Cascaded Localization Regression Neural Nets for Kidney Localization and Segmentation-free Volume Estimation

Mohammad Arafat Hussain, Ghassan Hamarneh, *Senior Member, IEEE,* and Rafeef Garbi, *Senior Member, IEEE*

***Abstract*—Kidney volume is an essential biomarker for a number of kidney disease diagnoses, for example, chronic kidney disease. Existing total kidney volume estimation methods often rely on an intermediate kidney segmentation step. On the other hand, automatic kidney localization in volumetric medical images is a critical step that often precedes subsequent data processing and analysis. Most current approaches perform kidney localization via an intermediate classification or regression step. This paper proposes an integrated deep learning approach for (i) kidney localization in computed tomography scans and (ii) segmentation-free renal volume estimation. Our localization method uses a *selection*-convolutional neural network that approximates the kidney inferior-superior span along the axial direction. Cross-sectional (2D) slices from the estimated span are subsequently used in a combined sagittal-axial Mask-RCNN that detects the organ bounding boxes on the axial and sagittal slices, the combination of which produces a final 3D organ bounding box. Furthermore, we use a fully convolutional network to estimate the kidney volume that skips the segmentation procedure. We also present a mathematical expression to approximate the 'volume error' metric from the 'Sørensen–Dice coefficient.' We accessed 100 patients' CT scans from the Vancouver General Hospital records and obtained 210 patients' CT scans from the 2019 Kidney Tumor Segmentation Challenge database to validate our method. Our method produces a kidney boundary wall localization error of $\sim$2.4mm and a mean volume estimation error of $\sim$5%.**

***Index Terms*—Mask-RCNN, FCN, CNN, kidney localization, kidney volume, Sørensen–Dice.**

I. Introduction

CLINICALLY, the reduced or absent functionality of a kidney for more than three months, is referred to as chronic kidney disease (CKD). Kidney disease, often resulting from the tumor, occurs in both hereditary and sporadic forms [1]. It is a significant risk factor for death worldwide [2]. The prevalence of CKD varies between 7-12% in different regions of the world. For example, China, Canada, Australia, the United States, Germany, Finland, Spain, and England reported 1.7%, 3.1%, 5.8%, 6.7%, 2.3%, 2.4%, 4.0%, and 5.2%, respectively [3]. Different CKDs, for example, Autosomal dominant polycystic kidney disease (ADPKD) and renal artery atherosclerosis (RAS), often lead to the end-stage-renal-disease (ESRD), which are associated with the change of kidney volume. However, CKD detection is complicated. Usual laboratory tests such as the estimated glomerular filtration rate (eGFR) and serum albumin-to-creatinine ratio often cannot detect early disease and known to be unreliable in tracking disease progression [4]. Several works suggested kidney volume as a potential surrogate marker for renal function. Thus, kidney volume is considered useful for predicting and tracking the progression of different CKDs [5], [6]. The *total kidney volume* is now considered as the gold standard imaging biomarker for ADPKD and RAS progression at the early stages of this disease [6]. In addition, renal volume measurement is an emerging alternative to renal scintigraphy, which is used in evaluating split renal function in kidney donors [5]. It is also considered the best biomarker in the follow-up evaluation of kidney transplants [5].

Kidney volume from 3D computed tomography (CT) data is typically estimated using different segmentation methods. These segmentation methods can be broadly categorized into two groups based on their use of any prior kidney localization step. Some methods use manual/(semi)automatic kidney localization before segmentation, while other methods directly perform segmentation without using a previous localization step. Although both types of methods have been reported in the literature, often, methods of the first category are preferred in the clinical environment because kidney localization facilitates better segmentation/volume estimation and can improve and speed up other algorithms as kidney lesion detection and registration [7].

Mohammad Arafat Hussain and Rafeef Garbi are with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada. E-mail: {arafat, rafeef}@ece.ubc.ca.

Ghassan Hamarneh is with the Department of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada. E-mail: hamarneh@sfu.ca.

II. Related Work

A. Kidney Localization

For the last two decades, medical imaging scientists have proposed several kidney localization approaches within the 3D volumetric medical images. This section surveyed the most relevant and recent machine learning-based kidney localization approaches in CT, divided into hand-engineered feature-based classical machine learning (ML) approaches and deep learning (DL) approaches.

1) Classical ML Approaches: Criminisi et al. [8], [9] predicted the locations of organ bounding box walls using regression-forest (RF)-based approaches and achieved a mean kidney bounding box wall localization error of $\sim$13mm. Cuingnet et al. [10] used an additional RF to fine-tune the method by Criminisi et al. [9] that improved the kidney localization accuracy by $\sim$60% (i.e., mean kidney localization error of 7mm). Gauriau et al. [11] estimated an organ bounding

box from the cascaded RF-based organ confidence map and achieved a mean wall localization error of ∼5.5mm for kidney bounding boxes. Recently, Samarakoon et al. [12] proposed a light RF that uses fewer nodes than regular RF to localize different organs in the CT scan and achieved a mean kidney localization error of ∼11mm. Zhou et al. [13] used ensemble learning-based multiple 2D detectors and combined their outputs using collaborative majority voting in 3D to accomplish robust kidney localization. In their subsequent work [14], [15], they localized the kidney in CT images using template matching, hand-crafted features, and local binary patterns. In summary, these classical ML approaches estimate the location of the kidney or its bounding box via predicting a continuous regression value or voxel-based voting.

2) DL Approaches: Humpire et al. [7], [16] proposed a convolutional neural network (CNN)-based approach to detect six organs, including the kidneys. They trained three separate CNNs to classify images taken from three orthogonal directions, where the classification of a slice is performed based on the presence or absence of a particular organ cross-section in that slice. The 3D organ bounding box is then generated by combining the classified labels of orthogonal images, which achieved a localization error of ∼2.6mm for kidneys. Lu et al. [17] proposed a method using a cross-sectional fusion of CNN and fully convolutional networks (FCN) for right kidney localization. In our previous work [18], we proposed a kidney localization approach that used a single CNN where 2D slice classifications from the three orthogonal views were concatenated into a fully connected layer to provide a voxel-wise kidney location prediction. In summary, these DL approaches determine the kidney locations via classifying the 2D CT slices, depending on the presence or absence of the kidney cross-section in that. In contrast, recently, Xu et al. [19] proposed a 3D region proposal network for localizing eleven abdominal organs, including the kidneys. Unlike classification CNN, region proposal networks usually propose a 2D/3D region around an object of interest, based on the learned feature maps. They achieved a localization error of ∼4mm for kidneys.

B. Kidney Volume Estimation

Kidney volume is typically estimated using different segmentation methods. Similar to the kidney localization approach, here we survey classical ML and DL methods for kidney segmentation from CT, though we note that other approaches exist for kidney segmentation from magnetic resonance (MR) images (e.g., [20]). Furthermore, we discuss the segmentation-free volume estimation approaches in the literature.

1) Classical ML Approaches: Zhou et al. used content-based image retrieval, group-wise organ location calibration, and 3D GrabCut techniques for kidney localization in [13], [14], [15], respectively. Cuingnet et al. [10] used a combination of RF and template deformation to segment kidneys. Glocker et al. [21] used a joint classification-RF scheme to segment different abdominal organs, including kidneys. Khalifa et al. [22] developed a 3D kidney segmentation framework integrating CT appearance features, higher-order appearance models, and adaptive shape model features into a random forest classification model. Hristova et al. [23] used vantage point trees to classify voxels for kidney segmentation. Zhao et al. [24] used CT intensity features from the image in an RF framework for voxel-level classification to segment kidneys.

2) DL Approaches: Chen et al. [25] proposed a 3D FCN based method for automatic multiorgan segmentation in dual-energy CT. Gibson et al. [26] proposed dense V-network FCN for multiorgan segmentation from abdominal CT images. Valindria et al. [27] investigated learning from multiple modalities for organ segmentation and showed effectiveness on kidney segmentation. Thong et al. [28] showed promising kidney segmentation performance using CNN. Keshwani et al. [29] proposed a multitask 3D CNN to segment ADPKD. Sharma et al. [30] performed the automated segmentation of ADPKD kidneys using FCN. Groza et al. [31] compared several CNN-based approaches for kidney segmentation and argued that foveal FCN is the most suitable deep architecture. Recently, more than two hundred deep learning methods have been proposed for kidney segmentation in the 2019 Kidney Tumor Segmentation (KiTS) Challenge [32], where most of the methods are variants of 3D U-Net [33] or V-Net [34]. Recent work on medical image segmentation tasks included many contributions based on 2D and 3D U-shaped networks (i.e., U-Nets) [35]. For example, to improve the U-Net architecture robustness to challenging organ and tumor segmentation scenarios, attention gates [36] and squeeze-and-excitation blocks [37] have been proposed.

3) Segmentation-free Approaches: Classical ML-based segmentation methods require hand-engineering features, which is often hard to design optimally. DL-based segmentation approaches showed better kidney segmentation performance than the classical ML approaches due to their capability of learning optimal features automatically. However, DL-based segmentation approaches often require training deep and complex dense prediction networks through expensive computation. Furthermore, it is often challenging to decide on the deep architecture and appropriate loss function. Besides, kidney cancer appears very heterogeneous on CT images. It often appears hyperdense (i.e., calcifications) as well as hypodense (e.g., cystic and necrotic tissue). This scenario introduces additional challenges to deep learning-based segmentation of kidneys [39]. Many organ functionality-related parameters are estimated using segmentation in clinical settings, although the ultimate aim is not producing a segmented organ. Thus, this segmentation procedure introduces additional challenges in estimating these vital parameters, e.g., total kidney volume. Avoiding the computational overheads and limitations associated with segmentation approaches, several segmentation-free ML approaches have been proposed for cardiac biventricular volume estimation from MR images [40]–[48], and direct tumor volume estimation from PET scans [49]. Recently, we proposed two segmentation-free kidney volume estimation approaches using a dual regression forest [38], and a CNN [18], respectively, which bypassed the segmentation step altogether. To the best of our knowledge, we are the first

TABLE I
LIST OF ML-BASED KIDNEY LOCALIZATION AND VOLUME ESTIMATION APPROACHES FOR CT. SHADED ROWS REPRESENT THE METHODS THAT INTEGRATE THE KIDNEY LOCALIZATION AND VOLUME ESTIMATION PROCESSES. ACRONYMS USED AS: SEG–SEGMENTATION, CM–CLASSICAL ML, AND DL–DEEP LEARNING.

	Localization		**Volume Estimation**			
			Seg		Seg-free	
Methods	CM	DL	CM	DL	CM	DL
Criminisi et al. [8]	✓					
Criminisi et al. [9]	✓					
Cuingnet et al. [10]	✓		✓			
Gauriau et al. [11]	✓					
Samarakoon et al. [12]	✓					
Zhou et al. [13]	✓					
Zhou et al. [14]	✓					
Zhou et al. [15]	✓					
Humpire et al. [16]		✓				
Humpire et al. [7]		✓				
Lu et al. [17]		✓				
Hussain et al. [18]		✓				✓
Xu et al. [19]		✓				
Zhou et al. [13]			✓			
Zhou et al. [14]			✓			
Zhou et al. [15]			✓			
Glocker et al. [21]			✓			
Khalifa et al. [22]			✓			
Hristova et al. [23]			✓			
Zhao et al. [24]			✓			
Chen et al. [25]				✓		
Gibson et al. [26]				✓		
Valindria et al. [27]				✓		
Thong et al. [28]				✓		
Keshwani et al. [29]				✓		
Sharma et al. [30]				✓		
Groza et al. [31]				✓		
Hussain et al. [38]					✓	
Proposed		✓				✓

to use CNN [18] for segmentation-free renal volumetry.

C. Contributions in the Proposed Method

This paper proposes an integrated approach for kidney localization and segmentation-free volume estimation using CNN-guided Mask-RCNN. We summarize the state-of-the-art ML-based methods for kidney localization and volume estimation from CT in Table I. In the table, we see that the proposed method is one of the two DL-based integrated methods (shown in shaded rows), where the other is our previous work [18]. The proposed method extends our previous work and achieves improved performance in kidney localization and volume estimation. Our proposed method's first module uses an effective CNN-guided Mask-RCNN approach for efficient kidney localization in CT. The second module subsequently uses the localized kidney data in an FCN for segmentation-free kidney volume estimation. Thus, the technical novelty of the proposed approach is four-fold:

1) we propose a novel CNN-based pipeline for kidney localization (Fig. 1). Although the underlying components of this pipeline are not novel, the way they are combined is what makes our method novel and effective,
2) we propose a new way to tackle a specific task in deep learning, i.e., area prediction without segmenting the kidney cross-section,
3) the design of the input-output relationship of the data in different parts of the proposed CNN pipeline is novel, and
4) for the first time, we derive a mathematical relation between two well established metrics in this field, the 'volume error' and 'Sørensen–Dice coefficient', to facilitate an approximate comparison between these two metrics.

The paper is organized as follows. Section III describes the proposed kidney localization-volume estimation technique. Section IV describes the validation and experiment setup. Section V presents the results to demonstrate the strength of our algorithm. Concluding remarks are presented in Section VI.

III. METHODS

In recent years, the increase in modern computers' computation capabilities has led to the development of complex 3D CNN models, which are very successful in 3D volumetric medical image-based applications. However, these complex models' inference also requires computers with extensive computation capabilities, which is not a match for point-of-care computers in typical clinical settings. That is why we explore the 2D CNN approaches in this paper, which are lighter in model complexity. Simultaneously, we also show better kidney localization and volume estimation performance than the state-of-the-art 3D CNN approaches.

A. Kidney Localization

Our kidney localization approach is a three-step pipeline (Fig. 1, box II-A). In the first stage, we use a kidney span detection CNN (S-CNN) classifying 2D axial slices, enabling a rough detection of the targeted kidney span along the axial direction. In the second stage, we use a Mask-RCNN to detect the 2D kidney bounding box along the coronal and sagittal directions. In the third and final stage, we use the same Mask-RCNN to detect the 2D kidney bounding box along the axial and sagittal directions. Since RCNN typically produces false-positive kidney bounding boxes in those slices that do not contain the kidney, the CNN pipeline of our method controls the choice of slices (fed to the RCNNs) by extracting those from kidney-containing regions by using S-CNN. We sequentially provide technical details and justification for different design choices in the following sections.

1) Kidney Span Detection using S-CNN: We use the S-CNN (ResNet-50 [50]) to classify 2D axial slices that enable a rough detection of the kidney span along the axial direction (Fig. 1, box II-A1). The initial slice classification labels (i.e., 0: kidney absent, 1: kidney present) may contain a few false positives and false negatives. To remove those, we perform a moving average over the label values along the axial direction with a moving window size of 12 cm as a typical kidney length is approximately 12cm [51]. Then we divide these average values by their maximum value to normalize between [0, 1] and estimate the organ span from the range of values ≥ 0.75. We empirically found that 0.75 is a robust threshold. This approximate span could be larger than the actual kidney span. If the estimated span comes out smaller than a typical kidney

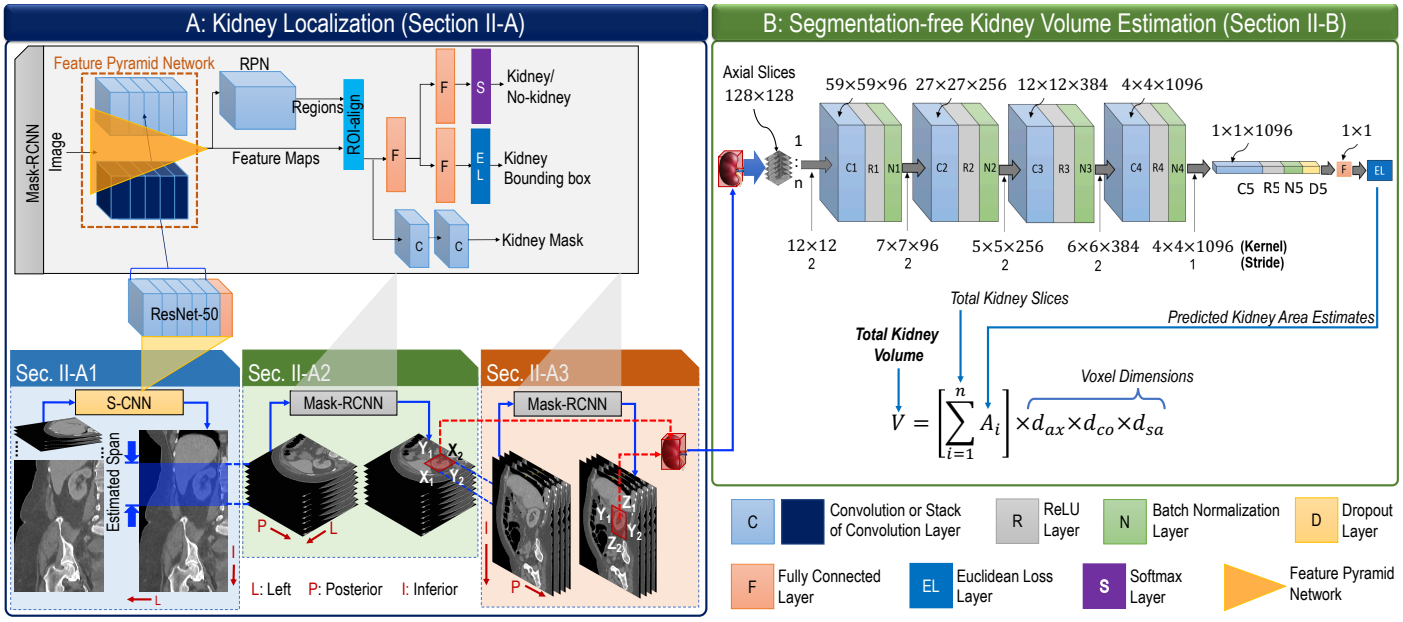

Fig. 1. Schematic diagram of our proposed method for analyzing renal CT scans: (A) The CNN-guided Mask-RCNN- and FCN-based integrated kidney localization component, and (B) the segmentation-free volume estimation component.

length (i.e., 12cm), we take extra slices into the span in the superior and inferior directions to ensure at least 12cm kidney height. This S-CNN (ResNet-50) network was pretrained on the ImageNet dataset, and we fine-tune the network weights on our kidney dataset. Although the S-CNN aims to detect the approximate kidney span in the axial direction, we fine-tune this network using cross-sectional 2D slices from all three orthogonal directions. For details of this network structure and function, we refer readers to [50].

2) Bounding Box Detection in the Coronal-Sagittal Direction: In this stage, we use a Mask-RCNN [52] to detect the 2D kidney bounding box along the coronal and sagittal directions (Fig. 1, box II-A2). The input to the Mask-RCNN is the 2D axial slices strictly taken from the inside of S-CNN's selected span.

The Mask-RCNN produces the classification, bounding box, and segmentation mask of an object in an image. It works in two stages: (1) first, it proposes regions on the input image where there is a likeliness of the presence of an object, and (2) second, the network classifies the object and produces the bounding box and segmentation mask of a particular object inside the proposed region. Mask-RCNN uses a feature pyramid network (FPN), referred to as the 'backbone.' The FPN has laterally connected bottom-up and top-down pathways. The bottom-up pathway is nothing but a CNN that extracts image features. The top-bottom pathway produces a feature pyramid of similar size to that of the bottom-up pathway. The FPN features are then used by a region proposal network (RPN) for proposing an object region. A pooling layer [region-of-interest (ROI)-align] subsequently extracts fixed-length feature vectors from the proposed regions. Each feature vector then goes through a sequence of fully connected/convolution layers and branches into three output layers: a classification layer, an object bounding box layer, and a segmentation masking layer. Mask-RCNN uses a multitask loss function $L = \lambda_1 L_{cls} + \lambda_2 L_{bbox} + \lambda_3 L_{mask}$, where λ_1, λ_2 and λ_3 are the balancing weights, and L_{cls}, L_{bbox} and L_{mask} are the class loss, bounding box loss and mask loss, respectively, defined as:

$$L_{cls} = -t_1 \log(s_1) - (1 - t_1) \log(1 - s_1), \tag{1}$$

$$L_{bbox} = q \sum_{i \in 1,2,3,4} \text{smooth}_{L_1}(b_i^t - b_i), \tag{2}$$

$$where\ \text{smooth}_{L_1}(u) = \begin{cases} 0.5u^2, & \text{if } |u| < 1 \\ |u| - 0.5, & \text{otherwise} \end{cases}$$

$$L_{mask} = q\{-\frac{1}{m^2} \sum_{1 \le i,j \le m} [-t_{i,j} \log(s_{i,j}) - (1 - t_{i,j}) \log(1 - s_{i,j})]\}, \tag{3}$$

where $t \in \{0$: background, 1: kidney$\}$ is the true label, s is the prediction score, q is equal 1 for positive anchor (i.e., kidney containing region proposal) or 0 otherwise, b is a vector representing the four parameterized coordinates of the predicted bounding box, and b^t is that of the ground-truth box associated with a positive (i.e., $q = 1$) anchor, and m^2 is the area of a mask of dimension $m \times m$ pixels for each ROI. Based on the kidney localization performance analysis on the validation data, we empirically set $\lambda_1 = \lambda_2 = \lambda_3 = 1$.

We use ResNet-50 [50] as the bottom-up network of FPN in our Mask-RCNN. This ResNet-50 is fine-tuned from S-CNN. For fine-tuning, we use a kidney containing 2D slices from all three orthogonal directions. During inference, we restrict the Mask-RCNN to produce a single bounding box and a kidney mask per slice. Although the Mask-RCNN produces a 2D bounding box around a kidney cross-section, in most cases, it does not tightly encompass a kidney cross-section.

Rather, gaps are seen between the predicted boundary line and the actual kidney boundary. Therefore, we use the predicted kidney mask to generate the rectangular kidney bounding box. Finally, we find the sagittal and coronal edges of a bounding box, which is the Union set of all axial bounding boxes, by $X_1 = min(x_1)$, $X_2 = max(x_2)$, $Y_1 = min(y_1)$ and $Y_2 = max(y_2)$, where min and max are the minimum and maximum operators, respectively, and X_1, X_2 and Y_1, Y_2 are the largest-box edges along the coronal and sagittal directions, respectively (Fig. 1, box II-A2). Note that finding the rough kidney span along the axial direction by the initial S-CNN is important for this stage, as false-positive bounding boxes may corrupt these estimates.

3) Bounding Box Detection in the Axial-Sagittal Direction: In this final detection stage, we use the same Mask-RCNN. In this stage, the input to the Mask-RCNN is the 2D sagittal slices strictly taken from $[X_1, X_2]$ (Fig. 1, box II-A3). In this stage, the Mask-RCNN detects the kidney bounding box along the sagittal and coronal directions. This stage updates the estimated axial kidney span in the previous step. Finally, we find the largest bounding box's axial edges, which is the largest of all the sagittal bounding boxes, by $Z_1 = min(z_1)$ and $Z_2 = max(z_2)$, where z_1 and z_2 are the edges along the axial directions. Lastly, we combine the final predicted spans in the second and third stages by the Mask-RCNN to produce the 3D bounding box around the kidney (Fig. 1, box II-A3).

B. Segmentation-free Kidney Volume Estimation

In Section III-A, we discussed the kidney encompassing tight ROI estimation process. Typically, there is a considerable variation in kidney size and shape across patients. It is often customary to feed images of similar size to a CNN during training. In this paper, we fix our input image patch size to 128×128 pixels, consistent across the training data. We use K_p/P_p as the output variable (label) for a particular image patch, where K_p is the total kidney pixels, and P_p is the total pixels in the image patch before getting resized to 128×128 pixels.

We use an FCN (Fig. 1, box II-B) to predict the cross-sectional area of a kidney in each patch. Our FCN is a regression network consisting of six layers, excluding the input. It has five convolutional layers, one fully connected layer (only to generate a single activation), and one Euclidean loss layer. To avoid overfitting, we use the dropout in the last convolution layer. During inference, the FCN predicts the kidney area in a particular image patch. Since all the input images were resized to 128×128 pixels, we multiply the FCN-predicted area estimate by a factor $(128 \times 128)/P_p$. Finally, we calculate a particular kidney's volume by adding the FCN-predicted areas for all of its axial image patches and multiplying by the voxel dimensions (Fig. 1, box II-B).

IV. Validation and Experiment Setup

A. Datasets

We used 100 patients' CT scans accessed from the Vancouver General Hospital (VGH) records with required ethics approvals by the UBC Clinical Research Ethics Board (CREB), certificate number: H15-00237. There were a total of 200 kidney samples, and we used 130 samples (from 65 randomly chosen patients) for training, 20 samples (from 10 randomly chosen patients) for validation, and the remaining 50 samples for testing. Our dataset included 12 pathological kidney samples (with endo- and exophytic tumors), and our training and test data contained six pathological cases each. We made sure that kidneys from the same patient were not split across the training, validation, and test cases. These data were acquired using a Siemens SOMATOM Definition Flash (Siemens Healthcare GmbH, Erlangen, Germany) CT scanner. Ground truth kidney bounding boxes were calculated from manual kidney delineation performed by an expert radiologist. We show a summary of these data in Table II.

TABLE II
SUMMARY OF RELEVANT AND AVAILABLE INFORMATION OF THE CT DATA FROM VGH.

Items	Descriptions
Pixel Dimensions	Axial: 1.5∼3mm Coronal: 0.5820∼0.9766mm Sagittal: 0.5820∼0.9766mm
Contrast Agent Used	45 cases
Total Patients	100
Number of Males	50
Number of Females	50
Age	Mean: 56.71±15.81 Y Minimum Age: 19 Y Maximum Age: 89 Y
Number of Pathological Kidneys	12 kidneys in 12 patients

We also used 210 patients' CT scans from the 2019 Kidney Tumor Segmentation (KiTS) Challenge database [32]. This database contains patients' scans accessed from the University of Minnesota Medical Center records. These patients underwent partial or radical nephrectomy for one or more kidney tumors between 2010 and 2018. We used 160 randomly chosen patients' data for training, 15 randomly chosen patients' data for validation, and the remaining 35 patients data (70 kidney samples) for testing. Here also, we made sure that kidneys from the same patient were not split into the training, validation, and test cases. We also collected the kidney segmentation data from the same database. A summary of these data is shown in Table III:

TABLE III
SUMMARY OF RELEVANT AND AVAILABLE INFORMATION OF THE CT DATA FROM THE KITS.

Items	Descriptions
Pixel Dimensions	Axial: 3mm (uniformed across cases) Coronal: 0.7816mm (uniformed across cases) Sagittal: 0.7816mm (uniformed across cases)
Total Patients	210
Contrast Agent Used	in all cases
Number of Males	Not available
Number of Females	Not available
Age	Mean: Not available Minimum Age: Not available Maximum Age: Not available
Pathological Kidneys	Either one or both kidneys in all patients

B. Competing Methods

We compare our kidney boundary localization performance to that by Cuingnet et al. 2012 [10] (M2), Criminisi et al. 2010 [8] (M1) and Criminisi et al. 2013 [9] (M3), Gauriau et al. 2015 [11] (M4), Hussain et al. 2017 [18] (M5), Samarakoon et al. 2017 [12] (M6), Humpire et al. 2018 [7] (M7), Xu et al. 2019 [19] (M8), and Jaeger et al. [53] (M9) in Table IV.

We further compare our kidney volume estimation performance to that by Zakhari et al. 2014 [54] (V1), Zhen et al. 2014 [46] (V2), Hussain et al. 2016 [38] (V3), Insensee et al. 2019 [55] (V4), Hou et al. 2019 [56] (V5), Mu et al. 2019 [57] (V6), and Hussain et al. 2017 [18] (V7) in Table V.

C. Implementation Details

Given that the left and right kidneys typically fall in symmetric half volumes of the abdomen area, we use an automatic routine to divide the CT volume medially along the left-right direction. CT intensity values are expected to be identical for the same organ regardless of the scan's origin. Therefore, we clipped the CT intensity to the range $[-200, 350]$HU, a typical range for kidneys. We performed our model training on a workstation with Intel 4.0 GHz i7 processor, an Nvidia Titan Xp GPU with 12GB of VRAM, and 32GB of host memory. While any CNN can be used as S-CNN, we found in our prototyping stage that ResNet-50 performs better than AlexNet, VGGNet, and ResNet-101. On the ground that ResNet-50 is a good trade-off between accuracy and model complexity, and it leads to outperforming other models, we used ResNet-50 in this work. The Mask-RCNN backbone network in Sections III-A2 and III-A3 is also ResNet-50, which are fine-tuned from S-CNN. We fine-tuned S-CNN and both Mask-RCNNs for about 100 epochs with an initial learning rate of 0.001 for the first 50 epochs and 0.0001 for the last 50 epochs. The batch size was 256 for S-CNN and 128 for Mask-RCNNs. We also trained the proposed FCN for about 200 epochs with an initial learning rate of 0.01 for the first 100 epochs and 0.001 for the last 100 epochs. All CNNs were trained using stochastic gradient descent with a momentum of 0.9. These parameters were found to be optimal for our validation data during prototyping. We tested our kidney localization networks on the VGH and KiTS data separately as well as in cross-domain fashion (i.e., trained and validated on VGH and tested on KiTS data, and vice versa). When we trained our localization networks on a particular dataset, the hyperparameters of the networks were empirically optimized based on the validation results for the same dataset. We also adopted a similar validation-based empirical hyperparameter optimization approach for the kidney volume estimation networks.

D. Sørensen–Dice Coefficient to Volume Error Approximation

State-of-the-art kidney volume estimation approaches are segmentation-based, and those report accuracy in terms of the Dice index. On the other hand, Dice cannot be calculated for segmentation-free methods as there is no voxel-level classification. However, the percentage of volume error is not linearly related to the Dice index. Therefore, it is difficult to directly compare the Dice index performance with the percentage of volume error. In Appendix, we derive a mathematical relation between the 'volume error' and 'Sørensen–Dice coefficient' to facilitate an approximate comparison between these two metrics and arrive at the following formula:

$$VE\ (\%) \approx \left| \frac{2}{S_{Dice}} - 2 \right| \times 100, \tag{4}$$

where VE is the volume error, and S_{Dice} is the Sørensen–Dice coefficient.

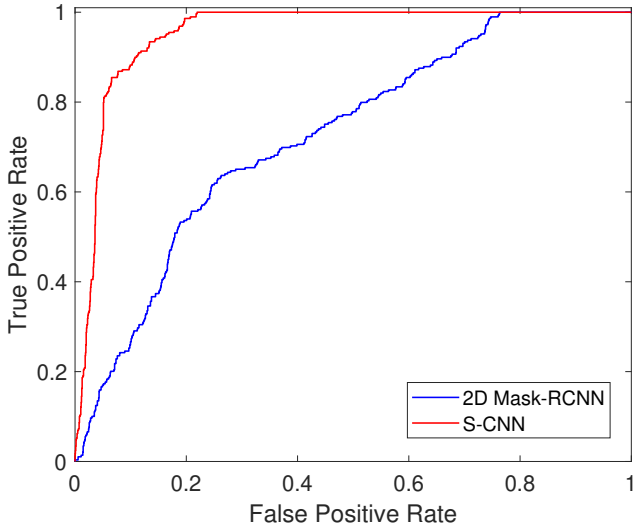

Fig. 2. ROC curves of the kidney cross-section detection performance by the Mask-RCNN and S-CNN on 2D CT slices.

V. Results

In the following sections, we provide a quantitative comparison of the kidney bounding box localization and volume estimation performance between the proposed and state-of-the-art methods.

A. Kidney Bounding Box Localization Performance

At first, we demonstrate the comparative cross-section detection performance by the Mask-RCNN and S-CNN in the axial CT slices in Fig. 2. As we discussed in Section III-A that RCNN typically produces false-positive kidney bounding boxes in those slices that do not contain the kidney, possibly because of lacking the global context in the proposed region, we see in Fig. 2 that false positive rate is higher for Mask-RCNN. We tackled this challenge by using a carefully designed CNN pipeline that starts with S-CNN, narrowing the kidney search region in the abdomen region. We see in Fig. 2 that S-CNN performed better in kidney cross-section detection than Mask-RCNN as depicted by the S-CNN ROC curve.

In this work, we fed axial slices as input to the S-CNN. However, we also checked the kidney localization performance with the coronal and sagittal slices as S-CNN inputs. We found a slightly higher error in kidney bounding box localization

TABLE IV
COMPARISON OF MEAN KIDNEY BOUNDING WALL LOCALIZATION ERROR. IN THE 'CONTRAST USED' COLUMN, 'MIXED' AND 'ALL' INDICATE SOME OR ALL PATIENTS HAVE BEEN ADMINISTERED WITH CONTRAST, RESPECTIVELY. NOT MENTIONED DATA ARE INDICATED WITH (-). X→Y INDICATES THE PROPOSED MODEL BEING TRAINED AND VALIDATED ON X, AND TESTED ON Y DATA. 5-FCV DENOTES 5-FOLD CROSS-VALIDATIONS.

Methods	**Short Name**	**Total Scans**	**Train (Val) Scans**	**Test Scans**	**Resolution (mm)**		**Contrast Used**	**Wall Error (mm)**	
					x, y	**z**		**Left Kidney**	**Right Kidney**
Criminisi et al. 2010 [8]	M1	100	55	45	∼0.5–1.0	∼1.0–5.0	Mixed	17.30±16.50	18.50±18.00
Cuingnet et al. 2012 [10]	M2	233	54	179	∼0.5–1.0	∼0.5–3.0	Mixed	7.00±10.00	7.00±6.00
Criminisi et al. 2013 [9]	M3	400	318	82	∼0.5–1.0	∼1.0–5.0	Mixed	13.60±12.50	16.10±15.50
Gauriau et al. 2015 [11]	M4	130	50	80	∼0.5–1.0	∼0.5–3.0	Mixed	5.50±4.00	5.60±3.00
Hussain et al. 2017 [18]	M5	100	65 (10)	25	∼0.6–1.0	∼1.5–3.0	Mixed	6.19±6.02	5.86±6.40
Samarakoon et al. 2017 [12]	M6	100	54	45	-	-	Mixed	11.52±9.60	10.98±9.60
Humpire et al. 2018 [7]	M7	1884	1130 (377)	377	-	∼1.0–2.0	-	2.67±7.18	3.03±9.30
Xu et al. 2019 [19]	M8	201	118 (13)	70	∼0.6–1.0	∼0.5–5.0	All	4.31±4.18	3.89±3.47
3D Mask-RCNN (KiTS) [53]	M9	210	160 (15)	35	0.7816	3	All	7.83±8.13	8.74±12.63
Proposed (on KiTS data)	-	210	160 (15)	35	0.7816	3	All	2.06±4.39	3.18±14.02
Proposed (KiTS/ResNet-101)	-	210	160 (15)	35	0.7816	3	All	2.55±6.53	3.72±16.33
Proposed (on VGH data)	-	100	65 (10)	25	∼0.6–1.0	∼1.5–3.0	Mixed	**1.93±1.21**	**2.45±1.75**
Proposed (VGH→KiTS)	-	310	90 (10)	210	∼0.6–1.0	∼1.5–3.0	Mixed	4.89±8.13	5.24±9.42
Proposed (KiTS→VGH)	-	310	190 (20)	100	∼0.6–1.0	∼1.5–3.0	Mixed	2.93±3.56	3.13±4.02
Proposed (on KiTS, 5-FCV)	-	210	168 (42)	-	0.7816	3	All	2.10±4.05	3.18±14.67

when S-CNN is applied on either coronal slices or sagittal slices, though not significantly different than the proposed approach. For example, kidney bounding box errors are 2.33 ± 4.58mm (KiTS-left kidneys) and 3.31 ± 14.81mm (KiTS-right kidneys) when S-CNN is applied on the coronal slices, and 2.41 ± 4.67mm (KiTS-left kidneys) and 3.39 ± 15.10mm (KiTS-right kidneys) when S-CNN is applied on the sagittal slices.

Then we quantitatively compared the performance of our proposed kidney localization method with those reported in recent kidney localization approaches M1-8 in Table IV. Our bounding box has six walls, and for a particular kidney sample, we used the mean of the Euclidean distance errors between the estimated and ground-truth locations for all six walls. Please note that each method was independently implemented and tested on different CT databases. However, the type of data these methods used are very similar to ours in terms of resolution, area scanned (i.e., abdominal CT), and scan quality (shown in Table IV). Therefore, our comparisons are conservative, and rather than using our implementation of other contrasting methods, we compare to each authors' best self-reported accuracy values.

The M2 method first, used the M-1 method for coarse localization of both left and right kidneys, then fine-tuned these locations using an additional RF per left/right kidney. The M3 method was an incremental work over the M1, and both used RFs for various organ localization tasks. The M4 method used an extended cascade of RFs to estimate an organ's confidence map, and the prediction was thresholded to obtain a final organ bounding box. The M5 method (our previous method [18]) used a deep CNN-based method for kidney 3D bounding box localization based on 2D orthogonal slice-based kidney candidacy decisions. The M6 method proposed a light RF consisting of fewer nodes than regular RF to localize different organs in the CT scans. The M7 method used separate deep CNNs for images from three orthogonal directions and performed better in organ boundary wall localization. The M8 method used a 3D region proposal network to detect eleven abdominal organs, including the left and right kidneys. The M9 method used the 3D implementation of the Mask-RCNN, designed for use on volumetric medical images. We trained this method on the KiTS data, and the mean boundary wall localization error by this method is ∼8mm. Finally, we show the results of our proposed method on the KiTS and VGH data. Although KiTS datasets contain tumors in the kidney, our method performs better in mean kidney boundary wall localization than other state-of-the-art techniques. However, we observed a higher standard deviation for the right kidney. It happened possibly because some of the right kidneys in this dataset have tumors in the upper pole, confusing boundary estimation. Here, we also show the performance of ResNet-101 as S-CNN and Mask-RCNN backbone. It is shown in [50] that ResNet outperforms other contemporary CNNs (e.g., VGG, GoogleNet) in ImageNet classification task. Therefore, we tested the classification performance between ResNet-50 and ResNet-101 in this work. In Table IV, we see that ResNet-50 marginally performs better than ResNet-101 in kidney boundary wall localization. We further see in Table IV that the proposed method produces the lowest mean wall localization error for both the left and right kidneys on the VGH data. We also checked the proposed method's performance by training it on VGH data and testing on KiTS data, and vice versa. The VGH database is $2\times$ smaller than the KiTS database. Besides, a contrast agent is used in all KiTS database scans, while only half of the scans in VGH have contrast applied. Furthermore, almost all scans in the KiTS database had kidney tumors. Therefore, these two databases are different in many aspects, and this cross-domain experiment yielded a comparatively worse performance in kidney bounding box localization than the in-domain performance, as seen in Table IV. Nevertheless, these cross-domain results are better than most of the state-of-the-art approaches. We finally checked the k-fold cross-validation performance of the proposed method on the KiTS data. In the last row of Table IV, we see that the mean kidney localization performance by the 5-fold cross-validation is similar to that of the non-cross-validation results on the

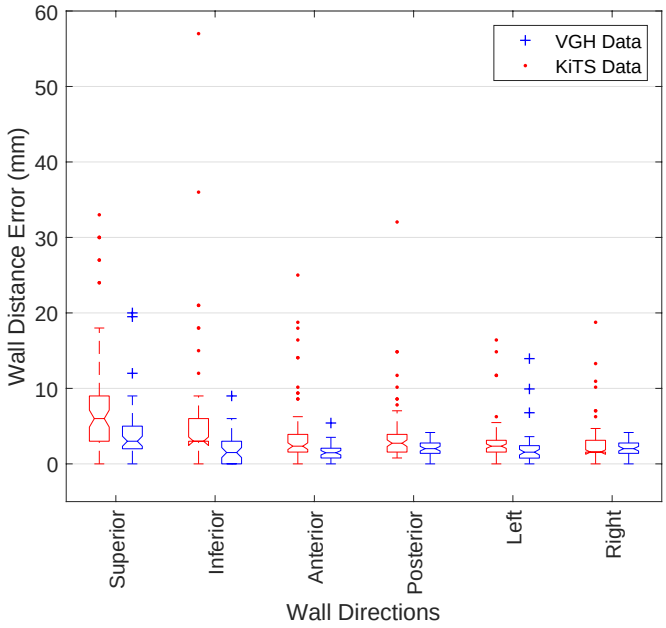

Fig. 3. Box-plot of wall distance error (mm) per wall side of the kidney by the proposed method on the KiTS and VGH data.

KiTS data.

We also estimated the intersection-over-union (IoU) performance of the proposed method for both VGH and KiTS datasets. For the VGH data, we achieved IoU of 0.83 ± 0.10 and 0.82 ± 0.13 for the left and right kidneys, respectively. Similarly, for the KiTS data, we achieved an IoU of 0.75 ± 0.10 and 0.68 ± 0.19 for the left and right kidneys, respectively. We further estimated the distance-to-centroid error by the proposed method for both datasets. For the VGH data, we achieved a distance-to-centroid error of 2.61 ± 1.58mm and 4.54 ± 7.43mm for the left and right kidneys, respectively. Similarly, for the KiTS data, we achieved a distance-to-centroid error of 5.36 ± 4.56mm and 11.92 ± 21.35mm for the left and right kidneys, respectively.

In Fig. 3, we show the box plot of the wall distance errors (mm) by the proposed method in the superior, inferior, anterior, posterior, left, and right directions of a kidney in the KiTS and VGH datasets. This figure further supports the mean error reported in Table IV. However, we also see in this figure that the errors for the KiTS in the superior-inferior direction are comparatively higher than those in the anterior-posterior and left-right directions. Plausible explanations for this scenario could be two-fold: (1) the slice thickness is higher in the axial direction than that in the coronal and sagittal directions, and (2) some of the right kidneys in the KiTS dataset have tumors in the upper pole, which is responsible for the higher boundary error in the axial direction. Although smaller than those for KiTS, we see a similar error pattern in the superior-inferior direction for the VGH data, which may be attributed to point (1) above.

B. Segmentation-free Kidney Volume Estimation Performance

We provide comparative results of our proposed method with those obtained by three generic approaches: a manual clinical method, two regression forest-based approaches, and four deep learning approaches. We estimated the ground truth kidney volumes for both VGH and KiTS data from the kidney annotations by expert radiologists. Nevertheless, first, we show the proposed method's performance quantitatively in Fig. 4, where we illustrate the correlation between the actual and estimated kidney volumes for the KiTS and VGH data, respectively. Estimated p values between the actual and estimated volumes are 0.9042 and 0.7521 for the KiTS and VGH data, respectively. It fails to reject the null hypothesis, and therefore, the actual and estimated kidney volumes do not statistically differ in both datasets.

We also show the quantitative comparative results of our segmentation-free volume estimation approach in Table V. Using (4), we convert the reported S_{Dice} into the percentage of volume error for three state-of-the-art kidney segmentation approaches that achieved ranks first, second, and third in the 2019 KiTS Challenge [32]. We also show these results in Table V.

First, we consider a manual approach V1, typically used

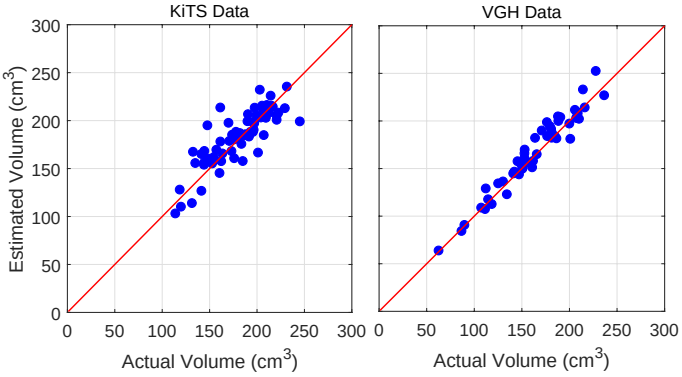

Fig. 4. Scatter plots showing the volume correlations between the actual and proposed FCN-based estimates for the KiTS and VGH data. The Pearson correlation coefficients are 0.9645 and 0.9714, and p-values are 0.9042 and 0.7521 for the KiTS and VGH data, respectively.

TABLE V
VOLUME ESTIMATION ACCURACY COMPARED TO STATE-OF-THE-ART COMPETING METHODS. ACRONYMS USED- SN: SHORT NAME, NTS: NUMBER OF TEST SAMPLES (THERE ARE 2 KIDNEY SAMPLES PER PATIENT), 5-FCV: 5-FOLD CROSS-VALIDATION.

Method Type	**Methods**	**SN**	**NTS**	**Mean Volume Error (%)**
Manual Ellipsoid Fitting	Zakhari et al. 2014 [54]	V1	44	14.20 ± 13.56
Regression Forest (Seg-free)	Zhen et al. 2014 [46]	V2	44	36.14 ± 20.86
	Hussain et al. 2016 [38]	V3	44	9.97 ± 8.69
Deep Learning (Segmentation on KiTS Data)	Insensee et al. [55]	V4	180	~ 5.40
	Hou et al. [56]	V5	180	~ 6.74
	Mu et al. [57]	V6	180	~ 5.58
Deep Learning (Seg-free)	Hussain et al. 2017 [18]	V7	50	8.05 ± 8.91
	Proposed with FCN VGH Data)	-	50	**4.80 ± 3.89**
	Proposed with FCL (VGH Data)	V8	50	5.92 ± 4.50
	Proposed with FCN (KiTS Data)	-	70	7.26 ± 6.80
	5-FCV with FCN (KiTS Data)	-	84	7.79 ± 6.70

by radiologists in clinical settings. The experts obtain three principal axes of a kidney, which correspond to a 3D ellipsoid that approximates that particular kidney. In Table V, we see that the estimated mean volume error (computed by expert radiologists) for this approach is approximately 15% with high standard deviation.

Next, we consider two conventional ML-based approaches, V2 and V3, for segmentation-free kidney volume estimation. The method V2 used a single regression forest, and the corresponding volume estimation error is the worst among the comparing methods. On the other hand, using dual regression forests, our initial work V3, shows better volume estimation accuracy than that by V2.

In Table V, we also included kidney volume estimation performance by three state-of-the-art DL-based approaches, V4, V5, and V6, that ranked first, second and third, respectively, in the 2019 KiTS Challenge. These methods initially reported the kidney segmentation performance in terms of Sørensen–Dice coefficient, which we converted into an approximate mean volume error using (4). Although these methods (V4-6) reported their results on the same KiTS dataset, we did not reimplement these methods as our implementation might result in reduced performance because of the lack of proper parameter tuning. On the other hand, our proposed Sørensen-Dice to volume error approximation gives benefit to the segmentation-based Dice scores as this conversion assumes that the segmentation method produces very low false negatives (see Appendix for details). Thus, the converted volume error by V4-6 methods would not be lower than what we showed in Table V.

Later, we showed the segmentation-free kidney volume estimation performance by a CNN-based approach V7. It showed better volume estimation accuracy than that by V2 and V3 as CNN better captured the rich and complex variability in the kidney anatomy and outperformed the hand-engineered feature representations in V2 and V3.

In this work, we further improve the volume estimation accuracy using a comparatively deeper network than that in V7. This network utilized fully convolution layers except for the layer before the loss calculation to accumulate the network activation as a single value. We see that the FCN approach shows the lowest volume estimation error on the VGH data and is also the lowest among all methods in Table V.

We replaced the final convolution layer 'C5' of size $1 \times 1 \times 1096$ (Fig. 1, box II-B) with a fully connected layer (FCL) of equal size (i.e., 1096×1). This change makes the FCN a CNN, which predicts worse kidney cross-sectional area estimates than that by the FCN (Table V: V8). We infer that the FCN performs better than CNN because of the better feature correspondence among convolution layers. In contrast, a fully connected layer typically learns a completely new set of weights based on the previous layer's activation.

Finally, we show the volume estimation performance of the proposed FCN approach on the KiTS data in Table V. Since almost all kidney samples in this dataset contain tumors of various sizes and shapes, the volume estimation error is slightly higher than that for the VGH data. Comparing this result with that by the state-of-the-art segmentation-based approaches V4-6, we see that V4-6 approaches perform slightly better than the proposed method. However, we emphasize that this small drop in accuracy compared to these state-of-the-art is overwhelmingly compensated by a substantial reduction in the number of model parameters (i.e., our FCN has $200\times$ fewer parameters than those by the state-of-the-art methods (i.e., V4-6), and lighter models are less prone to over-fitting). In the last row of Table V, we see that the mean kidney volume estimation error by the 5-fold cross-validation is similar to that by the non-cross-validation results on the KiTS data.

Although we used Mask-RCNN-produced kidney masks in estimating kidney bounding boxes, we did not use those masks in calculating the kidney volume because the contour of the masks does not follow the exact kidney cross-section. These masks eventually result in an overestimation of the cross-sectional area, and thus volume. For example, the mean volume error is 13.83 ± 12.73 for the KiTS dataset, when the volume is estimated from the Mask-RCNN produced masks. Rather, we used a separate FCN for segmentation-free kidney volume estimation, which can be detached from the proposed Mask-RCNN-FCN working pipeline for stand-alone use. In clinical settings, clinicians often localize kidney ROI manually. In that case, the FCN can be independently used for estimating the total kidney volume.

We also performed the Student t-test between the estimated volumes by the proposed FCN and V4-6 methods for KiTS data. The estimated p values are greater than 0.05. These statistical tests fail to reject the null hypothesis. Therefore, the estimated volumes by the proposed FCN and V4-6 methods do not statistically differ, although our approach uses $200\times$ fewer parameters than that by V4-6.

Next, we show the Bland-Altman plots for the actual and estimated volumes for the VGH and KiTS datasets in Figs. 5(a) and (b), respectively. From these plots, we see that very few samples are outside the limits of the agreement lines, thus proofing the robustness of our FCN approach.

We also visually demonstrate the comparison of the mean distribution of the ground truth and estimated kidney cross-sectional area in Fig. 6 for the VGH and KiTS data. Our FCN predicts the ratio between the kidney cross-sectional area and the 2D ROI area. So, in Fig. 6, we plot the mean kidney area to ROI area ratio for all test kidney samples along the axial direction. Since the kidney span along the axial direction varies across kidneys, we resample all the kidney spans to 25 slices to make those consistent across all samples. We can see in Fig. 6 that the mean and standard deviation of the kidney area to ROI area ratio by the proposed method follows the same trend as the ground truth. Besides, the Jarque-Bera test [58] indicates that these ratio data follow a normal distribution. Therefore, we performed the Student t-test on both samples, and the estimated p values are 0.775 and 0.6442 for the VGH and KiTS data, respectively. These statistical tests fail to reject the null hypothesis. Therefore, the ground truth and estimated kidney area to ROI area ratio do not statistically differ.

Our proposed FCN for segmentation-free kidney volume estimation is also very light in terms of the number of trainable parameters ($\sim$94,000). In contrast, one of the recent and popular segmentation-based organ volume estimation approaches, 3D U-Net [33], has $\sim$19,070,000 trainable parameters, which

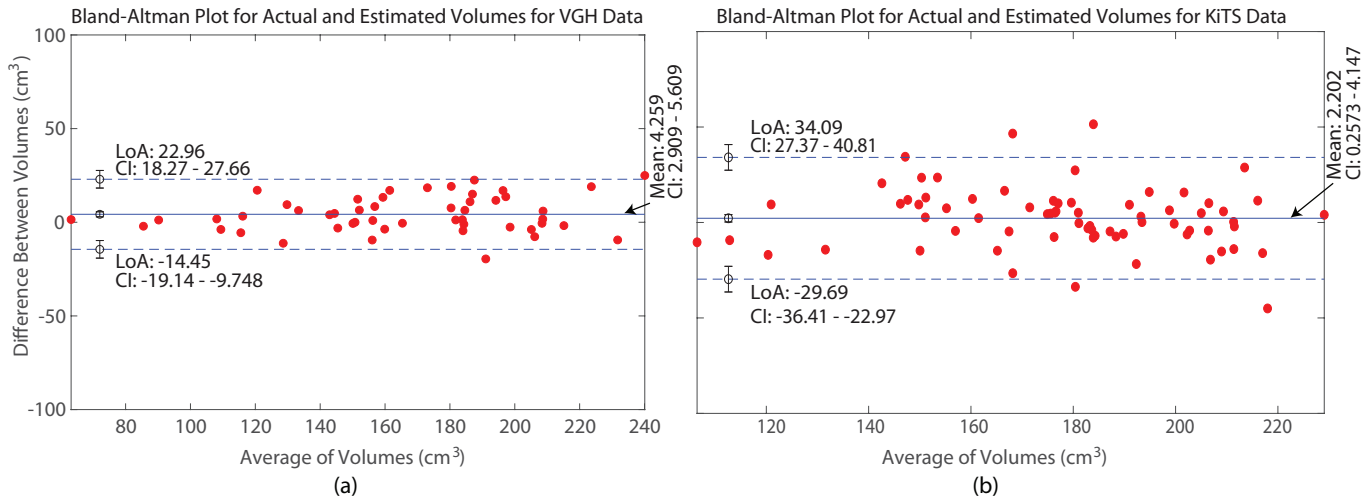

Fig. 5. Bland-Altman plot for the actual and estimated volumes for the (a) VGH and (b) KiTS datasets. LoA denotes limits of agreement and CI denotes confidence interval.

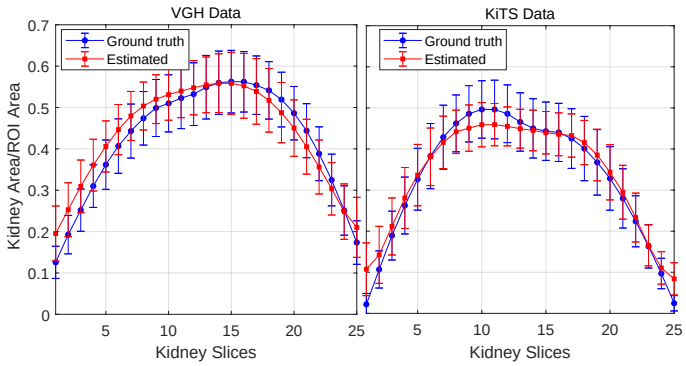

Fig. 6. Distribution of kidney cross-sectional areas for the VGH and KiTS data along the axial direction. The ground truth is estimated from the kidney delineations by an expert radiologist.

are approximately 200× more than that of our proposed FCN approach.

Overall, we see that our proposed method performs better in kidney localization and volume estimation on the VGH data than the KiTS data. The VGH dataset contains mostly healthy kidneys. As we mentioned in Table II that only 12 patients out of 100 had a tumor in one of their kidneys. On the other hand, all 210 patients of KiTS dataset had tumors in either or both kidneys (see Table III). Thus, KiTS dataset offers more challenges in image-based computation than that by the VGH data. This is the plausible reason that our method performs better on VGH dataset than the KiTS dataset.

VI. Conclusion

We proposed a deep learning framework for integrated kidney localization and volume estimation. Our method is capable of (i) kidney localization and (ii) segmentation-free volume estimation from the CT data. Our contribution to the localization step comprises a novel 3-step CNN-based architecture. In the first stage, S-CNN helps reduce false positives in detecting the targeted organ's bounding box. In the second stage, a Mask-RCNN operates strictly on the S-CNN selected slices. Similarly, the Mask-RCNN in the third stage operates strictly on the sagittal slices falling inside the span estimated in the second stage. In addition, we designed our segmentation-free volume estimation task as a 2D patch-based area prediction problem. Furthermore, we showed that an FCN performs better than a CNN of similar parameter numbers in a regression problem. We further derived a mathematical expression to approximate the volume error metric from the Sørensen–Dice coefficient. Our experimental results showed a kidney boundary wall localization error of ∼2.4mm and a mean volume estimation error of ∼5%, and demonstrated similar performance to that of recent segmentation-based approaches but using a simpler deep network. This performance might be improved further by using more fine-tuned deep architectures. That being said, the problem of comprehensive exploration of possible network architectures is beyond the scope of this work. Our future work envisions including the neural architecture search (NAS) to find the optimal network for this work.

We also emphasize that the kidney volume estimation part of our proposed method does not require segmentation, rather it requires a ground truth area value per slice, regardless of how that value was obtained. The area value may have been obtained via manual segmentation, but conceptually it may have been obtained without segmentation, e.g., by measuring some other proxy variables. Besides, we predict the volume directly in inference, without having to segment. We further emphasize that the 'segmentation-free' term is used in the literature, e.g., [40]–[49] in the same context as ours (i.e., no explicit segmentation is required in training), therefore we adopt that same established terminology.

Accurate estimation of total kidney volume is often very crucial, in particular CKD assessment, where the total kidney volume error needs to be lower than 5% to capture disease progression, particularly in the early stages. Although our proposed FCN approach shows an error below 5% for the VGH data, this dataset contains mostly healthy kidneys. In addition, our datasets (i.e., VGH and KiTS) do not include

any pathological cases of ADPKD. Therefore, although our method performs well on healthy and typical pathological kidneys (i.e., CKD without major 'morphological' changes due to disease), further studies are needed to validate the efficacy of the proposed volume estimation method on ADPKD cases. We also plan to incorporate shape priors into the deep neural network, which is known to improve organ-related prediction tasks [59]. We also envision to use the classification and contrastive semantic alignment (CCSA) loss [60] or domain invariant representations [61] that aims to learn domain-invariant features. This approach would allow to learn kidney-specific features irrespective of the data source, and thus, it would avoid a deep model being negatively affected by the domain-specific features. It is also widely accepted that it is easier to collect healthy data compared to pathological cases. We also showed in Table IV that cross-domain training testing leads to worse kidney localization performance when one domain contains a significant number of pathological kidneys. Therefore, to address this imbalance in training data, another possible future direction is to synthesize pathological renal cases, e.g., as done for lung nodules in [62] and brain tumor synthesis in [63]. This synthesis approach may enhance the result of cross-domain training testing.

We discussed in Section V-B that the contour of the Mask-RCNN generated kidney mask does not follow the exact kidney cross-section, which results in an overestimation of the kidney volume. In contrast, our proposed FCN was found to be better in kidney volume estimation than that by a Mask-RCNN. However, the accuracy of a kidney mask produced by the Mask-RCNN can be further improved, which we aim to investigate in future.

Finally, we envision to incorporate the proposed kidney localization technique in our future kidney radiomic studies, as previous deep learning studies of ours [64]–[67], require kidney localization as a preprocessing step. We believe that the proposed kidney localization approach would further contribute to improve the outcomes of those studies. Our future work can also explore the ability of segmentation-free methods to predict radiomic features, as recently explored by Klyuzhin et al. [68].

APPENDIX
SØRENSEN–DICE COEFFICIENT TO VOLUME ERROR APPROXIMATION

We derive the mathematical relation between the volume error and Sørensen-Dice coefficient to facilitate an approximate comparison these two metrics. The Sørensen-Dice coefficient for Boolean data is defined as:

$$S_{Dice} = \frac{2a}{2a+b+c}, \tag{5}$$

where a represents true positives, b represents false positives, and c represents false negatives. The volume error (*VE*) is defined as:

$$VE = \left| \frac{(V_{estimated} - V_{true}) \times \not{D}}{V_{true} \times \not{D}} \right|, \tag{6}$$

where $V_{estimated}$ and V_{true} are the estimated volume and true volume of an object, respectively, and $D = d_{ax} \times d_{co} \times d_{sa}$, and d_{ax}, d_{co} and d_{sa} are the voxel dimensions in the axial, coronal and sagittal directions, respectively. We can also write (6) in terms of a, b and c as:

$$VE = \left| \frac{(a+b)-(a+c)}{a+c} \right|. \tag{7}$$

Now, by rewriting (5), we get,

$$S_{Dice} = \frac{2a}{(a+b)-(a+c)+2(a+c)}. \tag{8}$$

We further rearrange (8) as:

$$(a+b)-(a+c) = \frac{2a}{S_{Dice}} - 2(a+c). \tag{9}$$

Now, dividing both sides of (9) by $(a+c)$ and taking the absolute value, we get,

$$\left| \frac{(a+b)-(a+c)}{a+c} \right| = \left| \frac{\frac{2a}{S_{Dice}} - 2(a+c)}{a+c} \right|. \tag{10}$$

By replacing the left side of (10) as *VE* from (7), we get:

$$VE = \left| \frac{2a}{S_{Dice}(a+c)} - \frac{2\cancel{S_{Dice}}\cancel{(a+c)}}{\cancel{S_{Dice}}\cancel{(a+c)}} \right|. \tag{11}$$

We see in (5) that when $S_{Dice} \to 1$, then the part of denominator $(b+c) \to 0$. Thus, for a state-of-the-art well-performing segmentation approach, we can assume that $a \gg c$. Based on these assumptions, we can also assume that $a \approx (a+c)$. Thus, we can rewrite (11) as:

$$VE\ (\%) = \lim_{b \to 0, c \to 0} \left| \frac{2}{S_{Dice}} - 2 \right| \times 100. \tag{12}$$

ACKNOWLEDGMENT

This work is supported by grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada. We also thank NVIDIA Corporation for supporting our research through their GPU Grant Program by donating the GeForce Titan Xp.

REFERENCES

[1] M. G. Linguraru, *et al.*, "Renal tumor quantification and classification in contrast-enhanced abdominal CT," *Pattern Recog.*, vol. 42, no. 6, pp. 1149–1161, 2009.

[2] A. C. Webster, E. V. Nagler, R. L. Morton, and P. Masson, "Chronic kidney disease," *Lancet*, vol. 389, no. 10075, pp. 1238–1252, 2017.

[3] P. Romagnani, *et al.*, "Chronic kidney disease," *Nature Rev. Disease Prim.*, vol. 3, no. 1, pp. 1–24, 2017.

[4] E. Porrini, *et al.*, "Estimated GFR: Time for a critical appraisal," *Nature Rev. Nephrol.*, vol. 15, no. 3, pp. 177–190, 2019.

[5] A. Diez, *et al.*, "Correlation between CT-based measured renal volumes and nuclear-renography-based split renal function in living kidney donors. Clinical diagnostic utility and practice patterns," *Clin. Transplan.*, vol. 28, no. 6, pp. 675–682, 2014.

[6] E. Widjaja, J. Oxtoby, T. Hale, P. Jones, P. Harden, and I. McCall, "Ultrasound measured renal length versus low dose CT volume in predicting single kidney glomerular filtration rate," *British J. Rad*, vol. 77, no. 921, pp. 759–764, 2004.

[7] G. E. Humpire-Mamani, A. A. A. Setio, B. van Ginneken, and C. Jacobs, "Efficient organ localization using multi-label convolutional neural networks in thorax-abdomen CT scans," *Phys. Med. Biol.*, vol. 63, no. 8, p. 085003, 2018.

[8] A. Criminisi, J. Shotton, D. P. Robertson, and E. Konukoglu, "Regression forests for efficient anatomy detection and localization in CT studies." in *Proc. Int. Work. Med. Comp. Vis.*, vol. 2010, pp. 106–117, 2010.

[9] A. Criminisi, *et al.*, "Regression forests for efficient anatomy detection and localization in computed tomography scans," *Med. Img. Anal.*, vol. 17, no. 8, pp. 1293–1303, 2013.

[10] R. Cuingnet, R. Prevost, D. Lesage, L. D. Cohen, B. Mory, and R. Ardon, "Automatic detection and segmentation of kidneys in 3D CT images using random forests," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Springer, 2012, pp. 66–74.

[11] R. Gauriau, R. Cuingnet, D. Lesage, and I. Bloch, "Multi-organ localization with cascaded global-to-local regression and shape prior," *Med. Imag. Anal.*, vol. 23, no. 1, pp. 70–83, 2015.

[12] P. N. Samarakoon, E. Promayon, and C. Fouard, "Light random regression forests for automatic multi-organ localization in CT images," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*. IEEE, 2017, pp. 371–374.

[13] X. Zhou, *et al.*, "Automatic organ segmentation on torso CT images by using content-based image retrieval," in *Med. Imag. 2012: Imag. Process.*, vol. 8314. International Society for Optics and Photonics, 2012, p. 83143E.

[14] X. Zhou, *et al.*, "Automatic anatomy partitioning of the torso region on CT images by using multiple organ localizations with a group-wise calibration technique," in *Med. Imag. 2015: Comp.-Aid. Diag.*, vol. 9414. International Society for Optics and Photonics, 2015, p. 94143K.

[15] X. Zhou, *et al.*, "A universal approach for automatic organ segmentations on 3D CT images based on organ localization and 3D GrabCut," in *Med. Imag. 2014: Comp.-Aid. Diag.*, vol. 9035. International Society for Optics and Photonics, 2014, p. 90352V.

[16] G. E. H. Mamani, A. A. A. Setio, B. van Ginneken, and C. Jacobs, "Organ detection in thorax abdomen CT using multi-label convolutional neural networks," in *Med. Imag. 2017: Comp.-Aid. Diag.*, vol. 10134. International Society for Optics and Photonics, 2017, p. 1013416.

[17] X. Lu, D. Xu, and D. Liu, "Robust 3D organ localization with dual learning architectures and fusion," in *Proc. Int. Work. Large-Scale Annot. Biomed. Data Expert Label Synthesis.* Springer, 2016, pp. 12–20.

[18] M. A. Hussain, A. Amir-Khalili, G. Hamarneh, and R. Abugharbieh, "Segmentation-free kidney localization and volume estimation using aggregated orthogonal decision CNNs," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Springer, 2017, pp. 612–620.

[19] X. Xu, F. Zhou, B. Liu, D. Fu, and X. Bai, "Efficient multiple organ localization in CT image using 3D region proposal network," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1885–1898, 2019.

[20] M. Shehata, *et al.*, "3D kidney segmentation from abdominal diffusion MRI using an appearance-guided deformable boundary," *PloS one*, vol. 13, no. 7, p. e0200082, 2018.

[21] B. Glocker, O. Pauly, E. Konukoglu, and A. Criminisi, "Joint classification-regression forests for spatially structured multi-object segmentation," in *Proc. Eur. Conf. Comp. Vis. (ECCV)*, pp. 870–881, 2012.

[22] F. Khalifa, A. Soliman, A. Elmaghraby, G. Gimel'farb, and A. El-Baz, "3D kidney segmentation from abdominal images using spatial-appearance models," Computat. Math. Meth. Medicine, vol. 2017, 2017.

[23] E. Hristova, H. Schulz, T. Brosch, M. P. Heinrich, and H. Nickisch, "Nearest neighbor 3D segmentation with context features," in *Med. Imag. 2018: Imag.Process.*, vol. 10574. International Society for Optics and Photonics, 2018, p. 105740M.

[24] F. Zhao, P. Gao, H. Hu, X. He, Y. Hou, and X. He, "Efficient kidney segmentation in micro-CT based on multi-atlas registration and random forests," *IEEE Access*, vol. 6, pp. 43712–43723, 2018.

[25] S. Chen, *et al.*, "Towards automatic abdominal multi-organ segmentation in dual energy CT using cascaded 3D fully convolutional network," *arXiv preprint*, arXiv:1710.05379, 2017.

[26] E. Gibson, *et al.*, "Automatic multi-organ segmentation on abdominal CT with dense V-networks," *IEEE Trans. Med. Imag.*, vol. 37, no. 8, pp. 1822–1834, 2018.

[27] V. V. Valindria, *et al.*, "Multi-modal learning from unpaired images: Application to multi-organ segmentation in CT and MRI," in *Proc. IEEE Winter Conf. Appl. Comp. Vis. (WACV)*. IEEE, 2018, pp. 547–556.

[28] W. Thong, S. Kadoury, N. Piché, and C. J. Pal, "Convolutional networks for kidney segmentation in contrast-enhanced CT scans," *Comp. Meth. Biomech. Biomed. Eng.: Imag. Visual.*, vol. 6, no. 3, pp. 277–282, 2018.

[29] D. Keshwani, Y. Kitamura, and Y. Li, "Computation of total kidney volume from CT images in autosomal dominant polycystic kidney disease using multi-task 3D convolutional neural networks," in *Proc. Int. Work. Mach. Learn. Med. Imag. (MLMI)*. Springer, 2018, pp. 380–388.

[30] K. Sharma, *et al.*, "Automatic segmentation of kidneys using deep learning for total kidney volume quantification in autosomal dominant polycystic kidney disease," *Scient. Rep.*, vol. 7, no. 1, p. 2049, 2017.

[31] V. Groza, T. Brosch, D. Eschweiler, H. Schulz, S. Renisch, and H. Nickisch, "Comparison of deep learning-based techniques for organ segmentation in abdominal CT images," in *Proc. Conf. Med. Imag. Deep Learn. (MIDL)*, 2018, pp. 1–3.

[32] N. Heller, *et al.*, "The KiTS19 Challenge Data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes," *ArXiv Preprint*, arXiv:1904.00445, 2019.

[33] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Springer, 2016, pp. 424–432.

[34] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*. IEEE, 2016, pp. 565–571.

[35] L. Liu, J. Cheng, Q. Quan, F. X. Wu, Y. P. Wang, and J. Wang, "A survey on U-shaped networks in medical image segmentations," *Neurocomp.*, vol. 409, pp. 244–258, 2020.

[36] J. Schlemper, *et al.*, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Imag. Anal.*, vol. 53, pp. 197–207, 2019.

[37] L. Rundo, *et al.*, "USE-Net: Incorporating Squeeze-and-Excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets," *Neurocomp.*, vol 365, pp. 31–43, 2019.

[38] M. A. Hussain, G. Hamarneh, T. W. O'Connell, M. F. Mohammed, and R. Abugharbieh, "Segmentation-free estimation of kidney volumes in CT with dual regression forests," in *Proc. Int. Work. Mach. Learn. Med. Imag. (MLMI)*. Springer, 2016, pp. 156–163.

[39] L. Rundo, *et al.*, "Tissue-specific and interpretable sub-segmentation of whole tumour burden on CT images by unsupervised fuzzy clustering," *Comp. Biol. Med.*, p. 103751, 2020.

[40] M. Afshin, I. B. Ayed, A. Islam, A. Goela, T. M. Peters, and S. Li, "Global assessment of cardiac function using image statistics in MRI," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Springer, 2012, pp. 535–543.

[41] M. Afshin, *et al.*, "Regional assessment of cardiac left ventricular myocardial function via MRI statistical features," *IEEE Trans. Med. Imag.*, vol. 33, no. 2, pp. 481–494, 2013.

[42] M. Afshin, *et al.*, "Automated assessment of regional left ventricular function for cardiac MRI with minimal user interaction," in *Proc. Annual Meet.-Radiol. Soc. North Amer. (RSNA)*, 2012.

[43] O. Zettinig, *et al.*, "Data-driven estimation of cardiac electrical diffusivity from 12-lead ECG signals," *Med. Imag. Anal.*, vol. 18, no. 8, pp. 1361–1376, 2014.

[44] Z. Wang, M. Salah, I. Ayed, A. Islam, A. Goela, and S. Li, "Biventricular volume estimation for cardiac functional assessment," in *Proc. Annual Meet.-Radiol. Soc. North Amer. (RSNA)*, 2013.

[45] Z. Wang, M. B. Salah, B. Gu, A. Islam, A. Goela, and S. Li, "Direct estimation of cardiac bi-ventricular volumes with an adapted Bayesian formulation," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 4, pp. 1251–1260, 2014.

[46] X. Zhen, Z. Wang, A. Islam, M. Bhaduri, I. Chan, and S. Li, "Direct estimation of cardiac bi-ventricular volumes with regression forests," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Springer, 2014, pp. 586–593.

[47] X. Zhen, Z. Wang, A. Islam, M. Bhaduri, I. Chan, and S. Li, "Direct volume estimation without segmentation," in *Med. Imag. 2015: Imag. Process.*, vol. 9413. International Society for Optics and Photonics, 2015, p. 94132G.

[48] X. Zhen, Z. Wang, A. Islam, M. Bhaduri, I. Chan, and S. Li, "Multi-scale deep networks and regression forests for direct bi-ventricular volume estimation," *Med. Imag. Anal.*, vol. 30, pp. 120–129, 2016.

[49] S. A. Taghanaki, *et al.*, "Segmentation-free direct tumor volume and metabolic activity estimation from PET scans," *Computer. Med. Imag. Graph.*, vol. 63, pp. 52–66, 2018.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2016, pp. 770–778.

[51] K.-Y. Kang, *et al.*, "A comparative study of methods of estimating kidney length in kidney transplantation donors," *Nephrol. Dial. Transplant.*, vol. 22, no. 8, pp. 2322–2327, 2007.

[52] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2017, pp. 2961–2969.

[53] P. Jaeger and G Kohler, "Medical detection toolkit," [Online]. Available: https://github.com/MIC-DKFZ/medicaldetectiontoolkit. Accessed on: Dec 10, 2020.

[54] N. Zakhari, B. Blew, and W. Shabana, "Simplified method to measure renal volume: The best correction factor for the ellipsoid formula volume calculation in pre-transplant computed tomographic live donor," *Urology*, vol. 83, no. 6, pp. 1444–e15, 2014.

[55] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, pp. 1–9, 2020.

[56] X. Hou, C. Xie, F. Li, and Y. Nan, "Cascaded semantic segmentation for kidney and tumor," [Online]. Available: https://kits.lib.umn.edu/wp-content/uploads/2019/11/PingAnTech_3e.pdf. Accessed on: Dec 10, 2020.

[57] G. Mu, Z. Lin, M. Han, G. Yao, and Y. Gao, "Segmentation of kidney tumor by multi-resolution VB-nets," [Online]. Available: https://kits.lib.umn.edu/wp-content/uploads/2019/11/gr_6e.pdf. Accessed on: Dec 10, 2020.

[58] T. Thadewald and H. Büningg, "Jarque–Bera test and its competitors for testing normality–a power comparison," *J. Appl. Stat.*, vol. 34, no. 1, pp. 87–105, 2007.

[59] Z. He, S. Bao, and A. Chung, "3D deep affine-invariant shape learning for brain MR image segmentation," in *Deep Learn. Med. Imag. Anal. Multi. Learn. Clin. Dec. Supp*. Springer, 2018, pp. 56–64.

[60] C. Yoon, G. Hamarneh, and R. Garbi, "Generalizable feature learning in the presence of data bias and domain class imbalance with application to skin lesion classification," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Springer, 2019, pp. 365–373.

[61] S. Aslani, V. Murino, M. Dayan, R. Tam, D. Sona, and G. Hamarneh, "Scanner invariant multiple sclerosis lesion segmentation from MRI," in *Proc. 2020 IEEE Int. Symp. Biomed. Imag. (ISBI)*. IEEE, 2020, pp. 781–785.

[62] C. Han, *et al.*, "Synthesizing diverse lung nodules wherever massively: 3D multi-conditional GAN-based CT image augmentation for object detection," in *Proc. 2019 Int. Conf. 3D Vis. (3DV)*. IEEE, 2019, pp. 729–737.

[63] B. Yu, L. Zhou, L. Wang, J. Fripp, and P. Bourgeat, "3D cGAN based cross-modality MR image synthesis for brain tumor segmentation," in *Proc. 2018 IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*. IEEE, 2018, pp. 626–630.

[64] M. A. Hussain, A. Amir-Khalili, G. Hamarneh, and R. Abugharbieh, "Collage CNN for renal cell carcinoma detection from CT," in *Proc. Int. Work. Mach. Learn. Med. Imag. (MLMI)*. Springer, 2017, pp. 229–237.

[65] M. A. Hussain, G. Hamarneh, and R. Garbi, "ImHistNet: Learnable image histogram based DNN with application to noninvasive determination of carcinoma grades in CT scans," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Springer, 2019, pp. 130–138.

[66] M. A. Hussain, G. Hamarneh, and R. Garbi, "Renal cell carcinoma staging with learnable image histogram-based deep neural network," in *Proc. Int. Work. Mach. Learn. Med. Imag. (MLMI)*. Springer, 2019, pp. 533–540.

[67] M. A. Hussain, G. Hamarneh, and R. Garbi, "Noninvasive determination of gene mutations in clear cell renal cell carcinoma using multiple instance decisions aggregated CNN," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Springer, 2018, pp. 657–665.

[68] I. S. Klyuzhin, Y. Xu, A. Ortiz, J. M. L. Ferres, G. Hamarneh, and A. Rahmim, "Testing the ability of convolutional neural networks to learn radiomic features," *medRxiv*, [Online]. Available: https://doi.org/10.1101/2020.09.19.20198077. Accessed on: Dec 10, 2020.