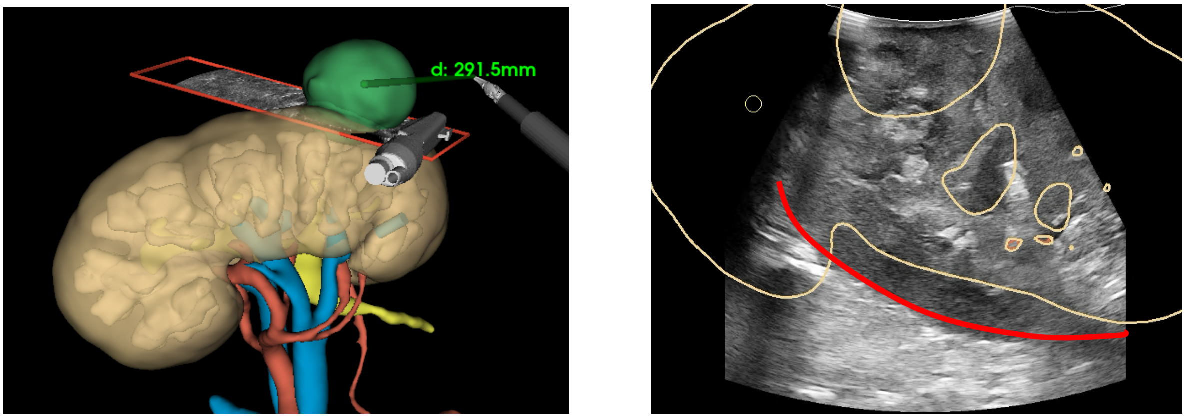
Appendices:

List of Abbreviations

PN	Partial nephrectomy
RAPN	Robotic-assisted partial nephrectomy
US	Ultrasound
CT	Computed tomography
DL	Deep learning
DLV3+	DeepLabV3+
SAMUS	Segment Anything Model - Ultrasound
IAKUS	Intra-abdominal kidney ultrasound
PKE	Posterior kidney edge
OKUS	Open kidney ultrasound
HD-95	Hausdorff Distance (95th percentile)
ASSD	Average symmetric surface distance
Dice	Dice similarity coefficient
FRE	Fiducial registration error
TRE	Target registration error
SC	Trained from scratch
PR	Pre-trained

1 Introduction

Partial nephrectomy (PN) has become the standard of care, alongside radical nephrectomy, for treatment of renal cell carcinoma with small renal masses, including T1a ($\leq$ 4 cm), T1b ($>$ 4 cm but $\leq$ 7 cm), and some T2 ($>$ 7 cm) tumors. Compared with radical nephrectomy, PN preserves renal function, reduces surgical trauma, and shortens hospitalization [1]. PN can be performed either via open, laparoscopic, or robotic-assisted approaches. Although laparoscopic PN has demonstrated several advantages over open PN, like reduced intra-operative blood loss and shorter hospital stay , it is typically associated with a longer operative time [2]. In comparison, robotic-assisted partial nephrectomy (RAPN) provides similar intra-operative benefits to laparoscopic PN, while also reducing warm ischemia time and enhancing postoperative renal function, without compromising short-term oncological or functional outcomes. Reflecting this shift towards minimally invasive techniques, a multicentre study of 7,869 PNs performed between 2010 and 2020 reported that 45.5% were robotic-assisted, 20.8% were laparoscopic, and 33.7% were open [2].



(a)

(b)

Figure 1: (a) Example of the 3D surgical navigation viewer. The surgeon can visualize important anatomical structures (beige: kidney surface, green: lesion, red: artery, blue: vein, yellow: ureter) in relation to surgical tools such as the ultrasound transducer and scissors. Distances between the surgical tools and the lesion are simultaneously displayed (green text). (b) Example of the ultrasound with an overlay of the registered kidney (beige) using the manual boundary selection. As seen the manual method currently provides an inaccurate registration, the actual posterior edge (red) does not match the registration.

During RAPN, a drop-in ultrasound (US) probe is deployed through a trocar port inside the patient abdomen and placed directly on the kidney, which is used for tumor localization, depth assessment and identification of blood vessels within the organ. However, during tumor excision, the surgeon relies on visual feedback from the robotic camera and memory instead, which limits the ability to assess depth and cutting angle [3]. This is especially challenging for endophytic tumors, as their inward growth provides fewer visual landmarks [4]. In order to provide real-time navigational cues during resection 3D virtual models, mixed reality, and augmented reality have been proposed. These navigation tools allow the surgeon to visualize anatomical structures and measure critical distances during the resection, as illustrated in Figure 1a, effectively addressing limitations posed by relying solely on traditional camera views [5–7].

For the navigation-guided surgery, a pre-operative 3D model of the patient is created using computed tomography (CT), including the kidney, tumor, cysts, ureter, and blood vessels. The model is used for preoperative planning and forms the basis of the surgical navigation viewer (Figure 1a). As patient anatomy can be significantly altered due to organ movement and influences of radiation or chemotherapy, the pre-operative model must be registered to intra-operative US to align anatomical structures accurately between both datasets [8]. US is preferred over cone-beam CT for intra-operative imaging due to its lower costs, minimal disruption to the clinical workflow, lack of radiation exposure, and ability to allow repeated re-registration without requiring a hybrid operating room [4]. During the procedure, points along the posterior kidney edge (the boundary furthest from the US probe) are manually selected, which are registered to the pre-operative model, transforming the 3D surgical navigation viewer. However, manual boundary selection is time-consuming, labor intensive, and prone to significant intra- and inter-observer variability, which affects the registrations consistency and accuracy (Figure 1b) [9].

To address these challenges, automatic segmentation methods have been considered. Nevertheless, this task can be challenging due to modality-specific issues like low contrast, uneven energy distribution, speckle noise, and overall low image quality [9]. Moreover, kidney-specific factors, such as blurred boundaries due to soft tissue transitions and variable object shapes, can further complicate segmentation [10–13]. Therefore, the aim of this study is to evaluate the feasibility of using deep learning (DL) models to automatically segment the posterior renal boundary in intra-abdominal US images and to assess their application for image registration.

1.1 Literature study

A systematic literature study was performed on (semi)-automatic segmentation techniques of renal US images between 2017 and 2024. Traditional methods were often user dependent, time-consuming to initialize, and performed poorly in real-time applications due to edge-based limitations and high inter-operator variability. Therefore, AI-assisted models became preferable over traditional image processing methods because of the need for automatic and precise applications [9, 11]. In line with this development, state-of-the-art DL models were compared, where U-Net remains the most widely used, though its performance varies across datasets [10, 14]. DeepLabV3+ (DLV3+) obtained superior results over other bench-marked models (including U-net), reaching a Dice of 89.76 %, Hausdorff distance (HD) of 9.91 mm, average symmetric surface distance (ASSD) of 3.03 mm, and an accuracy of 98.14 % [10]. Newer proposed DL models implement multi-scale feature fusion, often in combination with attention enhancement [15–20]. In addition, hybrid models often implemented U-net for the segmentation stage and followed different refinement techniques for the kidney edge [21–23]. Other studies implemented transfer learning techniques relying on U-Net, DeepLab or variants, and focused more on how the DL networks are trained, rather than the structural components of the networks. Approaches include cross-modal, intra-modal, and partial transfer learning [11–14, 19, 24–26]. All identified methods obtained good performance in terms of the Dice score (Figure 2). However, due to differences in datasets, data splitting techniques, and the large variation in applied metrics, no conclusion could be drawn about which method is superior.

A common challenge identified in the literature was the lack of a large publicly available dataset. In attempt to overcome this limitation, two recent studies constructed and publicly published large

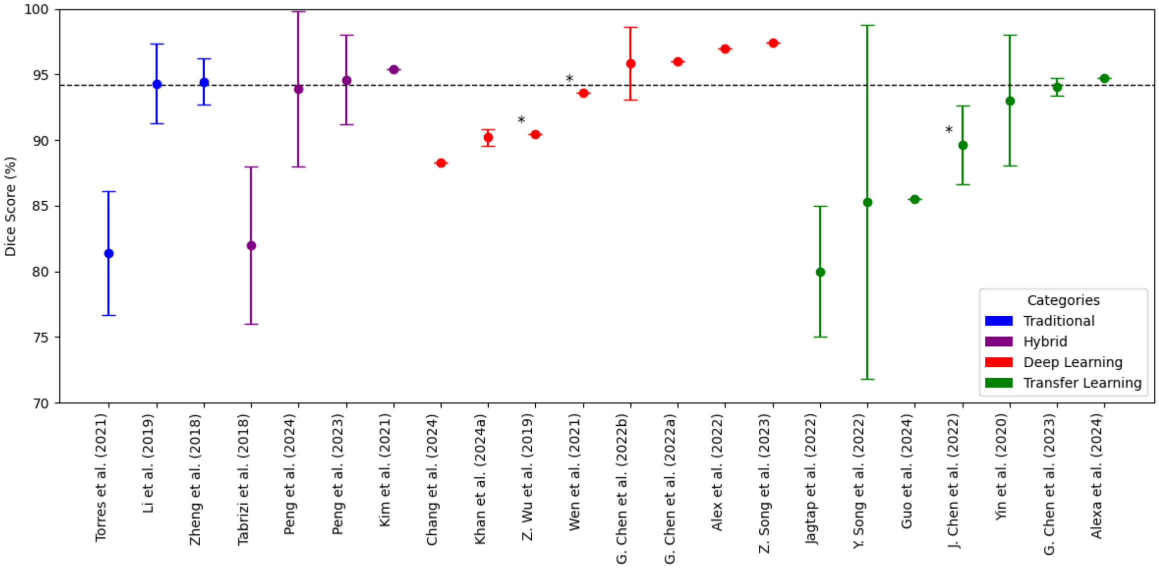


Figure 2: Performance of all the identified novel models categorized by methods in terms of the dice score, as this was the most commonly identified metric. Dice scores calculated from mean intersection over union are indicated with an asterix (*) on the left side. Error bars represent the standard deviation as reported in the respective studies. The dotted line indicates the mean value across cases. All models obtained competitive Dice scores, and no conclusion could be drawn as to which model is better as all models are trained on different datasets.

datasets, with 630 images of 500 patients and 44,880 images of 148 patients [18, 27]. Other methods proposed to overcome this limitation were applying data augmentation, implementing transfer-learning techniques [12, 13, 24–26], or creating training datasets from largely available renal CT data using cycle-generative adversarial networks [13]. Additionally, although most models were created to overcome segmentation accuracy disturbances caused by kidney morphology, heterogeneous structures and image quality, the models still exhibited susceptibility to these issues. Furthermore, the models had limited clinical applicability due to iterative processes [20], step-by-step mechanisms [15], or extensive pre- and post-processing steps [10, 16]. Although there is no established norm of inference time for clinical implementation, 10-20 frames per second (fps) is reported as ideal for continuous guidance [28, 29]. Finally, model availability was limited. Of the novel DL architectures, only MLAU-Net [18] was released publicly (CC BY-NC-ND license). Among transfer learning approaches, NU-Net [24] and DeepLabV3+ [25] were accessible via GitHub.

Most importantly, no research was published on the application of any of these models to intra-abdominal kidney US segmentation. All the mentioned models rather relied on standard trans-abdominal US, which is not feasible during RAPN due to the operating conditions due to the pneumoperitoneum and patient orientation [30]. This highlights a significant gap in the literature and the need for further research in this area.

1.2 Model selection

To evaluate the effectiveness of DL-based segmentation for registration, two state-of-the-art models, DLV3+ and SAMUS, were selected. Recent studies on trans-abdominal renal US segmentation reported that DLV3+ outperformed other bench-marked models [10, 14]. This is a semantic segmentation architecture that extends DeepLabV3 by adding a decoder module to improve spatial localization, especially along object boundaries. It consists of a backbone for the feature extraction, parallel atrous spatial pyramid pooling, and a decoder module followed by the final prediction layer [24, 31].

Since intra-modal transfer learning has been proven effective, models like SAMUS, pre-trained on a wide range of US data, may be used for kidney segmentation [24]. Lin et al. proposed SAMUS, a version of the Segment Anything Model (SAM), developed by Meta AI designed for universal image segmentation, tailored for US image segmentation [32]. SAM faced challenges in medical imaging, such as limited generalizability and dependence on accurate manual prompts. Key modifications consisted of a convolutional neural network branch, a vision transformer (ViT) branch and cross-branch attention. The model was further adapted to work in an end-to-end manner by addition of an auto prompt generator. SAMUS demonstrated strong performance obtaining a Dice of 83.1 % and HD of 28.8 mm on thyroid nodules, 84.5 % and 27.2 mm on breast cancer, and 83.1 % and 19.0 mm on myocardium data, surpassing prior SAM performances [33].

For both models, transfer learning strategies were adopted based on prior findings in the literature. For DLV3+, a partial transfer learning approach was used, employing a ResNet-50 backbone pre-trained on ImageNet to enhance feature representation and accelerate convergence. For SAMUS, the intra-modal transfer learning strategy was adopted, using pre-trained weights available from $\sim$ 30k images of publicly available datasets including thyroid nodules (TN3K), breast cancer (BUSI), left ventricle (CAMUS-LV), myocardium (CAMUS-MYO), and left atrium (CAMUS-LA).

1.3 Research goals

The goal of this study was to evaluate the performance of two DL models, DLV3+ and SAMUS, for boundary delineation of the kidney in intra-abdominal US images. Both models were trained on intra-abdominal kidney US data with the aim of improving surgical navigation during RAPN. Therefore the best-performing model was evaluated for its feasibility in registration tasks, where its performance was be compared against manual-based segmentation registration. This evaluation was motivated by the potential to reduce the time required for manual-based registration, which often takes several seconds per frame, and mitigate inter- and intra-observer variability.

2 Method and materials

2.1 Study design

This study is a sub-study of the prospective image-guided navigation during RAPN study at the Netherlands Cancer Institute, Amsterdam, the Netherlands. Patients who underwent RAPN between July 2024 and June 2025, 18 years or older and provided written informed consent were eligible for inclusion. The protocol was approved by the medical ethics committee (trial number: NL86425.041.24). Data were obtained during RAPN, with procedures performed on the da Vinci Xi surgical system (Intuitive Surgical, Inc., Sunnyvale, USA) and navigation displayed on TilePro [34]. Scissors and a drop-in US probe (BK Medical, Herlev, Denmark) were tracked using the Aurora electromagnetic system (Northern Digital Inc., Waterloo, Canada), which was also used to register the navigation system to the kidney. Additionally, a sensor was secured within the renal parenchyma in proximity to the lesion to compensate for intra-operative kidney motion. Data was stored as US recordings, starting at vessel exploration and ending after tumor excision.

2.2 Datasets

2.2.1 The intra-abdominal kidney ultrasound dataset

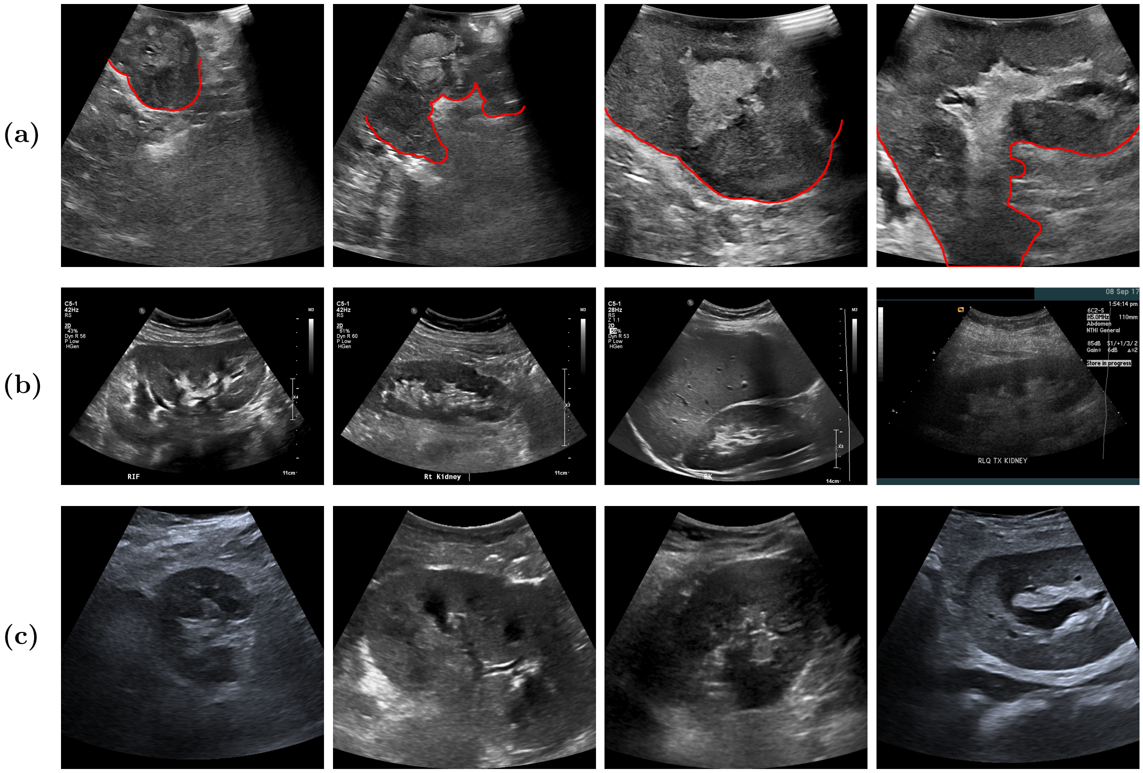


Figure 3: (a) Examples from the intra abdominal kidney ultrasound (IAKUS) dataset, with the posterior kidney edge overlaid in red. Reflecting different anatomical views within one patient. (b) Examples from the open kidney ultrasound (OKUS) dataset, reflecting different images sizes and user interface overlays. (c) Examples of pre-processed OKUS frames, adjusted to more closely resemble IAKUS frames.

The intra-abdominal kidney ultrasound (IAKUS) dataset was created retrospectively from recordings obtained during RAPN. The dataset comprised 2265 intra-operative renal US images collected from fifteen patients, with an average of 151 ± 25 images per patient. An example of the dataset is shown in Figure 3a. Since the recordings did not include a complete sweep of the kidney, a fixed frame rate could not be used for frame selection. Therefore, frames were manually extracted from the recordings to be included in the dataset. Minimal pre-processing was applied to ensure that the training data closely resembled the real-time data encountered in the operating room. However, DeepLabV3+ and autoSAMUS required resizing of 794 x 794 to 384 × 384 and 256 x 256 pixels, respectively, to reduce memory usage and minimize fragmentation artifacts. Ground truth annotations were created using 3DSlicer with a custom SingleSliceSegmentation module. Both kidneys and tumors were annotated, with contours drawn within the surrounding fat layer to capture the organ boundary. All masks were exported as binary labelmaps for subsequent model training and evaluation. Thirteen patients were processed by the author (J.O) a BioMedical Engineering student, and two patients were processed by supervisor (M.H) MSc Technical Medicine. Data augmentation was applied to increase the dataset, as it was proposed as a method to overcome limitations of data scarcity. Therefore, augmentation was applied by randomly selecting from horizontal flip ($p = 0.5$), scaling in [0.9, 1.1], rotation in $[-10°, +10°]$, color jitter (brightness = 0.2, contrast = 0.2), and random cropping to 384 × 384 with padding. Images were converted from grayscale to RGB before transforms and normalized to $[-1, 1]$ (mean = 0.5, std = 0.5 per channel).

2.2.2 The posterior kidney edge datasets

An additional dataset focusing on the posterior kidney edge (PKE) was derived from the IAKUS ground truth annotations using an in-house developed algorithm. First all connected components were identified and their outermost x-values were determined. Next, the bottom contours were segmented. Finally, if multiple objects were present, the two outer base objects were selected, and only the objects located below them were considered. The same algorithm was also applied to the models predictions to enable focused analysis of segmentation performance in the posterior region (Figure 3a).

2.2.3 The open kidney ultrasound dataset

The open kidney ultrasound (OKUS) dataset was obtained from Singla et al. (2022), which was used to pre-train the models. This dataset was created retrospectively and comprised 510 trans-abdominal B-mode renal US images from 500 patients with a mean age of 53.2 ± 14.7 years, body mass index of 27.0 ± 5.4 kg/m2, and most common primary diseases being diabetes mellitus, immunoglobulin A nephropathy, and hypertension [27]. The dataset was acquired using a wide range of US systems and transducers, resulting in varying image sizes, field of view, and anatomical presentation (Figure 3b). As a result, extensive pre-processing was necessary to standardize the data format and remove text overlays to match the IAKUS dataset (Figure 3c). This was done by generating a mask of the US beam consistent for IAKUS images, and rescaling the OKUS images until beam sizes were equal. The mask was applied to remove any text and user interface symbols outside the beam. Afterwards, the frames were cropped to 794 x 794 pixels at the mask location to ensure size consistency. Ground truth annotations were performed in the same manner as for the IAKUS dataset.

2.3 Soft- and Hardware

Annotations were performed using 3D Slicer [35], an open-source platform for biomedical image analysis and visualization. The models were altered, trained, and evaluated in Python [36] using PyCharm Community Edition 2024.3.3 [37] as integrated development environment. Since training a DL model requires large computational power, an external GPU was used. Specifications of each of these components are detailed in Appendix A.

2.4 Model application

The original DLV3+ model was introduced by Chen et al. [31], the PyTorch version by Viktor Zhou was used for this study [38]. The architecture was used as originally proposed, with only minor modifications to ensure compatibility with the dataset. These include: creating a new class for our dataset, enabling data-splitting for K-fold cross-validation, enabling early stopping to prevent overfitting (based on dice and validation loss), applying online data augmentations, incorporating boundary-specific metrics for evaluation and testing, and implementing boundary specific loss functions. Additionally, Python codes were added to run sanity checks, to allow distance based losses, to perform grid-searches, and to allow training over all folds. During finetuning, different loss functions were tested, including focal loss, dice loss, cross-entropy loss, boundary loss [39], Hausdorff loss [39], generalized surface loss [40] and combinations between them. From these loss functions, the best three options were selected based on boundary performance. Next, model parameters were finetuned using a grid-search for the validation batch size, training batch size, crop size (augmentation), and learning rate. During these trials, other parameters remained consistent and one fold was considered to ensure consistency. The finetuning process is further detailed in Appendix B. The cross-entropy (CE) loss and boundary (B) loss were selected as best performing configurations, with a training batch size of 8, validation batch size of 4 and learning rate of 0.005 and 0.001, respectively. The backbone was frozen for two epochs and CE was used as a warm-up for the first three epochs for the B loss to improve training stability.

The SAMUS architecture was also used as originally proposed, with only minor modifications to ensure compatibility with the dataset. Which include the same adaptations as detailed above, except for the addition of boundary-specific loss functions. The ViT-backbone variant ViT-Base, containing 12 layers, 768 dimensions, and $\sim$ 86 million parameters, was selected [41]. The available weights from pre-training were used as model initialization, following the intra-modal transfer learning approach. The SAMUS model was not finetuned due to time-restrictions, CE was selected as loss-function.

2.5 Training and testing

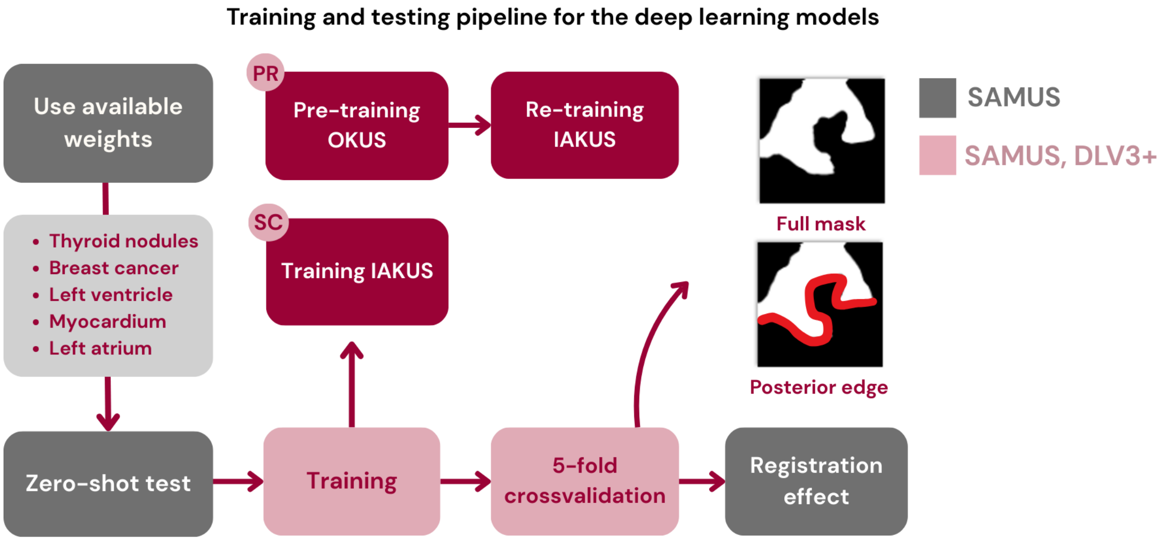


Figure 4: SAMUS started with a zero-shot test (gray), since the model had pre-trained weights available from publicly accessible ultrasound datasets. Both SAMUS and DeepLabV3+ (DLV3+) folloed the same training and testing strategies. Training was performed either directly on the intra-abdominal kidney ultrasound dataset (IAKUS), referred to as training from scratch (SC), or by pre-training on the open kidney ultrasound dataset (OKUS) followed by fine-tuning on IAKUS, referred to as pre-training (PR). Testing was done using a five-fold cross-validation on both the full masks and the posterior kidney edges. Finally, the best model was used to test the effect on registration.

The SAMUS and DLV3+ followed identical training and testing approaches, as illustrated in Figure 4. For SAMUS, however, pre-trained weights were available from training on various US datasets. This enabled an additional zero-shot evaluation, where the model was tested on the IAKUS data without any fine-tuning on that specific dataset.

Both models underwent two training strategies: from scratch (SC) using IAKUS, and pre-training on OKUS followed by fine-tuning on IAKUS (PR). Here, SC refers to training starting from the default weight initialization provided for each model, rather than from randomly initialized weights. The predictions were then evaluated using 5-fold cross-validation, where data was split on a patient level, in which for every fold 20 % (three patients) of the data was separated for testing. The performance was evaluated using full-mask and PKE predictions, and the effect of using the DL-based segmentation for registration was compared with manual boundary selection.

2.6 Segmentation accuracy

Region-based metrics were included to allow comparison with previously bench-marked models, providing an overall assessment of segmentation overlap. Values range from 0 to 100%, with a higher value indicating better predictions. Below, each metric is formally defined.

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP, TN, FP, and FN denote the true positives, true negatives, false positives, and false negatives, respectively. This reflects the ratio of correctly predicted pixels to the total pixels [10].

Dice Similarity Coefficient (Dice):

$$\text{Dice} = \frac{2 \cdot |A \cap B|}{|A| + |B|} = \frac{2TP}{2TP + FP + FN} \tag{2}$$

where A and B denote the predicted and ground truth segmentation. The Dice reflects the area of intersection between these segments, divided by the total number of pixels in them [10].

Contour-based metrics were additionally selected to offer a more precise evaluation of segmentation quality, with respect to the clinical interest in this study. Distances were measured in pixels and converted to millimeters using the pixel spacing (0.15 mm/pixel), with lower values indicating better predictive performance. Below, each metric is formally defined.

95th Percentile Hausdorff Distance (HD-95):

$$\text{HD}_{95}(A, B) = \max\left\{\text{percentile}_{95}\left(\{d(a, B) \mid a \in A\}\right),\ \text{percentile}_{95}\left(\{d(b, A) \mid b \in B\}\right)\right\} \tag{3}$$

where $d(x, Y) = \min_{y \in Y} \|x - y\|$ is the minimum Euclidean distance from point x to set Y. HD-95 reflects the 95th percentile of the distances from points on one boundary to their closest points on the other boundary in both directions [10].

Average Symmetric Surface Distance (ASSD):

$$\text{ASSD}(A, B) = \frac{1}{|A| + |B|}\left(\sum_{a \in A} d(a, B) + \sum_{b \in B} d(b, A)\right) \tag{4}$$

where A and B are the contours of the predicted and ground truth segmentations. ASSD reflects the average mutual distance between all points on these contours [10].

2.7 Registration accuracy

For the evaluation of the registration, a separate test set was created from the IAKUS data, where frames with available registration were only included. Two patients were excluded from this analysis, as no registration data or landmarks were available. These frames were passed through the best-performing model (SAMUS) to retrieve the predicted masks, from which five fiducial points along the PKE per frame were selected. The fiducials formed a point cloud representation of the surface, which was registered to the kidney model using 3DSlicer, resulting in a rigid registration transformation. The registration was evaluated using:

Fiducial Registration Error (FRE):

$$\mathrm{FRE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|Tp_i - q_i\|^2} \tag{5}$$

FRE is the root mean square of point-to-surface distances after registration. In which p_i are the fiducial points selected in the US space, q_i current nearest surface point on the CT-based kidney-tumor model, T the registration transformation and N the number of fiducial across all US frames. FRE is measured in mm, a lower value indicates the sampled posterior surface lies close to the CT surface under that transform [42, 43]. Literature indicates an FRE in the range of 2-5 mm is clinically acceptable in a kidney phantom study [44].

The surgeon selected landmarks that represented the kidney-tumor border in the physical space using an EM tracked tool. The registration transformation was applied to the landmarks, and distances of these landmarks to the CT-based tumor model were used to evaluate the accuracy of the registration.

Target Registration Error (TRE):

$$\mathrm{TRE\ (LM)} = \sigma(LM) \min_{x \in \mathcal{S}} \parallel LM - x \parallel_2 \tag{6}$$

The TRE is defined as the euclidean distance (mm) of transformed landmarks to the tumor. Where LM are the landmarks, and S represents the tumor surface. To capture whether landmarks were located inside (-) or outside (+) the tumor, a signed distance function was used, where $\sigma(LM) \in \{-1, +1\}$ sets the sign.

$$\Delta\ \mathrm{TRE} = TRE_{DL} - TRE_{Manual} \tag{7}$$

Δ TRE represented the difference in TRE of specific LMs between the DL- and manual-based registrations. A negative Δ TRE indicates a more accurate registration, meaning the landmark was localized closer to the lesion.

$$\mathrm{MAD_{TRE}} = \frac{1}{N}\sum_{i=1}^{N} |\mathrm{TRE}_i| \tag{8}$$

Finally, the mean absolute distance (MAD) of the TRE was evaluated.

Studies reported ∼3–4 mm TRE in kidney phantom setups as clinically acceptable [45]. Moreover, TRE and FRE have been shown to be uncorrelated, indicating that although FRE provides relative fiducial information, it does not reflect registration accuracy like the TRE [46, 47].

3 Results

3.1 Patient characteristics

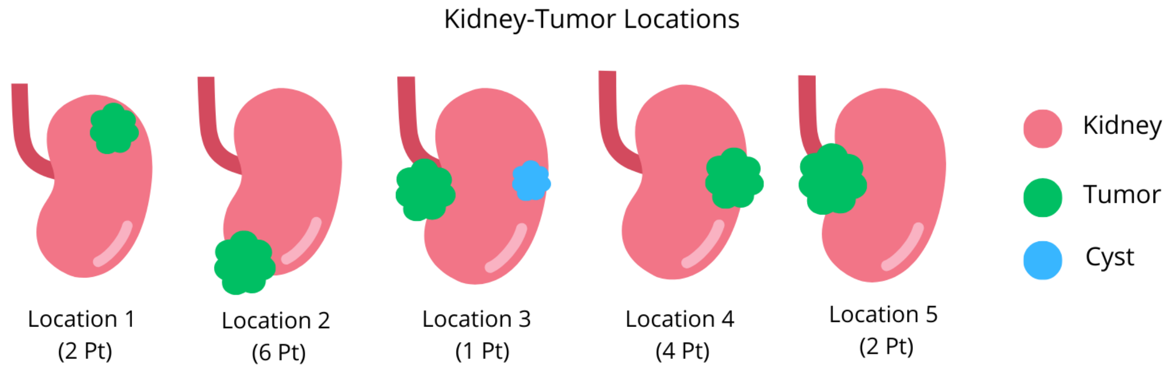


Figure 5: Overview of the five recognized kidney–tumor locations observed across patients. Each schematic illustrates the relative position of the tumor (green) or cyst (blue) with respect to the kidney (pink). Numbers below each location indicate how many patients were identified within. Tumor size in the schematics is not to scale and is shown only to indicate location.

Fifteen patients (7 males and 8 females) with a mean age of 58.8 ± 8.2 years were included in this study. Among them, 9 patients (60%) had tumors located on the left kidney, and 6 (40%) on the right kidney. In one case, a cyst was also present on the same kidney as the tumor. Tumor staging showed 11 cases of pT1a, 3 cases of pT1b, and 1 case of metanephric adenoma, located on varying parts of the kidney surface (Figure 5). The maximum tumor diameter on pre-operative CT ranged from 1.7 to 6.1 cm, with a mean of 3.3 ± 1.2 cm. The tumor volume ranged from 4.12 to 94.6 cm^3 pre-operatively (mean 24.5 ± 23.9 cm^3), and from 2.9 to 101.1 cm^3 post-operatively (mean 21.6 ± 24.3 cm^3).

3.2 Segmentation performance

3.2.1 Zero-shot test

The SAMUS (SC) model achieved an average HD-95 of 13.7 mm, representing a 10.4 mm improvement from the zero-shot test. Moreover, SAMUS obtained an average Dice of 88.0%, a 20% improvement compared with zero-shot testing. Notably, even without training on kidney-specific ultrasound data, SAMUS obtained an average HD-95 of 24.1 ± 0.9 mm, which was comparable with DLV3+ performance after training (CE-SC: 21.9 ± 3.4 mm; CE-PR: 22.2 ± 4.8 mm).

3.2.2 Full mask predictions

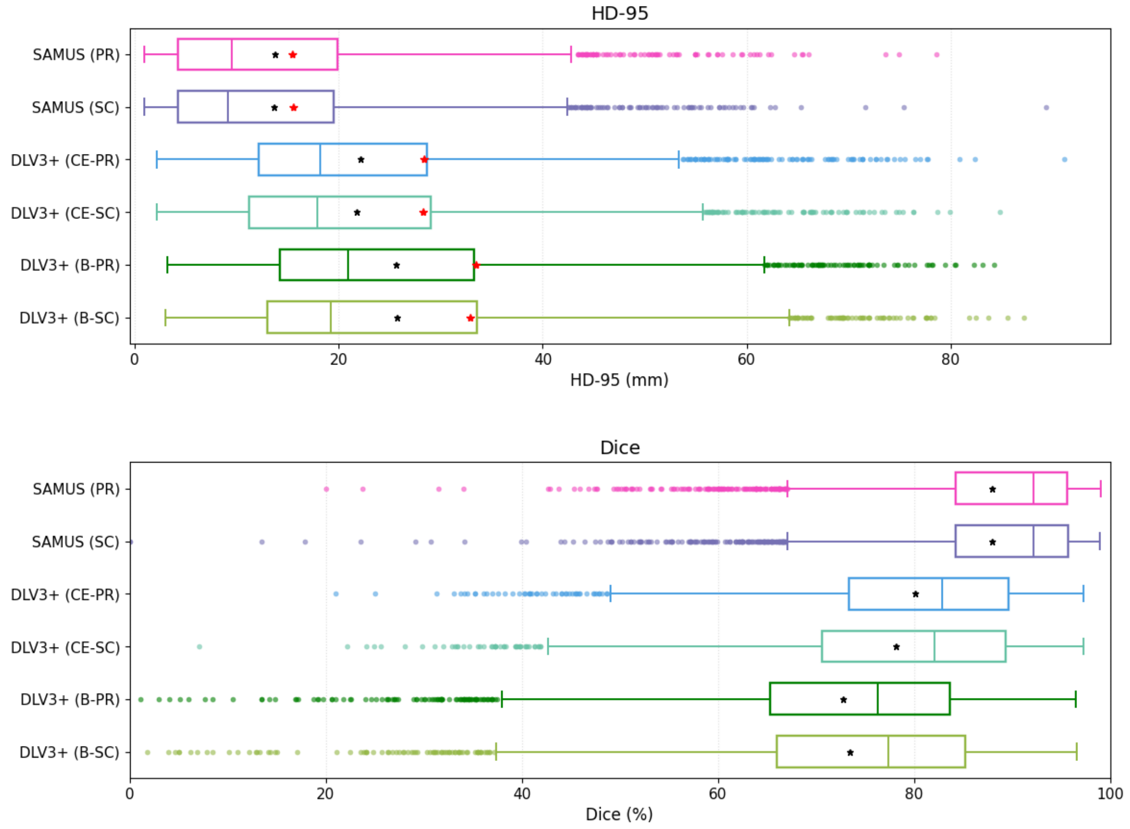


Figure 6: Performance on pooled image level across models. The black asterisk (*) marks the mean across folds, and the red asterisk marks the mean of the posterior kidney edge evaluation. The pooled median is marked with a vertical line within the IQR. Where SAMUS (PR) is resembled in pink, SAMUS (SC) in purple, DLV3+ (CE-PR) in blue, DLV3+ (CE-SC) in light blue, DLV3+ (B-PR) in dark green and DLV3+ (B-SC) in light green.

SAMUS-based models consistently outperformed DLV3+-based models across all metrics (Figure 6). Both SAMUS (SC) and SAMUS (PR) obtained lower mean HD-95 values of 13.7 ± 3.8 mm and 13.8 ± 3.9 mm, respectively, compared with 21.9 ± 3.4 mm for the best-performing DLV3+ variant (CE-SC). Notably, pooled median values fell considerably lower, at $\sim$ 9 mm for SAMUS and $\sim$ 19 mm for DLV3+. Furthermore, variability and outliers were apparent across all models, with SAMUS and DLV3+ yielding similar number of outliers ($\sim$ 80). In the best SAMUS configuration (SC), the whisker range reached 40 mm, with individual outliers exceeding 70 mm, although the 75th percentile remained below 20 mm. A similar trend was observed for the ASSD (Appendix C). SAMUS-based methods reached a mean of 3.5±1.0 mm, approximately half of the DLV3+ variants, but also obtained $\sim$ 1.4x more outliers ($\sim$ 165).

Moreover, SAMUS obtained superior average Dice scores of 88.0 ± 2.8% (SC) and 88.2 ± 3.0% (PR), compared to 78.2 ± 5.0% (CE-SC) and 80.0 ± 3.6% (CE-PR) for the best DLV3+ configurations (Figure 6). Median Dice values for SAMUS even exceeded 90%. SAMUS-based models also demonstrated more consistent performance, achieving a smaller inter quartile range (IQR $\approx$ 11.5%) and narrower whisker range ($\approx$ 33%) compared with the best DLV3+ variant (IQR $\approx$ 12.2%, whisker $\approx$ 40%). However, SAMUS obtained approximately three times as many of outliers ($\sim$ 150). Accuracy results were comparable to the Dice score; however, SAMUS produced $\approx$ 1.4x the number of outliers observed for DLV3+ (Appendix C). Finally, performances of SAMUS-SC and SAMUS-PR were largely comparable, as were those of and DLV3+ (CE) and DLV3+ (PR) configuration.

3.2.3 Qualitative performance

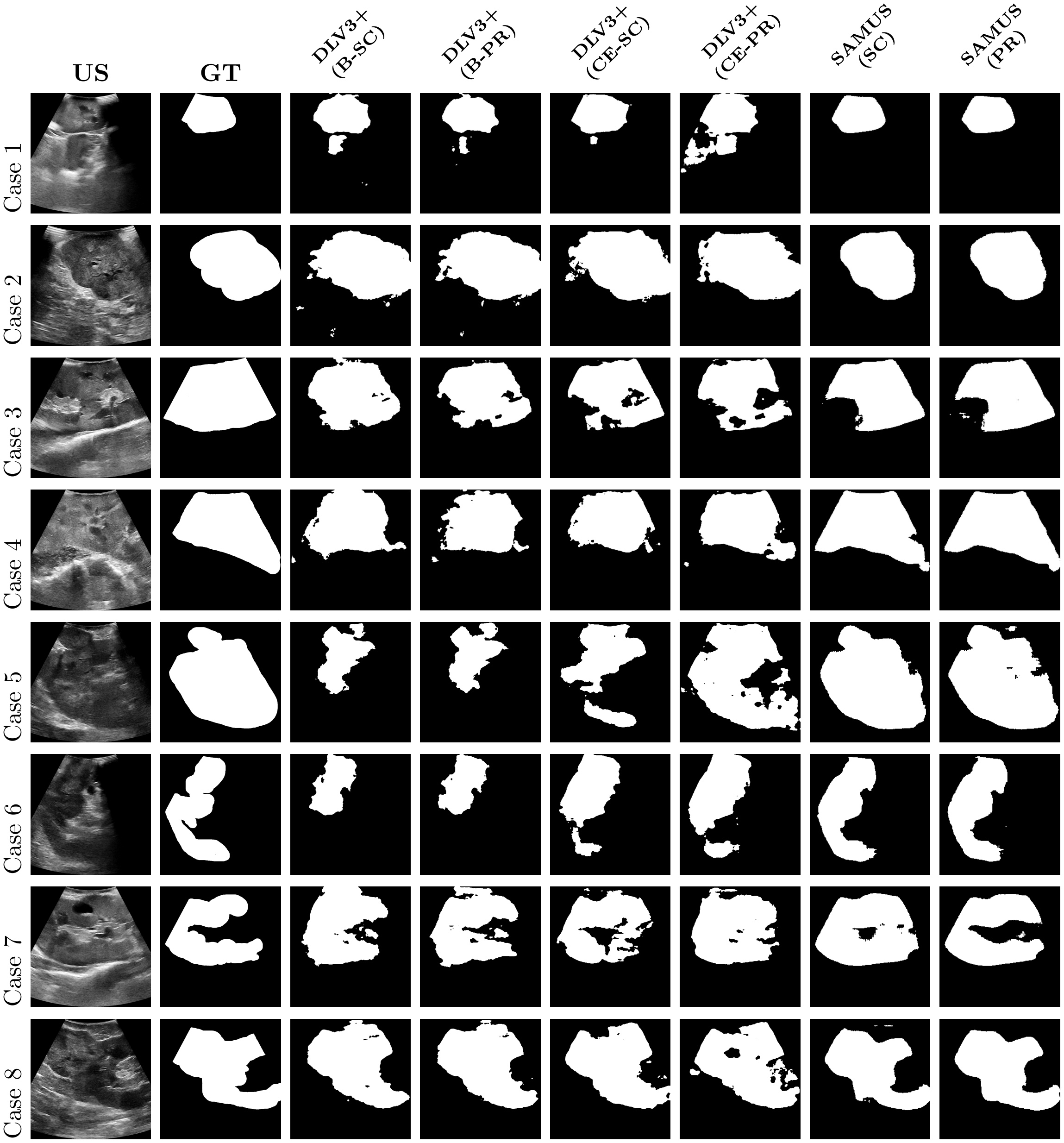


Figure 7: Qualitative results showing ultrasound images (US), corresponding ground truth (GT), and predicted segmentations from all models. Rows 1–4 resemble easier cases (clear kidney boundaries, relatively uniform shape). Rows 5–8 illustrate more complex cases, with challenging contrast and irregular shapes.

The qualitative examples, shown in Figure 7, illustrate the diversity of cases and differences between model predictions. SAMUS (SC) and SAMUS (PR) generally produced segmentations that closely matched the ground truth in both shape and extent. With fewer false-positive regions, smoother boundaries and holes compared to DLV3+ variants, even in the challenging cases with low contrast or irregular kidney shapes (case 5 - 8). The DLV3+ configurations often exhibited under-segmentation (case 5 - 6) or fragmented boundaries (case 1 - 4 & 7). However, occasional small errors remained for SAMUS, including slight under-segmentation (case 2 & 3) or over-segmentation into surrounding tissue (case 7, SAMUS-SC).

3.2.4 Case specific performance

The distributions of all reported metrics exhibited a light skew (Figure 6). To investigate the source of the large variability within one model, the distributions of the HD-95 per fold for SAMUS (SC) and DLV3+ (CE-SC) were evaluated (Appendix C). This further illustrated the variability, with median values varying ∼ 7.6 mm between folds for both models. Across all folds, SAMUS generally achieved lower median HD-95 values than DLV3+, although in two folds its variability was higher. Notably, in the fold where DLV3+ exhibited its highest median (20 mm) and widest variability, SAMUS reached its lowest median (6 mm) with relatively small variability. One fold also contained a particularly high number of outliers for both models.

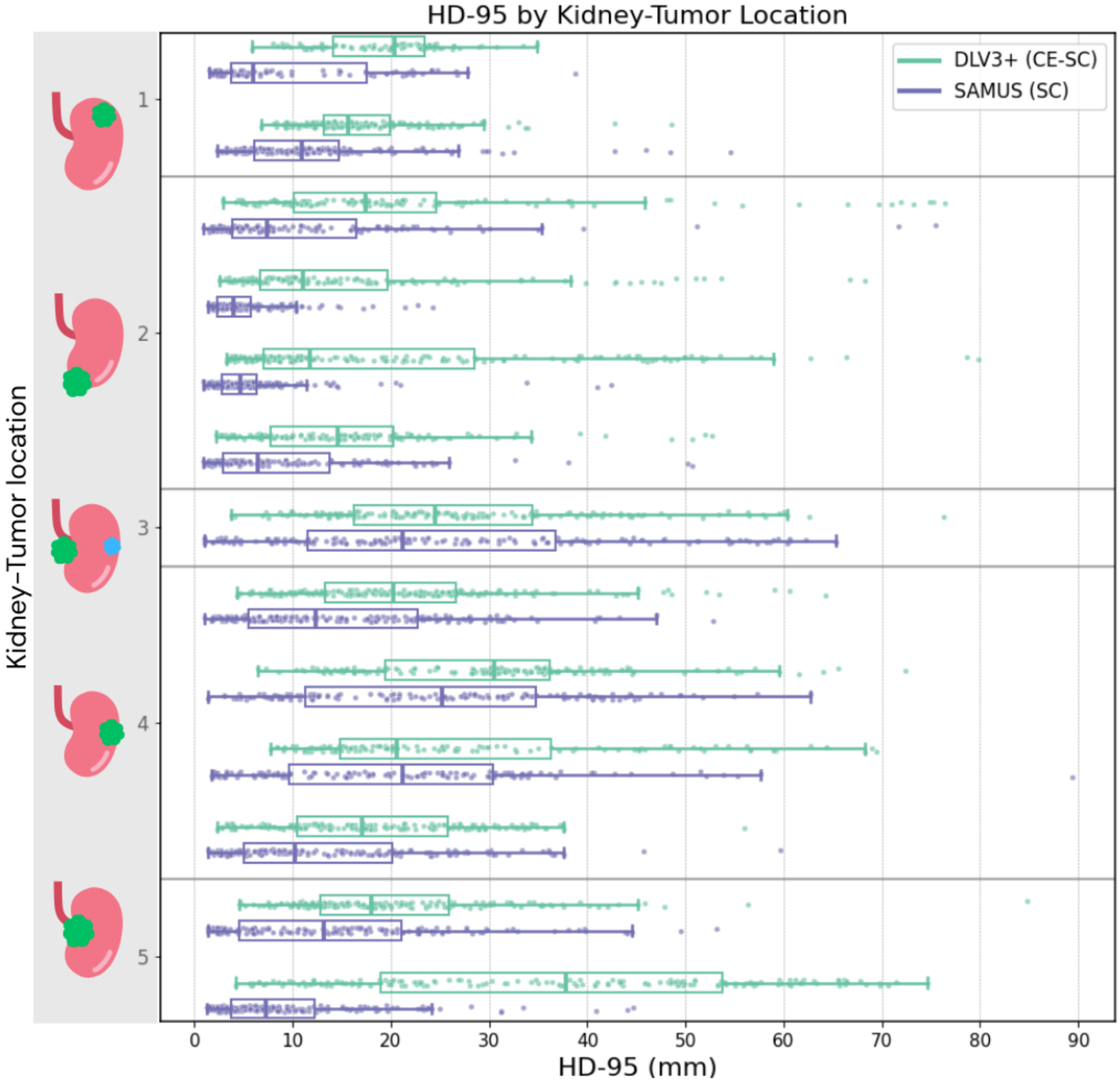


Figure 8: Boxplots for the HD-95 per identified orientation, comparing models DeepLabV3+ (DLV3+, green) and SAMUS (purple). Each boxplot resembles a case within that orientation, ordered by ascending tumor volume. The vertical line in the box represents the median value.

Case-wise analysis, categorized by kidney-tumor locations and ascending tumor size, further illustrated intra- and inter patient variability (Figure 8). The SAMUS model outperformed DLV3+ in all individual cases, with an average median value of 11.3 ± 6.6 mm, ~ 8 mm lower compared with DLV3+. Moreover, IQR and whisker ranges were consistently smaller for SAMUS. In two kidney-tumor locations (upper pole and lower pole), all cases scored below the SAMUS median. In contrast, in the case where the tumor was located on the medial border and a cyst was present on the lateral border, the median was higher (23.3 mm) and the whisker range approximately doubled. Cases where the tumor was present on the lateral border also showed larger IQRs and higher medians, with maximum values of ~ 15 mm. In these challenging kidney-tumor locations, whisker ranges extended on average ~ 30 mm higher than in the other tumor–kidney locations. Notably, in certain cases where SAMUS achieved high performance, DLV3+ did not show a similar results compared to their performance in other cases.

3.3 Posterior kidney edge performance

Since the PKE is considered for registration, performance on this specific region must be evaluated and compared to performance on the full-mask predictions (Figure 6, asterisks). Mean HD-95 and ASSD values for all models increased compared to the full masks. DLV3+ configurations exhibited an increase of 6.3 mm (CE) and 7.6 mm (B) in HD-95, while SAMUS configurations increased only by 1.8 mm. Furthermore, all standard deviations increased slightly (≈ 1.0 mm). For the ASSD the same pattern was apparent (Appendix C). However, standard deviations for the DLV3+ configurations increased by ~ 2.4 mm, while for the other models only a slight increase was apparent (< 1.0 mm).

3.4 Inference time

The inference time is an important aspect of feasibility into clinical practice. The SAMUS (SC) model obtained an average inference time of 0.23 ± 0.01 seconds on the GPU, corresponding to a 4.35 fps. This equates to processing approximately one in seven frames at 30 Hz or one in nine frames at 40 Hz. The DLV3+ (CE-SC) model was slower, it obtained an average inference time of 0.29 ± 0.03 seconds (3.45 fps), which translates to processing one in nine frames at 30 HZ and one in twelve frames at 40 Hz.

3.5 Registration performance

On average, 278.8 ± 67.8 fiducials were selected for the DL-based registration, approximately 200% more compared with manual-based registration, resulting in an average FRE of 3.4 ± 1.4 mm. Manual fiducial selection, in contrast, resulted in an FRE of 2.3 ± 0.7 mm. Finally, no trend was observed between the number of fiducials or tumor size and the FRE.

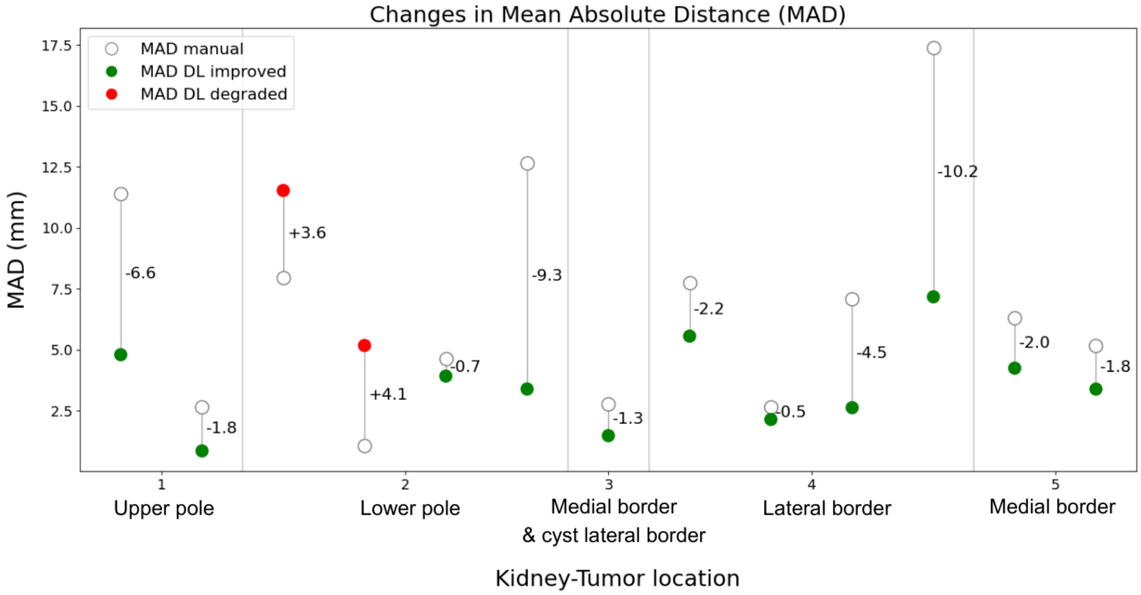


Figure 9: Changes in mean absolute distance (MAD) of the target registration error (TRE). White markers indicate results of the manual-based selection method, while green and red markers represent improvements or degradations, respectively, when using the deep learning-based selection. Vertical lines connect the paired values, with the numerical MAD difference shown to the right of each line. Cases are grouped by kidney-tumor location and ordered by ascending kidney surface fiducial count.

The number of tumor landmarks ranged from 4 to 38 between cases, and were used to obtain TRE evaluation (Figure 9). Across all cases, the average MAD_{TRE} was 4.3 ± 2.8 mm, which is an improvement of 2.6 ± 4.2 mm compared with the manual-based registration. In eleven out of thirteen cases, the MAD_{TRE} improved when using DL-selected fiducials, with an average reduction of 3.7 ± 3.5 mm compared to manual selection. Moreover, no trend was apparent between TRE improvement and kidney-tumor location, tumor size, number of landmarks or overall number of fiducials. Although the two cases where MAD did not improve both had the tumor located on the lower pole, and had the lowest number of fiducials within that location (230 and 285). Finally, the two cases in which the highest MAD improvements were seen (9.3 and 10.2 mm) were cases with the highest number of fiducials (405).

Landmark-level analysis showed that in three cases, all from different kidney-tumor locations, all landmarks improved (Appendix D). In the remaining cases where TRE was reduced (green), most landmarks improved, although the degree of improvement varied. Notably, two cases with highest number of landmarks showed more consistent improvements across landmarks. However, in the two cases with degraded TRE (red), errors increased consistently across all landmarks.

4 Discussion

4.1 Clinical Application

This study demonstrated that DL-based registration achieved clinically feasible accuracy in intra-abdominal kidney ultrasound. The TRE improved by 2.6 ± 4.2 mm compared with manually-based registration, with values comparable to the TRE reported in kidney phantom studies (~3-4 mm) [45]. The FRE for DL-based registration (3.4 ± 1.4 mm) was also consistent with literature (~2-5 mm) [44]. However, manual-based registration obtained a lower FRE (~ 1 mm), but values still overlapped with those from the DL-based selection, indicating comparable registration performance. The difference likely reflects occasional fiducial misidentification in the DL-based approach rather than a true performance advantage of manual registration. Moreover, since FRE and TRE are uncorrelated, a lower FRE does not necessarily indicate superior registration accuracy. Finally, these results were obtained under rigid transformations, making perfect alignment inherently limited as pressure applied with the US probe deformed the kidney surface.

Another key aspect of clinical feasibility is inference time. In diagnostic US or continuous guidance systems where overlays follow every probe movement, 10–20 fps is reported as ideal [28, 29]. However, intra-operative kidney registration does not require continuous segmentation of every frame. Instead, manual PKE selection is typically performed on selected key frames. The SAMUS-based model obtained a frame-rate of 4.35 fps, in this context this is sufficient for clinical application, especially compared with manual boundary selection, which typically takes several seconds per frame. Alternatively, frame skipping strategies could be used, which results in an effective 20–25 fps when every 5th frame is segmented. Together, these findings demonstrate the clinical feasibility of registration supported by DL-based segmentation.

4.2 Performance Characteristics

DL-based fiducial selection generally improved the TRE, where the largest improvements were observed in the two cases with the highest number of fiducials overall and within the kidney-tumor locations. Furthermore, in the two cases where delta TRE was degraded, the least fiducials within the kidney-tumor locations were obtained. DL-based segmentation resulted in a 200% increase of fiducials compared to the manual-based method. These findings may suggest that the number of fiducials have a positive impact on registration accuracy. This is likely because a higher number of fiducials covered a larger portion of the kidney surface, resulting in a more accurate alignment between the US and CT space. However, this increase in number of fiducials did not improve the FRE, which was on average worse than manual registration. This may be due to occasional segmentation errors and more fiducials increased the residual error. Despite this the global organ alignment improved, resulting in a lower TRE. This is consistent with literature reporting that the FRE and TRE are uncorrelated. Moreover, in the two degraded cases, TRE of all individual landmarks were degraded, while in the other cases more variability was seen. This reflects degradation was not random but rather systematic. Interestingly, segmentation quality (HD-95) did not consistently reflect registration accuracy, as degraded cases showed good HD-95 while improved cases varied. Since the TRE analysis was performed using a separate dataset, it is also possible that specific frames used for TRE evaluation were not included in the IAKUS test set. Therefore, this may not fully reflect the segmentation performance observed in cross-validation. Finally, no trends were found between registration performance and number of landmarks, differences in included fiducials or tumor size. Therefore, DL-based fiducial selection generally reduced TRE, with the largest improvements outweighing the few systematic degradations.

In terms of segmentation SAMUS obtained superior results over DLV3+ for intra-abdominal kidney US, and obtained more consistent results, especially in complex anatomical or low-quality image scenarios. This difference is likely because SAMUS benefits from larger-scale pre-training and an architecture better suited to capturing ultrasound-specific features. Although, both models obtained comparable Dice results to the literature (Figure 2. Moreover, SAMUS outperformed previously reported performances on other US datasets such as BUSI, TNK3 and CAMUS-MYO. Of which best performance was obtained on the BUSI dataset, reaching a HD-95 of 19.0 mm and Dice of 81.3%. Comparably on the IAKUS dataset a HD-95 of 13.6 ± 3.8 mm and Dice of $88.0 \pm 3.6\%$ were obtained. While results were lower compared with DLV3+ on trans-abdominal kidney ultrasound (Dice 89.8%, HD-95 9.91 mm, ASSD 3.03 mm, accuracy 98.1%), the performance on intra-abdominal data is nevertheless highly encouraging. Especially given that intra-abdominal ultrasound exhibited far greater variability in organ shape and annotation was considerably more challenging than in trans-abdominal imaging. Therefore, the comparable performance underscores the robustness and clinical potential of SAMUS in more demanding intra-operative conditions. Furthermore, the average HD-95 and ASSD values in the PKE evaluation increased for both models. This reflects the fact that the anterior kidney surface, being closest to the probe, is more frequently represented and more straightforward to segment, while the posterior edge is less consistently visualized and more difficult to annotate, leading to higher errors in the focused evaluation. The SAMUS-based configurations were less affected by this increase, indicating that their segmentations were more robust in the posterior kidney edge evaluation.

Despite the superior performance of SAMUS, the results illustrated large variability across all metrics. Moreover, median values were consistently better than average values for all metrics, indicating that the mean was pulled up by a subset of poor-performing cases. However, the SAMUS configurations obtained consistently smaller variability, with the 75th percentile HD-95 remaining below 20 mm. This reflects SAMUS obtained more consistent and accurate surface delineations with occasional large segmentation errors. Fold- and case-wise analyses highlighted both intra- and inter-patient variability. In particular, locations with the tumor located on either poles of the kidney were segmented more accurately. Possibly because cases where tumor was present on the medial or lateral border contained vessels or ureter on the PKE, which were harder to delineate and such views may have been less represented in training data. The inter-patient variability suggests specific views within one patient were more challenging, such as the non-uniform shapes seen in Figure 7. Notably, in cases where SAMUS obtained its best results the DLV3+ obtained remarkably poor performance. This indicates that one model may struggle with specific cases or orientations the other model excels, reflecting model-specific challenges. Likely arising due to differences in architecture or specific cases where US features were especially important.

The trained SAMUS configurations achieved substantial improvement from the zero-shot test, indicating that for this model kidney-specific US images were beneficial to the refinement. Although, even before training on kidney-specific ultrasound data, SAMUS obtained comparable results with lower variation to the DLV3+ after training. This indicates that SAMUS was able to generalize well, suggesting US specific features may be more important than organ specific features for training, and data diversity may be more important than strict anatomical similarity. This needs to be further tested by also training the DLV3+ on the publicly US datasets for comparison, as differences in performance may also be derived from differences in architecture. Finally, no substantial differences were apparent between SC and PR trained configurations. Indicating that pre-training on OKUS did not lead to measurable improvements over direct training on IAKUS. Possibly due to the size of the OKUS dataset or resemblance to the IAKUS dataset. Taken together, these findings imply that the limitation commonly identified in literature, the lack of publicly available kidney US data, may not be as limiting as thought when other US data is available for transfer learning.

In conclusion, these findings show that while variability across cases remained a challenge, SAMUS achieved more accurate delineation and robust performance over DLV3+. Its strong zero-shot results further suggest that ultrasound-specific features and dataset diversity may be more influential than organ-specific pre-training, as also reflected in the lack of added benefit from OKUS pre-training. Moreover, case-wise HD-95 evaluation suggests differences in performance between kidney-tumor locations while this same trend is not reflected in the TRE evaluation.

4.3 Limitations and Future Directions

Despite the improvements achieved with DL-based segmentation, registration may benefit from non-rigid transformations that better account for probe-induced kidney surface deformation or perform a probe-pressure deformation correction. Moreover, the SAMUS model may be finetuned to find the optimal hyperparameters, which could increase performance in the challenging cases. Additionally, implementing boundary-specific loss functions could further increase performance. Alternatively, other non-medical segmentation architectures may be valuable, as SAM was also not originally designed for medical segmentation, but proved effective on various US datasets and on the IAKUS data after training on $\sim$ 30k images. Alternatively, a patient specific approach may be adopted, where models are finetuned on data of specific tumor-kidney locations.

In addition to other US specific artifacts like low contrast and speckle noise, the IAKUS was faced by frames where probe was not in full contact with the lesion or kidney, and kidneys not completely being visible in the field of view. Moreover, IAKUS frames were harder to annotate compared to the OKUS set, due to varying shapes and blurry boundaries especially in PKE contained close to vessels and the ureter. Annotation quality may be improved by involving expert radiologists or additional annotators for validation.

Models may improved by increasing the dataset, more data could be obtained ex-vivo from radical nephrectomies or from phantoms. This would provide more US and organ-specific data important for adequate feature learning. Furthermore, the IAKUS dataset contained images from the entire recording, including explorative frames used by the surgeon, rather than frames used explicitly for registration. For future application, data collection should contain more registration-specific frames, and possibly more angles around the tumor location to increase kidney surface coverage. However, angles around the kidney are limited by the surgical field and layer of fat around the kidney. Finally, standardized data acquisition and landmark selection should be considered. This may be helpful to get more even distribution of frames within the dataset by being able to apply a fixed frame-rate for frame selection.

5 Conclusion

This study demonstrated that DL-based segmentation can be integrated into intra-operative kidney registration with clinically acceptable accuracy. The DL-based fiducial selection generally improved the TRE with 2.6 ± 4.2mm compared with manual-based registration. With the largest improvements outweighing the few systematic degradations. For the segmentation the SAMUS model consistently outperformed the DeepLabV3+ model across all metrics, obtaining a Dice of $88.0 \pm 2.8\%$, HD-95 of 13.7 ± 3.8 mm , ASSD of 3.5 ± 1.0 mm and accuracy exceeding 90%. Moreover, SAMUS proved to be more more robust in the posterior kidney edge evaluation further illustrating it superior performance. SAMUS obtained a inference time of 4.35 fps, which is sufficient for key-frame delineation or may be adopted when frame skipping techniques are applied.

Despite overall clinically acceptable performance of SAMUS, large variability across all metrics was apparent, reflecting consistent and accurate surface delineations with occasional large segmentation errors. Fold- and case-wise analysis further demonstrated inter- and intra-patient variability, which may be driven by large differences between tumor–kidney locations and the variety of views within a single patient. Finally, SAMUS without kidney-specific training achieved competitive results to trained DLV3+ configurations, suggesting that ultrasound-specific features and dataset diversity may be more critical than organ-specific data. The lack of measurable improvements from OKUS pretraining further implies that dataset quality, size, and diversity are more influential than strict anatomical similarity.

Overall, SAMUS provides a more accurate, efficient, and clinically relevant approach to intra-operative kidney segmentation, supporting its potential to improve registration workflows in RAPN.

References

[1] Andrea Baudo et al. "Other-Cause Mortality, According to Partial vs. Radical Nephrectomy: Age and Stage Analyses". In: *Clinical Genitourinary Cancer* 22.2 (Nov. 2023), pp. 181–188. DOI: 10.1016/j.clgc.2023.10.011.

[2] Gavin G. Calpin et al. "Comparing the outcomes of open, laparoscopic and robot-assisted partial nephrectomy: a network meta-analysis". In: *BJU International* 132.4 (June 2023), pp. 353–364. DOI: 10.1111/bju.16093.

[3] Chongyun Wang et al. "Ultrasound 3D reconstruction of malignant masses in robotic-assisted partial nephrectomy using the PAF rail system: a comparison study". In: *International Journal of Computer Assisted Radiology and Surgery* 15.7 (May 2020), pp. 1147–1155. DOI: 10.1007/s11548-020-02149-4.

[4] Rohit Singla et al. "Intra-operative ultrasound-based augmented reality guidance for laparoscopic surgery". In: *Healthcare Technology Letters* 4.5 (Aug. 2017), pp. 204–209. DOI: 10.1049/htl.2017.0063.

[5] Wangmin Liu, Enchong Zhang, and Mo Zhang. "ASO Author Reflections: Clinical Value of Navigation Systems for RAPN and LPN Procedures". In: *Annals of Surgical Oncology* 31.3 (Dec. 2023), pp. 2175–2176. DOI: 10.1245/s10434-023-14819-z.

[6] Le Li et al. "Three-dimensional (3D) reconstruction and navigation in robotic-assisted partial nephrectomy (RAPN) for renal masses in the solitary kidney: A comparative study". In: *International Journal of Medical Robotics and Computer Assisted Surgery* 18.1 (Sept. 2021). DOI: 10.1002/rcs.2337.

[7] Alberto Piana et al. "Automatic 3D Augmented-Reality Robot-Assisted partial Nephrectomy using Machine Learning: our pioneer experience". In: *Cancers* 16.5 (Mar. 2024), p. 1047. DOI: 10.3390/cancers16051047.

[8] M. A. J. Hiep et al. "Feasibility of tracked ultrasound registration for pelvic–abdominal tumor navigation: a patient study". In: *International Journal of Computer Assisted Radiology and Surgery* 18.9 (May 2023), pp. 1725–1734. DOI: 10.1007/s11548-023-02937-8.

[9] Helena R. Torres et al. "Kidney segmentation in ultrasound, magnetic resonance and computed tomography images: A systematic review". In: *Computer Methods and Programs in Biomedicine* 157 (Jan. 2018), pp. 49–67. DOI: 10.1016/j.cmpb.2018.01.014.

[10] Rashid Khan et al. "Transformative Deep Neural Network Approaches in Kidney Ultrasound Segmentation: Empirical Validation with an Annotated Dataset". In: *Interdisciplinary Sciences Computational Life Sciences* 16.2 (Feb. 2024), pp. 439–454. DOI: 10.1007/s12539-024-00620-3.

[11] Deepthy Mary Alex and D. Abraham Chandy. *Investigations on performances of pre-trained U-Net models for 2D ultrasound kidney image segmentation.* Springer Nature, Jan. 2020, pp. 185–195. DOI: 10.1007/978-3-030-60036-5_13.

[12] Radu Alexa et al. "Harnessing Artificial Intelligence for Enhanced Renal Analysis: Automated Detection of Hydronephrosis and Precise Kidney Segmentation". In: *European Urology Open Science* 62 (Feb. 2024), pp. 19–25. DOI: 10.1016/j.euros.2024.01.017.

[13] Shuaizi Guo et al. “Cross-modal Transfer Learning Based on an Improved CycleGAN Model for Accurate Kidney Segmentation in Ultrasound Images”. In: *Ultrasound in Medicine & Biology* 50.11 (Aug. 2024), pp. 1638–1645. DOI: 10.1016/j.ultrasmedbio.2024.06.009.

[14] Sang Hoon Song et al. “Deep-learning segmentation of ultrasound images for automated calculation of the hydronephrosis area to renal parenchyma ratio”. In: *Investigative and Clinical Urology* 63.4 (Jan. 2022), p. 455. DOI: 10.4111/icu.20220085.

[15] Gongping Chen et al. “MBANet: Multi-branch aware network for kidney ultrasound images segmentation”. In: *Computers in Biology and Medicine* 141 (Feb. 2022), p. 105140. DOI: 10.1016/j.compbiomed.2021.105140.

[16] Gongping Chen et al. “A novel convolutional neural network for kidney ultrasound images segmentation”. In: *Computer Methods and Programs in Biomedicine* 218 (Feb. 2022), p. 106712. DOI: 10.1016/j.cmpb.2022.106712.

[17] Deepthy Mary Alex et al. “YSegNet: a novel deep learning network for kidney segmentation in 2D ultrasound images”. In: *Neural Computing and Applications* 34.24 (Aug. 2022), pp. 22405–22416. DOI: 10.1007/s00521-022-07624-4.

[18] Rashid Khan et al. “MLAU-Net: Deep supervised attention and hybrid loss strategies for enhanced segmentation of low-resolution kidney ultrasound”. In: *Digital Health* 10 (Jan. 2024). DOI: 10.1177/20552076241291306.

[19] Pengceng Wen et al. “A-PSPNet: A novel segmentation method of renal ultrasound image”. In: *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (Oct. 2021), pp. 40–45. DOI: 10.1109/smc52423.2021.9658740.

[20] Zhengxuan Song et al. “A Two-Stage Framework for Kidney Segmentation in Ultrasound Images”. In: *International Conference on Neural Computing for Advanced Applications (NCAA 2023)*. Ed. by H. Zhang et al. Vol. 1870. Communications in Computer and Information Science. Singapore: Springer, 2023, pp. 60–74. DOI: 10.1007/978-981-99-5847-4_5.

[21] Dong-Wook Kim et al. “Advanced Kidney Volume Measurement Method Using Ultrasonography with Artificial Intelligence-Based Hybrid Learning in Children”. In: *Sensors* 21.20 (Oct. 2021), p. 6846. DOI: 10.3390/s21206846.

[22] Tao Peng et al. “Novel Solution for Using Neural Networks for Kidney Boundary Extraction in 2D Ultrasound Data”. In: *Biomolecules* 13.10 (Oct. 2023), p. 1548. DOI: 10.3390/biom13101548.

[23] Tao Peng et al. “Coarse-to-fine approach: Automatic delineation of kidney ultrasound data”. In: *Big Data Mining and Analytics* (Jan. 2024), pp. 1–12. DOI: 10.26599/bdma.2024.9020008.

[24] Gongping Chen et al. “Rethinking the unpretentious U-net for medical ultrasound image segmentation”. In: *Pattern Recognition* 142 (May 2023), p. 109728. DOI: 10.1016/j.patcog.2023.109728.

[25] Jifan Chen et al. “Auto-Segmentation Ultrasound-Based Radiomics Technology to Stratify Patient With Diabetic Kidney Disease: A Multi-Center Retrospective Study”. In: *Frontiers in Oncology* 12 (July 2022). DOI: 10.3389/fonc.2022.876967.

[26] Jaidip M. Jagtap et al. “Automated measurement of total kidney volume from 3D ultrasound images of patients affected by polycystic kidney disease and comparison to MR measurements”. In: *Abdominal Radiology* 47.7 (Apr. 2022), pp. 2408–2419. DOI: 10.1007/s00261-022-03521-5.

[27] Rohit Singla et al. "The Open Kidney Ultrasound data set". In: *arXiv (Cornell University)* (Jan. 2022). DOI: 10.48550/arxiv.2206.06657.

[28] Mario Muñoz et al. "Deep Learning-Based Algorithms for Real-Time Lung Ultrasound Assisted Diagnosis". In: *Applied Sciences* 14.24 (Dec. 2024), p. 11930. DOI: 10.3390/app142411930.

[29] Qinghua Huang and Zhaozheng Zeng. "A Review on Real-Time 3D Ultrasound Imaging Technology". In: *BioMed Research International* 2017 (Jan. 2017), pp. 1–20. DOI: 10.1155/2017/6027029.

[30] Giacomo Di Cosmo et al. "Intraoperative ultrasound in robot-assisted partial nephrectomy: State of the art". In: *Archivio Italiano di Urologia e Andrologia* 90.3 (Sept. 2018), pp. 195–198. DOI: 10.4081/aiua.2018.3.195.

[31] Liang-Chieh Chen et al. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Vol. 11211. Lecture Notes in Computer Science. Springer, 2018, pp. 833–851. DOI: 10.1007/978-3-030-01234-2_49.

[32] Meta AI. *Segment Anything*. https://github.com/facebookresearch/segment-anything. Accessed: 2025-06-04. 2023.

[33] Xian Lin et al. " Beyond Adapting SAM: Towards End-to-End Ultrasound Image Segmentation via Auto Prompting ". In: *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Vol. LNCS 15008. Springer Nature Switzerland, Oct. 2024.

[34] Intuitive Surgical. *da Vinci Xi Surgical System*. https://www.intuitive.com/en-us/products-and-services/da-vinci/systems/xi. Sunnyvale, CA, USA. Accessed: 2025-06-12. 2025.

[35] Andriy Fedorov et al. "3D Slicer as an Image Computing Platform for the Quantitative Imaging Network". In: *Magnetic Resonance Imaging* 30.9 (2012), pp. 1323–1341. DOI: 10.1016/j.mri.2012.05.001.

[36] Python Software Foundation. *Python Language Reference, version 3.10.16*. https://www.python.org/. Accessed: 2025-06-04. 2021.

[37] JetBrains. *PyCharm: Python IDE for Professional Developers 2024.3.3*. https://www.jetbrains.com/pycharm/. Accessed: 2025-06-04. 2024.

[38] Viktor Zhou. *DeepLabV3Plus-Pytorch*. https://github.com/VainF/DeepLabV3Plus-Pytorch. Accessed: 2025-06-03. 2020.

[39] Hoel Kervadec et al. "Boundary loss for highly unbalanced segmentation". In: *Medical Image Analysis* 67 (Oct. 2020), p. 101851. DOI: 10.1016/j.media.2020.101851.

[40] Adrian Celaya, Beatrice Riviere, and David Fuentes. *A Generalized Surface Loss for Reducing the Hausdorff Distance in Medical Imaging Segmentation*. Feb. 2023.

[41] Nikhila Ravi et al. "SAM 2: Segment Anything in Images and Videos". In: *arXiv (Cornell University)* (Aug. 2024). DOI: 10.48550/arxiv.2408.00714.

[42] Elvis C.S. Chen, Andras Lasso, and Gabor Fichtinger. *External tracking devices and tracked tool calibration*. Oct. 2019, pp. 777–794. DOI: 10.1016/b978-0-12-816176-0.00036-3.

[43] J.M. Fitzpatrick, J.B. West, and C.R. Maurer. “Predicting error in rigid-body point-based registration”. In: *IEEE Transactions on Medical Imaging* 17.5 (Jan. 1998), pp. 694–702. DOI: 10.1109/42.736021.

[44] Xiaoyao Fan et al. “Intraoperative fiducial-less patient registration using volumetric 3D ultrasound: a prospective series of 32 neurosurgical cases”. In: *Journal of neurosurgery* 123.3 (July 2015), pp. 721–731. DOI: 10.3171/2014.12.jns141321.

[45] James M. Ferguson et al. “Toward Practical and Accurate Touch-Based Image Guidance for Robotic Partial Nephrectomy”. In: *IEEE Transactions on Medical Robotics and Bionics* 2.2 (May 2020), pp. 196–205. DOI: 10.1109/tmrb.2020.2989661.

[46] Reuben R. Shamir and Leo Joskowicz. “Geometrical analysis of registration errors in point-based rigid-body registration using invariants”. In: *Medical Image Analysis* 15.1 (Aug. 2010), pp. 85–95. DOI: 10.1016/j.media.2010.07.010.

[47] J. Michael Fitzpatrick. “Fiducial registration error and target registration error are uncorrelated”. In: *Proceedings of SPIE, the International Society for Optical Engineering/Proceedings of SPIE* (Feb. 2009). DOI: 10.1117/12.813601.

Appendix A. Soft- and hardware overview

This appendix provides an overview of software libraries and hardware resources used during model development, training and evaluation.

Table 1: Software overview with the most important modules and packages mentioned.

3D Slicer (v5.8.1)	**Python (v3.10.16)**
Modules:	Packages:
Kidney Segmentation	PIL
Single Slice Segmentation	Pandas
Segmentations	PyTorch
Segmentation Editor	Scikit-learn
Python Console	Matplotlib
Fiducial to model registration	NumPy
Fiducial to model distance	Torchvision.models
	tqdm
	Seaborn
	MedPy

Table 2: GPU overview with specifications mentioned.

In-house GPU (NVIDIA)	**Snellius GPU (NVIDIA A100)**
Specifications:	Specifications:
2× Intel Xeon E5-2630v4 2.20 GHz CPU	Node type: gcn
8× 16GB (128GB) DDR4 ECC REG RAM	72 cores per node
2× 8TB 7200RPM HDD Seagate IronWolf	480 GiB per node
1× 960GB SSD Intel S4600	Smallest allocation 1/4 node
4× GTX 1080 Ti	Max wall time 120 h
1× Supermicro CSE-M14TQC Mobile Rack	Intel Xeon CPUs
10× KAB SATA III 1.0m	40 GB VRAM
1× Supermicro Rear Fan Kit	
Windows 10 64-bit	
1× LG DVD-RW	

Appendix B. DeepLabV3+ finetuning

The DeepLabV3+ model was finetuned to find usable hyperparameters and to evaluate to possibility of implementing boundary specific loss functions. To minimize overfitting an early stop was implemented where both the Dice had to keep improving and the validation loss had to keep decreasing in order for the model to continue training, with a patience of two. Additionally, the model's backbone (ResNet50) was frozen for the first two epochs and the first three epochs were used as a warm-up with cross-entropy as loss function, this was needed to limit overfitting and improve training stability. All loss types and combination between them were evaluated during finetuning. The other parameters were set to: fold 0, training batch size 8, validation batch size 4, crop size 254, maximum augmentation, learning rate 0.003.

Table 3: Comparison of loss functions. Cross-entropy (CE), Hausdorff (HD), and Generalized Surface Loss (GSL). Top-3 values in Dice, Accuracy, HD95, and ASSD are shown in bold.

Loss Type	**Dice (%) ↑**	**Accuracy (%) ↑**	**HD95 (mm) ↓**	**ASSD (mm) ↓**
CE	**76.8**	**87.5**	22.5	**7.1**
Focal	72.2	86.0	34.1	11.4
Dice	76.0	**87.2**	24.1	8.4
Boundary	73.9	86.8	**19.9**	**6.0**
HD	67.5	84.6	28.7	9.6
GSL	71.0	85.5	28.6	10.5
Dice + HD	76.6	**87.2**	34.1	9.2
Dice + B	**77.3**	87.0	42.7	13.0
CE + HD	73.3	86.6	21.8	7.2
CE + B	76.4	87.1	44.4	14.2
CE + GSL	**76.7**	87.1	42.0	12.9
GSL + HD	69.5	84.9	**21.5**	8.3
GSL + B	74.4	86.6	**18.1**	**6.2**

There was a clear definition between losses that perform well in terms of region-based metrics and ones that perform better in terms of contour-based metrics (Table 3). The Cross-entropy (CE) loss, Dice Loss and Dice loss in combination with boundary (B) loss were the top three loss functions in terms of surface metrics, although results are competitive. Boundary loss, Generalized surface loss (GSL) in combination with Boundary loss and GSL in combination with Hausdorff loss were the top three loss functions in terms of boundary metrics. Moreover, large differences were apparent between models.

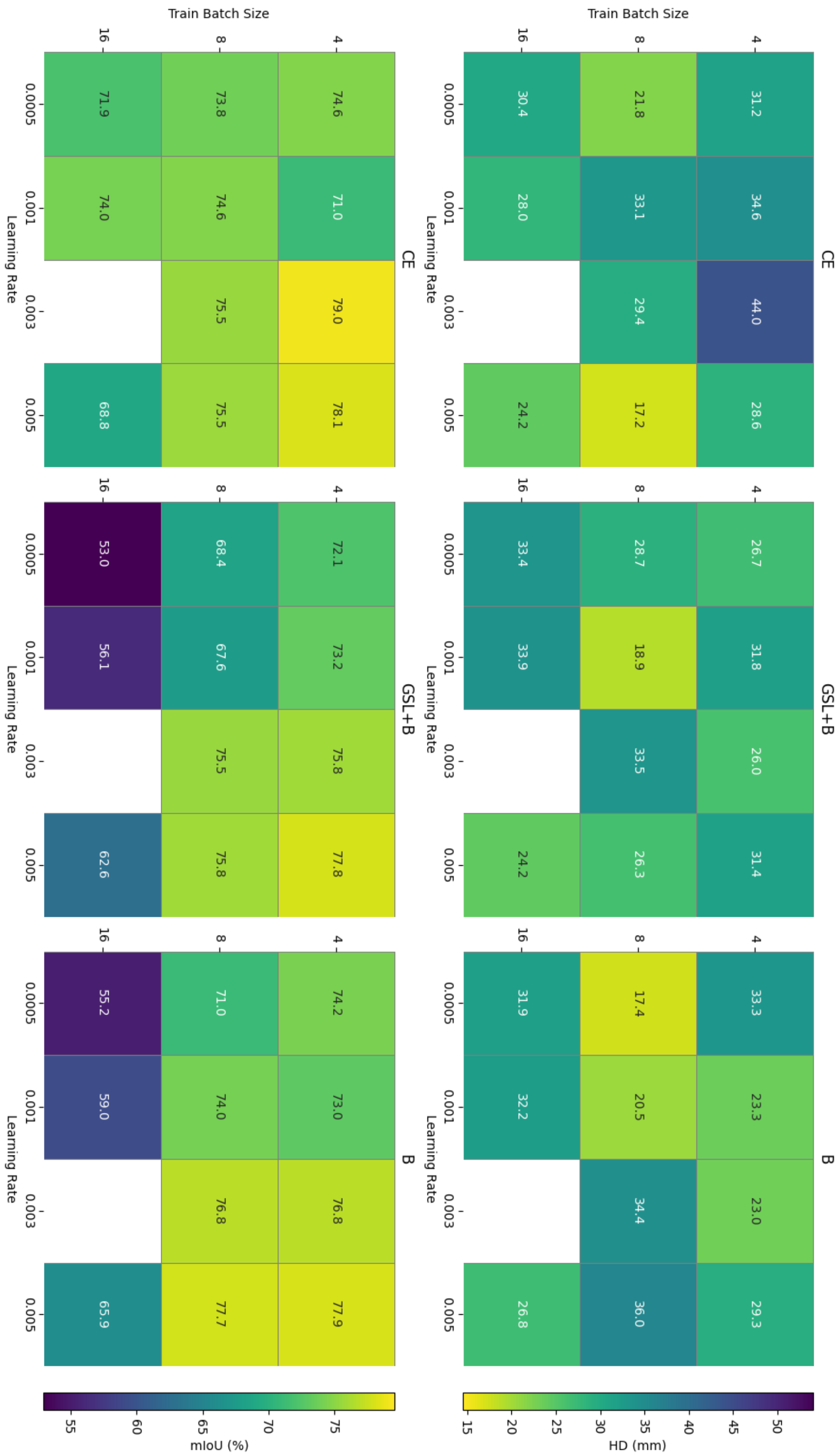


Figure 10: Heatmaps of hyperparameters: learning rate, training batch size, and learning batch size (set to half the training batch size) for CE loss, B loss, and GSL + B loss. Configurations that failed due to fragmentation errors are shown as empty cells. Best configurations are shown in yellow, resembling low values for HD-95 and high values for the mean intersection over union (mIoU).

Boundary loss and GSL + B loss were selected to be further finetuned, due to their superior performance in the contour-based metrics. Additionally, CE loss was selected due to its superior performance in the region-based metrics, and comparable performance to best contour loss types. For the finetuning the following search space for the grid-search was selected: learning rates $0.0005, 0.001, 0.003, 0.005]$, loss types[CE, Boundary, GSL + Boundary] and training batch size $[4, 8, 16]$ (where the validation batch size was $1/2$ of the training batch size). The grid-searches are visualized in heatmaps, for both metrics mean intersection over union (mIoU) and Haussdorff distance (HD) in Figure 10. The best state in terms of mIoU was obtained by the CE loss, where the best value was 79.5%. And the best states in terms of the HD were also obtained by the CE loss, where the best obtained value was 7.31 mm. Although this state did not obtain the best mIoU value (74.9%) it was competitive with other states and there were no large variations in the mIoU values. Moreover, both CE loss and B loss reached six epochs of training before overfitting while GSL + B was stopped after four epochs by the early stopping mechanism. Therefore, cross-entropy and boundary loss were implemented as loss functions for the DeepLabV3+ model, and the parameters were set to a training batch size of 8, a validation batch size of 4 and a learning rate of 0.005 and 0.001, respectively.

Appendix C. Supplementary figures: Full mask results

This appendix contains supplementary figures to the results section, full mask performance. In Figure 11 the distributions of the ASSD and accuracy are visualized in boxplots for every model. In Figure 12 the HD-95 distributions per fold for DeepLabV3+ and SAMUS both trained from scratch is shown.

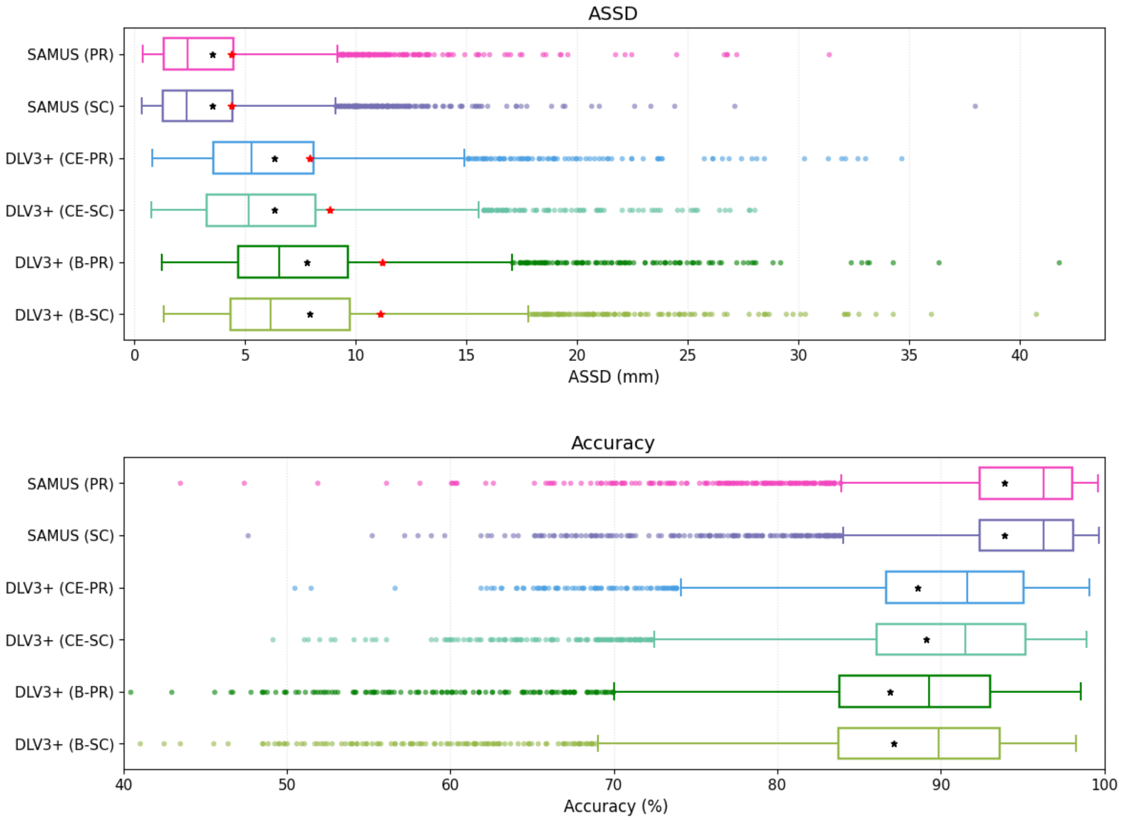


Figure 11: Performance on pooled image level across models. The black asterisk (*) marks the mean across folds, and the red asterisk marks the mean of the posterior kidney edge evaluation. The pooled median is marked with a vertical line within the IQR. Where SAMUS (PR) is resembled in pink, SAMUS (SC) in purple, DLV3+ (CE-PR) in blue, DLV3+ (CE-SC) in light blue, DLV3+ (B-PR) in dark green and DLV3+ (B-SC) in light green.

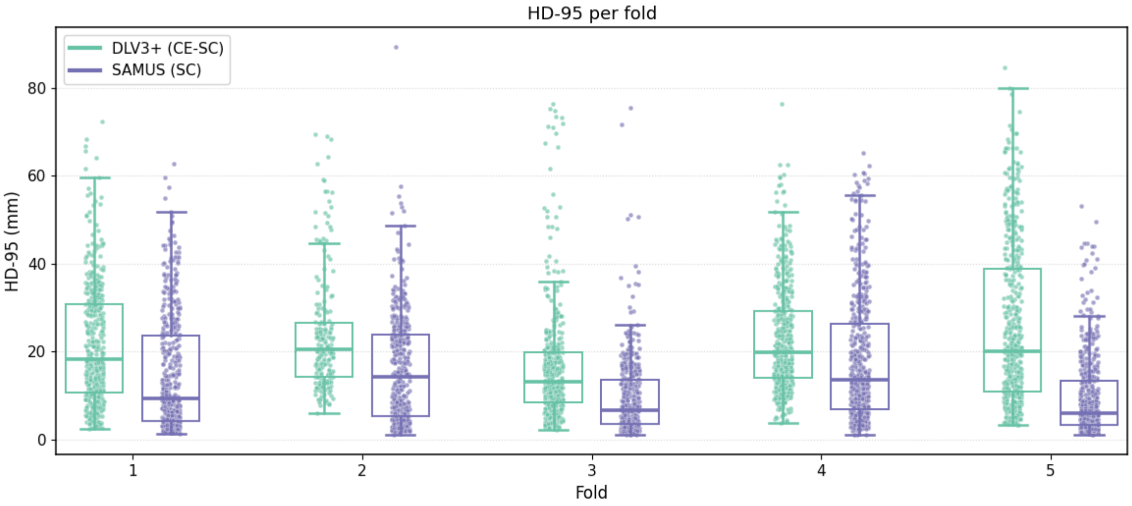


Figure 12: Performance on fold-level for the DLV3+ (CE-SC) (turquoise) and SAMUS (SC) (purple) models.

Appendix D. Supplementary figure: Registration performance

This appendix contains the supplementary figure for the registration performance. Figure 13 resembles the difference in TRE for reach landmark between the manual- and DL-based registration.

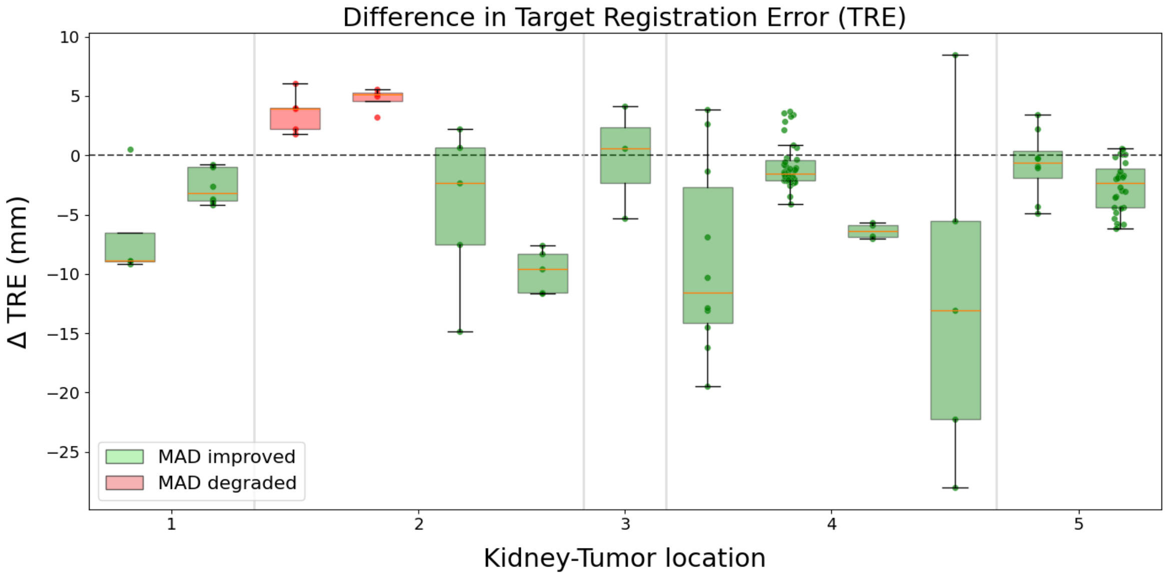


Figure 13: Differences in target registration error (TRE) per landmark beteen manual- and deep learning-based registration, each boxplot resembles an individual case and each marker represents a landmark. Green boxplots resemble cases where MAD was improved and red boxplots resemble cases where MAD was regraded, in both the median is represented in orange. Cases are grouped by kidney-tumor location and ordered by ascending fiducial count.