IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 February 2023, accepted 5 March 2023, date of publication 14 March 2023, date of current version 17 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3257049

RESEARCH ARTICLE

Urinary Stones Segmentation in Abdominal X-Ray Images Using Cascaded U-Net Pipeline With Stone-Embedding Augmentation and Lesion-Size Reweighting Approach

WONGSAKORN PREEDANAN[1], KENJI SUZUKI[1], TOSHIAKI KONDO[2], MASAKI KOBAYASHI[3], HAJIME TANAKA[3], JUNICHIRO ISHIOKA[3], YOH MATSUOKA[3], YASUHISA FUJII[3], AND ITSUO KUMAZAWA[1], (Member, IEEE)

[1]Department of Information and Communications Engineering, School of Engineering, Tokyo Institute of Technology, Yokohama 226-8503, Japan
[2]School of Information and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Prathumthani 12121, Thailand
[3]Department of Urology, Tokyo Medical and Dental University, Tokyo 113-8519, Japan

Corresponding author: Wongsakorn Preedanan (job.preedanan@gmail.com)

This work was supported in part by the Japan Science and Technology Agency Department of R&D for Future Creation (JST-Mirai) Program, Japan, under Grant JPMJMI20B8.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of Tokyo Institute of Technology under Application No. JB0000797174, and performed in line with the Declaration of Helsinki.

ABSTRACT In this research, we proposed a two-stage pipeline for segmenting urinary stones. The first stage U-Net generated the map localizing the urinary organs in full abdominal x-ray images. Then, this map was used for creating partitioned images input to the second stage U-Net to reduce class imbalance and was also used in stone-embedding augmentation to increase a number of training data. The U-Net model was trained with the combination of real stone-contained images and synthesized stone-embedded images to segment urinary stones on the partitioned input images. In addition, we proposed to use an inverse weighting method in the focal Tversky loss function in order to rebalance lesion size. The U-Net model using our proposed pipeline produced a 71.28% pixel-wise F_2 score and a 69.82% region-wise F_2 score, which were 2.88% and 7.63%, respectively, higher than those of a baseline method. Experimental results showed that the proposed method improved urinary stone segmentation results, especially for small stones and stones in uncommon locations.

INDEX TERMS Computer-aided detection and diagnosis, urinary stone, deep learning, image segmentation, abdominal X-ray imaging.

I. INTRODUCTION

A urinary stone, or renal calculi, is one of the most frequent abnormalities in the urinary system. These hard mineral deposits form in the kidneys and can travel down the urinary tract into the ureters and bladder, causing severe discomfort, as well as other complications if left untreated [1]. Each year, more than half a million people visit emergency rooms for urinary stone problems. The accurate and early detection of urinary stones is a crucial task for medical doctors for the diagnosis and management of this condition. [2]. Urinary stones can be detected by various medical imaging modalities such as CT-scanning, ultrasonography, and x-ray imaging. An abdominal x-ray or KUB (Kidney, Ureter, Bladder) radiography can detect urinary stones because most stones are calcified. Although radiography is not frequently used for stone detection, advantages of this method include relatively lower radiation exposure than CT imaging and a lower cost than ultrasonography and CT imaging [3]. However, stone detection in plain x-ray images is often difficult for radiologists and

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Sharif.

other medical doctors because of the following challenges. In radiography, stones and other anatomical structures are projected in a 2D image; hence small stones are difficult to identify due to the overlaps, and some types of stones is poorly visible.

In this paper, we propose the pipeline of a cascaded framework based on the U-Net model for the urinary stones segmentation in plain x-ray images. The significant contributions of our work are summarized as follows:

1.) We propose the pipeline of urinary stone segmentation by using two stages of U-Net models, reducing class imbalance and improving segmentation performance.

2.) We utilize the stone-free images by proposing the stone-embedding augmentation implementing during training the second stage U-Net.

3.) We modify the training loss function by implementing the lesion-size reweighting approach, improving the recall rate of small stones.

II. RELATED WORKS

A computer-aided diagnosis (CAD) for urinary stones is demanded, because it can support radiologists and other medical doctors in various processes, such as screening, treatment planning and treatment follow-up. As such, many researchers have proposed approaches to detect or segment urinary stones in ultrasonography [4], [5] or CT-scan imaging [6], [7]. However, as the limitation of x-ray images for urinary stones diagnosis, there was only a few work proposing CAD for this modality [8].

An increased use of medical images causes more burden of their interpretation for medical experts. Recently, deep learning has been widely used to support medical doctors in various medical imaging tasks due to its high accuracy comparing with traditional methods. The performance of deep learning is typically dependent on the amount of training data [9]. However, the availability of medical image datasets is usually limited compared with other domains because the data acquisition and the preparation of image ground truth are usually costly and need experts. Generally, the number of medical anomaly samples is less than the normal data; accordingly, many techniques to generate new positive samples have been proposed and used to create extra training samples. For example, the new lesions are simulated using a mathematical model and inserted into the existing medical images such as the study in [10] for lung nodules in CT, the one in [11] for breast lesions in mammography, and the one in [12] for digital breast tomosynthesis (DBT). In [13], [14], [15], and [16], an actual lesion is firstly extracted from real CT-scan images and then inserted into a new location on other images using image-processing techniques. Our previous works [17], [18] proposed augmentation techniques for creating synthetic images to improve the segmentation network's performance by increasing the number and diversity of training data. However, our methods still had a limitation for detecting stones in some cases, particularly small stones and stones at bladder region.

Class imbalance is common problem in many medical imaging applications. The lesion region can be extremely smaller than the background region; therefore, the small lesions are more likely not to be detected or well-segmented because their information is lost during the downsampling in deep learning models. The researches such as the kidney tumor segmentation [19], [20], [21] and brain tumor segmentation [22] proposed the pipeline consisted of multiple stages. With these approaches, a small lesion can be segmented more precisely than using a full image as the input.

III. METHODOLOGY

The overview of proposed pipeline of this study is shown in Fig. 1.

A. ABDOMINAL X-RAY IMAGES DATASET

In this work, we used our private dataset, consisting of 1,156 abdominal x-ray images containing urinary stone(s) called stone-contained images (I_{sc}) and 1,200 abdominal x-ray images without any urinary stone called stone-free images (I_{sf}), as shown in Fig. 2 (left) and Fig.2 (right), respectively. The ground-truth masks of urinary stones (Fig. 2 (middle)), which require medical knowledge and precise annotation skills, were manually drawn by the urology experts for every stone-contained image. Only 600 stone-free images were randomly chosen for using in the first stage, whereas the other 600 stone-free images and all stone-contained images were used in the second stage.

B. KUB REGION MAP GENERATION STAGE

1) STONE LOCATION MAP

Based on medical domain knowledge, urinary stones can only be found in kidneys, ureters, and bladder. In this stage, we generated the stone location maps (Fig.3 (2^{nd} image)), representing the approximate locations of these organs in an abdominal X-ray image (Fig.3 (1^{st} image)). Firstly, all training images were resized to 256×256 pixels. We used the U-Net model [23] for training a network to generate coarse stone location maps. The network was trained with 500 full abdominal x-ray images and their manual segmentation of stone location maps from scratch for 100 epochs and used Adam optimizer [24] with a learning rate of 10^{-3} to minimize the Dice coefficient loss (DL). In post-processing, the output images were binarized using a 0.5 threshold value and implemented morphological operations to connect all white components and remove the small ones.

2) KUB REGION MAP

Next, we used stone location maps to create KUB region maps (Fig.3(4^{th} image)), representing kidneys, ureters, and

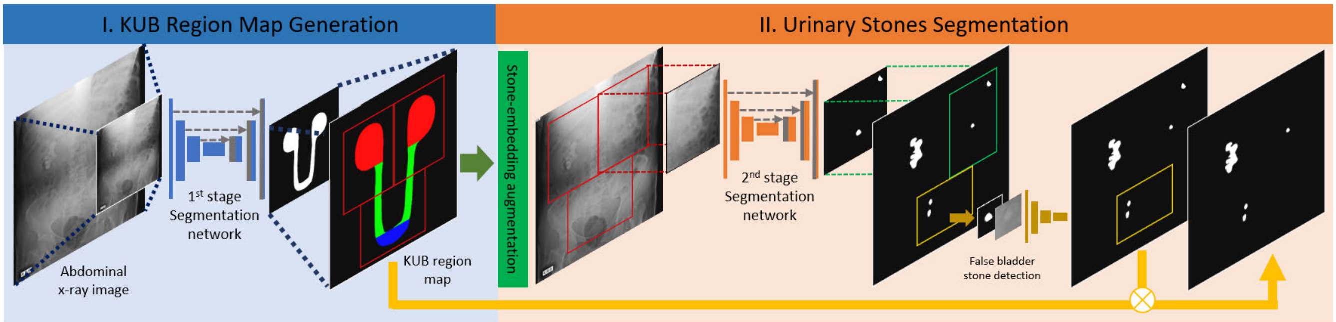


FIGURE 1. The overview of proposed pipeline for segmenting urinary stones. The 1^{st} stage U-Net generates KUB region maps from downsampled abdominal x-ray images. The results from this stage are upsampled and used for stone-embedding augmentation, and cropping a full image into 3 partitions based on the anatomical region. The 2^{nd} stage U-Net processes the partitioned images and generates the segmented stones results. Post-processing consists of the detection of false bladder stones and the removal of lesions outside the stone localization map.

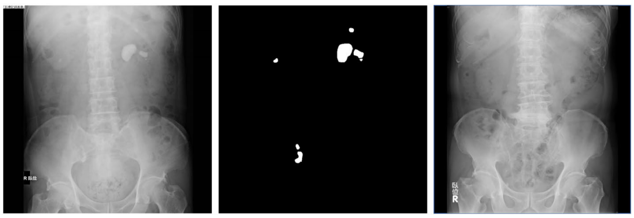


FIGURE 2. Illustration of an abdominal x-ray image with stones (left), corresponding gold standard manual segmentation of the stones (middle) and a stone-free abdominal x-ray image (right).

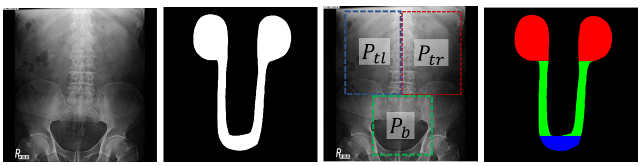


FIGURE 3. An abdominal x-ray image (1^{st} image), its segmentation of stone location map (2^{nd} image), its partitioned bounding boxes (3^{rd} image) and its KUB region map (4^{th} image), where kidneys, ureters, and bladder regions are represented in red, green, and blue, respectively.

bladder regions. Let (x^m_{tl}, y^m_{tl}) be the top-left coordinate and (w^m, h^m) be its width and height of the bounding box of a stone location map. We cropped this bounding box into 3 partitions; top-left partition (P_{tl}), top-right partition (P_{tr}), and bottom partition (P_b), as shown in Fig.3(3^{rd} image). The coordinates of top-left partition (x^p_l, y^p_l), top-right partition (x^p_r, y^p_r), and bottom partition (x^p_b, y^p_b) are defined as Eqs.(1) - (3), respectively. The width and height of each partition (w^p, h^p) are defined as Eq.(4).

$$(x^p_l, y^p_l) = (x^m_{tl} - b_x, y^m_{tl} - b_y) \tag{1}$$
$$(x^p_r, y^p_r) = (x^m_{tl} + w^m/2, y^m_{tl} - b_y) \tag{2}$$
$$(x^p_b, y^p_b) = (x^m_{tl} + w^m/4, y^m_{tl} + h^m/2 + b_y) \tag{3}$$
$$(w^p, h^p) = (w^m/2 + b_x, h^m/2 + b_y) \tag{4}$$

where b_x and b_y are the border size in the vertical and horizontal direction, respectively, which are set to 10% of the width and height of the stone location map's bounding box. Then, we split the stone location map into kidneys, ureters, and bladder regions. S_{tl} and S_{tr} which are the region separating lines used in P_{tl} and P_{tr} are defined as Eqs. (5) and (6), respectively, while the separating line S_b used in P_b is defined as Eq. (7).

$$S_{tl} = \arg\min_j \Delta_j(\sum_{i=0}^{w^p} P_{tl}(i,j)) \tag{5}$$
$$S_{tr} = \arg\min_j \Delta_j(\sum_{i=0}^{w^p} P_{tr}(i,j)) \tag{6}$$
$$S_b = \arg\max_j \Delta_j(\sum_{i=0}^{w^p} P_b(i,j)) \tag{7}$$

C. URINARY STONES SEGMENTATION STAGE

1) DATA AUGMENTATION APPROACHES

In this stage, the traditional augmentation method, including rotation [-5°, 5°], horizontal flipping was implemented for both stone-contained and stone-free samples, while the proposed augmentation method was implemented for only stone-free samples, as shown in Fig. 4(left). We proposed to use stone-embedding algorithm to generate new training images, which urinary stones were inserted while preserving the background texture of the target image. Firstly, urinary stone images were cropped and multiplied with cropped stone mask to remove the region outside stone pixels, as shown in the 2^{nd} row in Fig. 4(right). By using KUB region maps, all stones were separated into three categories based on their location: kidney stones, ureteral stones, and bladder stones. We selected only small and medium stones, which have size between 20 to 500 pixels, which are the hard samples to use in this augmentation.

During the augmentation process, we randomly selected 1 to 3 target location(s) (x_t, y_t) from the KUB region map of each target image to be the center of a cropped region of the target image (f_t) that has the same size as the selected source image (f_s). Then, the source stone image f_s was randomly selected based on the region of selected locations in the KUB

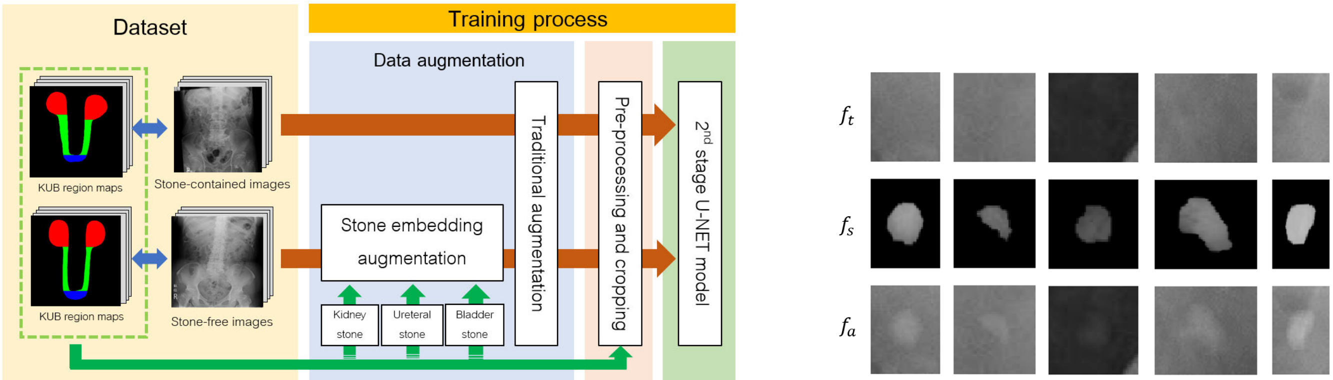


FIGURE 4. Training framework in the 2nd stage (left), and cropped targets (f_t), cropped sources (f_s) and results (f_a) by stone-embedding augmentation (right).

region map (kidneys, ureters, or bladder) and applied with simple augmentation methods: rotation [-10°, 10°], vertical flip, and horizontal flip. The augmented source image ($A(f_s)$) was multiplied by λ_{stone}, which has a random value [0.1, 0.2] to control the intensities of stone pixels and combined with f_t as shown in Eq. (8).

$$f_c = \lambda_{stone} A(f_s) + f_t \tag{8}$$

where f_c is a combined region of a source and target image. Gaussian filter G (window size 3 × 3) was applied on f_c to make the synthetic stone looks more natural. Then, the distance map (w_{dist}) calculated by the Euclidean distance transform was used to calculate the weighted sum between $G(f_c)$ and (f_t) as shown in Eq. (9).

$$f_a = G(f_c) w_{dist} + f_t (1 - w_{dist}) \tag{9}$$

where f_a is a final stone-embedded image, as shown in Fig. 4(right) (3rd row).

2) PRE-PROCESSING AND IMAGE PARTITIONING

All samples were normalized to [-1, 1] before the training process, while ground-truth images were converted to binary images where 1s pixels represent the stone region, and 0s pixels represent the background region. Full images were partitioned based on the stone location map into 3 local images including P_{tl}, P_{tr}, and P_b as described in the first-stage section. Then, all partitioned images were resized to 256 × 256 pixels.

3) LESION-SIZE REWEIGHTING APPROACH TO BALANCE STONE SIZE INEQUALITY

The model trended to miss small stones when training with the traditional dice coefficient loss or binary cross entropy because large lesions overshadow the small ones in loss calculation. Most of the recent loss functions try to solve the data imbalance between classes [25], but ignore imbalance between lesion size in the same class. In our case, abdominal x-ray images usually have multiple stones per image, and some stones can be much larger than small ones. Therefore, we proposed the lesion-size reweighting method, inspired from [26], to reduce the lesion size imbalance problem during training process. The difference is that our inverse weighting method does not include the background component because highly imbalance between the background and stone region makes the weights of stone pixels too high, which reduced segmentation performance in our case. During the training process, we generated the tensor of weights for every batch. We split a tensor of ground-truth into N 2D connected components and the weight for every pixel inside each component (w_j) can be computed by Eq.(10).

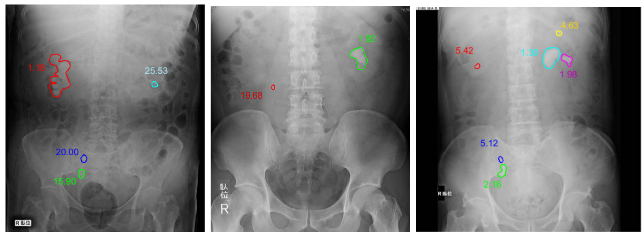


FIGURE 5. Illustration of an inverse weighting result calculated using our modified formulas in Eq.(11). A Weight for every stone is shown near the stone contour.

$$w_j = \begin{cases} 1, & \text{if } j = 0 \\ 1 + \dfrac{\sum_{n=1}^{N} | C_n |}{N \cdot | C_j |}, & \text{otherwise} \end{cases} \tag{10}$$

where C_0 is the background component, and $C_1, \ldots, C_N$ are the connected components [27] of stones in the current batch. This inverse weighting method assigns the higher weights to small stones (Fig.5), which will be used in loss calculation during the training stage.

4) TRAINING METHODOLOGY

In each epoch of training process, all stone-contained images (I_{sc}) and 1/4 stone-free images implemented stone-embedding augmentation ($S(I_{sf})$) were used for

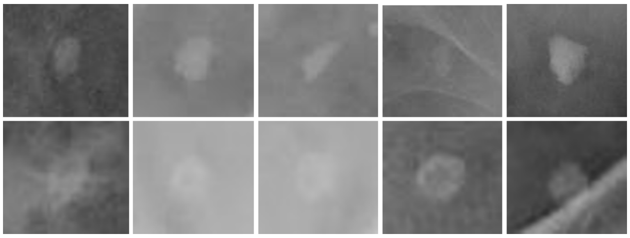


FIGURE 6. Illustration of a comparison between bladder stones (1st row) and phleboliths (2nd row) from our dataset.

training the U-Net model. We used focal Tversky loss (FTL) applied with the proposed inverse weight map (iw.) to overcome these challenges. Focal Tversky loss is the generalization of the Dice loss (DL) balancing importance between FN and FP by α and β, respectively [28]. Furthermore, it also has γ for controlling between easy and hard training samples [29]. We used $\alpha = 0.7$, $\beta = 0.3$, and $\gamma = 2.0$ in all experiments. The calculation of TI_{iw} and FTL_{iw} is defined as Eqs. (11) and (12), respectively.

$$TI_{iw} = \frac{\sum_{i=1}^{N} w_i p_{1i} g_{1i}}{\sum_{i=1}^{N} w_i p_{1i} g_{1i} + \alpha \sum_{i=1}^{N} w_i p_{0i} g_{1i} + \beta \sum_{i=1}^{N} w_i p_{1i} g_{0i}}, \quad (11)$$

$$FTL_{iw} = (1 - TI_{iw})^{1/\gamma} \quad (12)$$

where p_{1i} and p_{0i} are the probability of pixel i being a stone and non-stone, respectively. g_{1i} is 1 for a stone pixel and 0 for a non-stone pixel, and g_{0i} vice versa. w_i is the inverse weight of pixel i as described in previous section. Total number of pixels in a current batch is denoted by N.

We trained the network from scratch for 150 epochs with a batch size of 12 images, and used the Adam optimizer to minimize FTL_{iw} with an initial learning rate of 10^{-3}. Whenever validation loss has not decreased by at least 10^{-4} for 10 epochs during training, the learning rate is divided by two and the minimum learning rate is set to 5×10^{-4}.

5) POST-PROCESSING

Calcifications of tiny veins or phleboliths, as shown in Fig. 6 (bottom), are prevalent in bladder region and can be difficult even for an expert to identify from urinary stones in this location (top). Several studies have reported that urinary stones and phleboliths present different morphological structures and characteristics, however, the classification is still challenging especially for the x-ray modality [30], [31]. In post-processing, we also proposed the detection of false bladder stone by training the classification model to distinguish between bladder stones and phleboliths. We manually cropped 150 images of the bladder stone and phleboliths as well as the paired stone masks. The pre-trained VGG16 network was fine-tuned only fully-connected layers with the concatenation of the cropped image and stone mask input from our dataset using Focal binary cross entropy loss for 150 epochs. Then, we used this trained model to detect and remove the false-positive lesions from the 2nd stage network in the bladder partition.

Lastly, the output images from 2nd stage network were binarized using a 0.5 threshold value, then multiplied with the corresponding stone location maps from 1st stage network to remove the false predicted lesions outside of the urinary organ region.

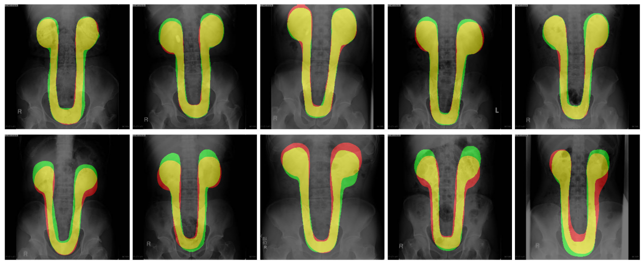


FIGURE 7. Illustration of stone location map results from the 1st stage U-NET; plain x-ray images are overlaid with the predicted map and ground-truth map where TP, FP, and FN pixels are shown in yellow, red, and green, respectively. The first row images are the top-five highest F-score results and the second row images are the top-five lowest F-score results.

TABLE 1. Pixel-wise evaluation of the stone location map segmentation measured by recall, precision, and F_1 score (average B1 S.D.).

	Recall Average (±S.D.)	Precision Average (±S.D.)	F_1 score Average (±S.D.)
1st stage U-Net	0.84 (±0.05)	0.90 (±0.04)	0.87 (±0.03)

D. EXPERIMENTAL SETUP AND EVALUATION METHOD

We evaluated urinary stones segmentation performance using five-fold cross-validation. Stone-contained (I_{sc}) samples were divided into 64% training images, 16% validating images, and 20% testing images. Stone-free (I_{sf}) samples were used only in experiments using stone-embedding augmentation. All experiments were conducted using TensorFlow 2.1.0 and the models were trained on an Nvidia GeForce 1080Ti (12GB) GPU.

We used simple pixel-wise metrics including recall, precision, and F-score to evaluate segmentation results like other researches in lesion segmentation tasks. However, this metric has a drawback in the multiple objects task because big object overshadows small ones. Therefore, we also evaluated by the region-wise metrics, measuring the detection performance based on actual stone(s) and predicted stone(s). In every testing image, each connected component of stone-ground truth (G_i) is compared with the predicted stone connected component that overlaps G_i. If the overlap area over the area of G_i is equal or greater than 0.5, the result will be counted as true positive (TP_r). Otherwise, the result will be counted as false negative (FN_r). To compute false positive (FP_r), each predicted connected component (P_j) is compared with the ground truth that overlaps P_j. If the overlap area over the area

TABLE 2. Pixel-wise and Region-wise evaluation of segmentation results (averageB1S.D.%) by different training methods. The highlight cells represent the scores that difference compared with the baseline are statistically significant ($p < 0.05$).

Model	Pixel-wise evaluation				Region-wise evaluation			
	Recall (%)	Precision (%)	F_1 score (%)	F_2 score (%)	Recall (%)	Precision (%)	F_1 score (%)	F_2 score (%)
Baseline	69.94 (±1.17)	62.94 (±2.26)	66.22 (±0.96)	68.40 (±0.65)	61.04 (±1.49)	67.27 (±2.15)	64.00 (±1.05)	62.19 (±1.17)
Baseline + iw.	72.29 (±1.72)	61.07 (±1.69)	66.18 (±0.54)	69.70 (±0.92)	66.79 (±1.59)	61.85 (±1.34)	64.23 (±0.77)	65.74 (±1.12)
Baseline + aug.	71.81 (±0.67)	62.59 (±1.47)	66.87 (±0.85)	69.74 (±1.16)	65.26 (±1.15)	66.17 (±2.04)	65.71 (±1.40)	65.44 (±1.18)
Baseline + aug. + iw.	74.10 (±1.94)	60.77 (±1.24)	66.76 (±0.93)	70.97 (±1.38)	69.10 (±1.88)	61.09 (±1.45)	64.85 (±0.96)	67.33 (±1.36)
Baseline + part.	70.42 (±0.86)	**64.57 (±1.77)**	67.36 (±1.29)	69.16 (±1.01)	64.77 (±0.79)	**72.16 (±3.52)**	68.26 (±1.48)	66.12 (±0.69)
Baseline + part. + iw.	73.45 (±1.15)	61.87 (±0.56)	67.16 (±0.45)	70.80 (±0.80)	68.82 (±0.98)	67.24 (±0.82)	68.02 (±0.60)	68.50 (±0.77)
Baseline + part. + aug.	71.72 (±0.33)	64.00 (±0.63)	67.64(±0.30)	70.03 (±0.21)	66.96 (±0.92)	70.76 (±1.17)	68.81 (±0.83)	67.69 (±0.84)
Proposed	**73.86 (±1.08)**	62.57 (±0.73)	**67.74 (±0.17)**	**71.28 (±0.64)**	**70.36(±1.18)**	67.76 (±1.87)	**69.03 (±1.21)**	**69.82 (±1.08)**

of G_i is less than 0.5, the result will be counted as FP_r. Then, these values were used for computing recall, precision, and F_B score for region-wise metrics, as shown in Eqs. (13), (14), and (15), respectively.

$$Recall = \frac{TP_r}{TP_r + FN_r} \tag{13}$$

$$Precision = \frac{TP_r}{TP_r + FP_r} \tag{14}$$

$$F_B = \frac{(B^2 + 1) \cdot Precision \cdot Recall}{(B^2 \cdot Precision) + Recall} \tag{15}$$

IV. RESULTS AND DISCUSSION

A. OVERALL STONE LOCATION MAP SEGMENTATION RESULTS

Pixel-wise results of the stone location map segmentation (mean ± s.d.) is presented in Table 1. Our first stage U-Net can produce 0.84% recall, 0.90% precision, and 0.87% F_1 score. Examples of stone location map result are displayed in Fig.7. The top-five best results, showing in the first row, demonstrate that these maps can represent the kidneys, ureters, and bladder region and our stone location map generated by the U-Net model corresponds to the characteristics of input abdominal x-ray images. The top-five lowest F-score results, showing in the second row, demonstrate that although the stone location map results are not segmented precisely compared with the ground-truth, the overall results can represent the estimated location of the urinary organs.

B. URINARY STONES SEGMENTATION RESULTS

1) OVERALL RESULTS

We evaluated U-Net model using different proposed methods including partitioned input from two-stage pipeline (part.), stone-embedding augmentation (aug.), and inverse weight maps (iw.). Post-processing method by false bladder stones detection was implemented only for two-stage pipeline experiments. Based on region-wise results in Table 2, all experiments could outperform the baseline in recall and F_2 scores with statistical significance ($p < 0.05$). The model trained with partitioned $I_{sc} + S(I_{sf})$ samples implemented lesion-size reweighting approach (Proposed.) achieved the highest pixel-wise and region-wise F_1 and F_2 score. Although this method produced a low precision, it significantly improved the recall as a trade-off, which outperformed the baseline 2.88 % pixel-wise (68.40 % to 71.28%) and 7.63% region-wise F_2 score (62.19 % to 69.82 %), respectively. In overall results, the baseline method and our proposed method segmented the large stones very well as shown in Fig.8 (1^{st} column); however, our method could improve the segmentation performance in difficult cases, such as small stones or obscure stones located near other anatomical structures. This improvement is demonstrated in the comparison between baseline and our proposed method in Fig.8 (2^{nd}-6^{th} column).

2) FULL VS. PARTITIONED INPUT TYPE

All experiments of the U-Net model trained with partitioned images demonstrated a significant improvement in pixel-wise and region-wise scores when compared with their paired experiments trained with full image inputs. Instead of receiving entire images as inputs, the second stage U-Net in our cascaded U-Net pipeline processed each partition cropped by KUB region maps. This approach can preserve more information, especially in pixels of small stone, which can be lost during the image scaling and downsampling. Furthermore, the usage of KUB region maps derived from the 1^{st} stage U-Net model can alleviate the imbalance problem between stones and background by removing irrelevant pixels outside urinary tract region.

3) EFFECT OF STONE-EMBEDDED TRAINING IMAGES

Our proposed stone-embedding augmentation reduced the need for actual positive samples and utilized normal images to improve the performance of the deep learning model. When compared to those without this augmentation, the experimental results show that this method significantly improved recall and F_2 score. This method increases the number and variety of positive training samples, which is important for training deep learning models to detect urinary stones in unusual shapes, locations, or background properties. Lower precision results, on the other hand, indicated that the model trained with stone-embedded images was increasingly predicting false positives. This increased false prediction is thought to be due to the fact that some training augmented stones may not appear realistic enough.

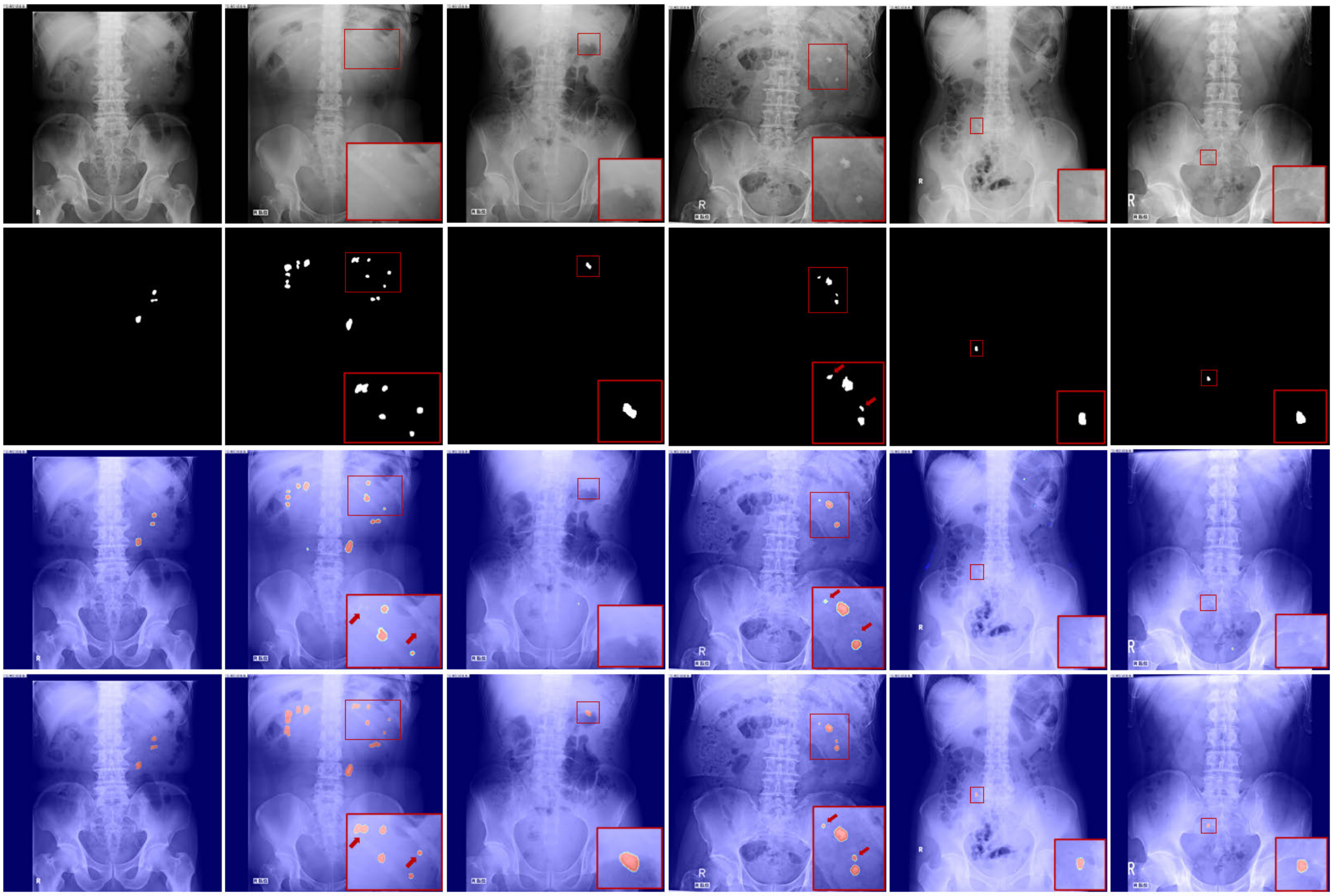


FIGURE 8. Illustration of a comparison between urinary stone segmentation results by a baseline method (3^{rd} row) and those by our proposed method (4^{th} row), displaying predicted stone regions via heatmap visualization.

4) WITH VS. WITHOUT INVERSE WEIGHT MAP (iw.)

The inverse weighting method compensates for the effect of stone size imbalance on loss calculation by multiplying the high weight assigned to small connected components and the low weight assigned to large connected components. Although the precision was decreased when applying this method, the recall was increased as well. These results indicated that the model could detect more stones while also predicting false ones. The results show that using this method with the *FTL* significantly improved the F_2 score when compared to those without it.

5) EVALUATION BASED ON STONE's SIZE AND STONE's REGION

We also investigated the effect of the stone's size on the segmentation performance (region-wise F_1 score). Firstly, all stones in testing data were separated into 3 categories based on their size, including small-sized stones (0-200 pixels), medium-sized stones (201-500 pixels), and large-sized stones (> 500 pixels) from image's resolution of 1,024 × 1,024 pixels. The result in Fig. 9 shows that the region-wise F_1 score was relative to the stone's size, which the larger stones are more detected than the small ones in all experiments. This result also indicated that U-Net model implemented all proposed method (Proposed.) could significantly enhance F_1 score, particularly for small-sized and medium-sized stones, which produced the highest F_1 score in these categories.

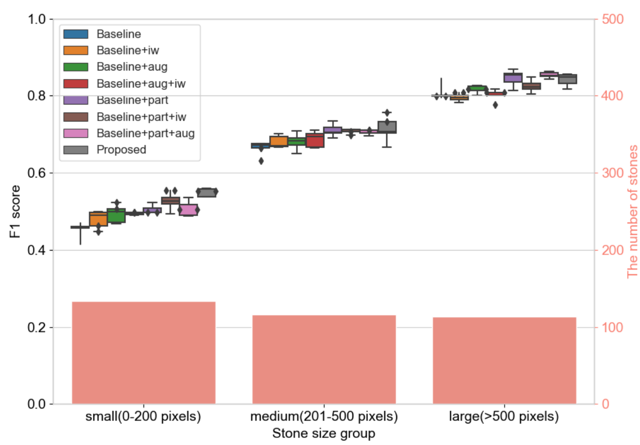


FIGURE 9. The comparison of region-wise F_1 score in different stone size groups.

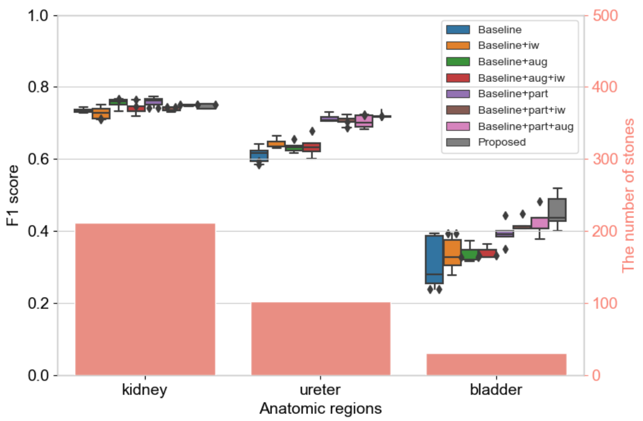


FIGURE 10. The comparison of region-wise F_1 score in different anatomic regions.

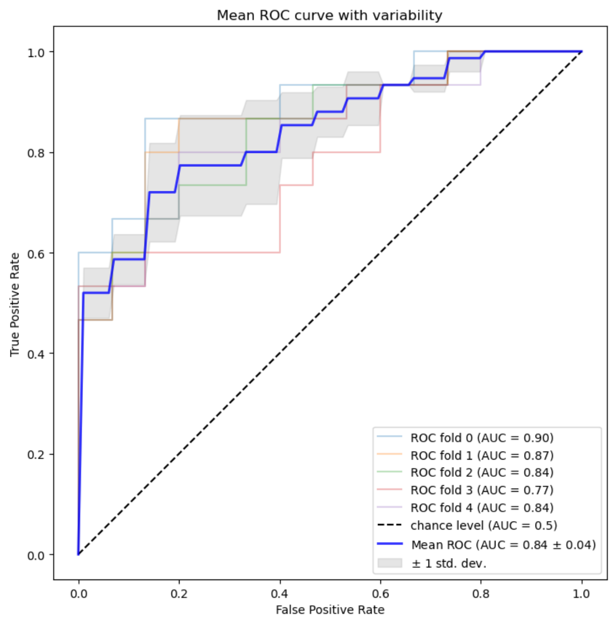


FIGURE 11. Mean ROC curve of bladder stone classification model for 5-fold cross validation.

TABLE 3. Bladder stone classification results measured by recall, precision, and accuracy (average B1 S.D.).

Recall Average (±S.D.)	Precision Average (±S.D.)	Accuracy Average (±S.D.)
0.76 (±0.09)	0.83 (±0.03)	0.80 (±0.05)

For evaluating the stone in different anatomical regions and region-wise F_1 score, we separated all urinary stones in testing data into 3 categories based on their location, including kidney, ureteral, and bladder stones, by using the KUB region maps. The result in Fig.10 shows that stone detection performance was decreased significantly in the bladder region, which has the lowest number of stones. The results in the ureters and bladder region demonstrate that U-Net model implemented all proposed method (Proposed.) produced the highest F_1 score score.

TABLE 4. Comparative stones segmentation results between the proposed method with and without false bladder stone detection measured by region-wise recall, precision, and F_1 score (average B1 S.D.%).

False stones detection	Recall (%) Average (±S.D.)	Precision (%) Average (±S.D.)	F_1 score (%) Average (±S.D.)
x	71.84 (±1.42)	65.55 (±2.41)	68.55 (±1.44)
✓	70.36 (±1.18)	67.76 (±1.87)	69.03 (±1.21)

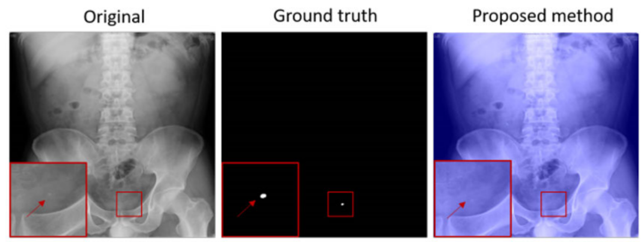


FIGURE 12. False-negative example by our method (the heatmap visualization displays predicted stone regions). Red boxes show enlarged regions containing urinary stone that were missed.

C. FALSE BLADDER STONES DETECTION

The pre-trained VGG16 fine-tuned with our dataset was evaluated using 5-fold cross validation, with 120 cropped-stone images for training and validating, and 30 images for testing. The classification model achieved 0.84 (± 0.04) AUC, as shown in Fig. 11, and 0.76 (± 0.09) recall 0.83 (± 0.03), precision, and 0.80 (± 0.05) accuracy, as shown in Table 3. The detection of false bladder stones was implemented in the post-processing of urinary stones segmentation by 2^{nd} stage network to reduce false positive results. Although, this method reduced the region-wise recall (71.84% to 70.36%), it improved the precision (65.55% to 67.76%) and F_1 score (68.55% to 69.03%), as shown in Table 4.

D. EXPERIMENTAL RESULTS OF U-NET VARIANTS

Furthermore, we compared U-Net-based models implemented with our proposed method (partitioned input + stone-embedding augmentation + inverse weighting map) and the baseline U-Net-based models, which were trained using full images without any proposed method. The U-Net variants that we experimented included U-Net, ResUnet [32], Unet++ [33], Attention Unet [34], MultiResUnet [35], and TransUnet models [36].

Base on the pixel-wise and region-wise evaluation results, as shown in Table 5, Unet++ model with the proposed methods has the highest pixel-wise F-score, while MultiResUnet model has the highest region-wise F-score for both baseline approach and the one employing the proposed methods. Plain U-Net model implementing the proposed methods has better pixel-wise and region-wise F-scores than other baseline U-Net-based variants. Overall, our proposed pipeline can significantly improve F-scores in both pixel-wise and region-wise evaluations as shown by the improvement when compared to the baselines of all Unet-based models.

TABLE 5. Pixel-wise and region-wise evaluation of segmentation results measured by recall, precision, and F_B score (average B1 S.D. %) by Unet-based models with and without our proposed pipeline.

Model	Pixel-wise evaluation			Region-wise evaluation			
	Recall (%)	Precision (%)	F_1 score (%)	Recall (%)	Precision (%)	F_1 score (%)	F_2 score (%)
Unet	69.94 (±1.17)	62.94 (±2.26)	66.22 (±0.96)	61.04 (±1.49)	67.27 (±2.15)	64.00 (±1.05)	62.19 (±1.17)
Unet w/ proposed.	73.86 (±1.08)	62.57 (±0.73)	67.74 (±0.17)	70.36 (±1.18)	67.76 (±1.87)	69.03 (±1.21)	69.82 (±1.08)
ResUnet	68.42 (±1.03)	63.23 (±0.93)	65.72 (±0.36)	60.00 (±1.15)	66.65 (±2.08)	63.15 (±1.05)	61.22 (±0.96)
ResUnet w/ proposed.	70.94 (±1.50)	64.27 (±1.25)	67.43 (±0.64)	68.22 (±1.03)	69.51 (±1.70)	68.86 (±0.41)	68.47 (±0.54)
Attention Unet	67.91 (±2.90)	61.62 (±1.57)	64.59 (±1.80)	61.37 (±1.56)	62.89 (±2.77)	62.12(±1.77)	61.67 (±1.51)
Attention Unet w/ proposed.	**73.90 (±0.67)**	60.90 (±1.12)	66.76 (±0.62)	**72.00 (±1.14)**	62.81 (±2.35)	67.09 (±1.47)	69.95 (±1.09)
Unet++	68.47 (±1.13)	63.35 (±1.62)	65.79 (±0.84)	59.18 (±0.95)	64.90 (±1.80)	61.91 (±1.21)	60.24 (±1.00)
Unet++ w/ proposed.	69.00 (±1.85)	**66.87 (±1.91)**	**67.88 (±0.57)**	65.04 (±1.82)	**71.21 (±2.78)**	67.98 (±1.34)	66.19 (±1.43)
MultiResUnet	72.71 (±1.24)	62.29 (±3.12)	67.05 (±1.49)	65.15 (±0.41)	66.72 (±2.06)	65.93 (±0.98)	65.46 (±0.47)
MultiResUnet w/ proposed.	72.64 (±1.62)	63.57 (±1.18)	67.79 (±0.71)	71.34 (±1.35)	67.32 (±1.78)	**69.27 (±0.43)**	**70.50 (±0.72)**
TransUnet	67.83 (±1.25)	60.94 (±2.83)	64.16 (±1.38)	58.14 (±1.96)	61.05 (±5.20)	59.56 (±2.37)	58.70 (±1.60)
TransUnet w/ proposed.	67.86 (±2.59)	65.02 (±1.28)	66.39 (±1.41)	65.10 (±1.31)	67.50 (±2.42)	66.28 (±1.27)	65.56 (±1.10)

E. LIMITATIONS OF CURRENT WORK AND OUR FUTURE WORK

Although the stone-region evaluation shows that our proposed method can detect the large stones and the stones in kidney region very well, there are some cases that the model cannot detect them. Based on our stone-size and stone-region evaluations, the small stones in the lower ureters or bladder region are the most challenging case that shows the lowest recall results compared with other cases. In this case, as shown in Fig. 12, our model is unable to detect the small stone that is barely visible in the bladder region. In addition, although the detection of false bladder stones could decrease false-positive results, there is room for improvement in the classification accuracy. Our future work will focus on improving the performance of bladder stone classification and stones segmentation, and further implement our proposed method with the larger dataset.

V. CONCLUSION

We proposed a two-stage pipeline for automatically segmenting urinary stones in abdominal x-ray images. The proposed method produced a 71.28% pixel-wise F_2 score and a 69.82% region-wise F_2 score, which were higher than 2.88% and 7.63% produced by the baseline method, respectively. The urinary stones segmentation network in the cascaded framework, processed partitioned images instead of full images, could improve segmentation results by reducing class imbalance problem and processing images at higher resolution. Stone-embedding augmentation was implemented to increase the number and variety of positive training samples during the training process, which was important for improving the performance, especially for stones in rare locations. Our lesion-size reweighting approach used with the focal Tversky loss could significantly improve the detection performance for small stones.

ACKNOWLEDGMENT

The authors would like to thank the medical doctors with the Department of Urology, Tokyo Medical and Dental University (TMDU), for providing the dataset and their insightful suggestions.

REFERENCES

[1] S. R. Khan, M. S. Pearle, W. G. Robertson, G. Gambaro, B. K. Canales, S. Doizi, O. Traxer, and H. G. Tiselius, "Kidney stones," *Nature Rev. Disease Primers*, vol. 2, no. 1, 2016, Art. no. 16008.

[2] T. Alelign and B. Petros, "Kidney stone disease: An update on current concepts," *Adv. Urol.*, vol. 2018, pp. 1–12, Feb. 2018.

[3] W. Brisbane, M. R. Bailey, and M. D. Sorensen, "An overview of kidney stone imaging techniques," *Nature Rev. Urol.*, vol. 13, no. 11, pp. 654–662, Nov. 2016, doi: 10.1038/nrurol.2016.154.

[4] Tamilselvi, "Computer aided diagnosis system for stone detection and early detection of kidney stones," *J. Comput. Sci.*, vol. 7, no. 2, pp. 250–254, Feb. 2011, doi: 10.3844/jcssp.2011.250.254.

[5] K. Viswanath and R. Gunasundari, "Design and analysis performance of kidney stone detection from ultrasound image by level set segmentation and ANN classification," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2014, pp. 407–414.

[6] P. T. Akkasaligar, S. Biradar, and V. Kumbar, "Kidney stone detection in computed tomography images," in *Proc. Int. Conf. Smart Technol. Smart Nation (SmartTechCon)*, Aug. 2017, pp. 353–356.

[7] T. Shah and S. Kadge, "Analysis and identification of renal calculi in computed tomography images," in *Proc. Int. Conf. Nascent Technol. Eng. (ICNTE)*, Jan. 2019, pp. 1–4, doi: 10.1109/ICNTE44896.2019.8945877.

[8] M. Kobayashi, J. Ishioka, Y. Matsuoka, Y. Fukuda, Y. Kohno, K. Kawano, S. Morimoto, R. Muta, M. Fujiwara, N. Kawamura, T. Okuno, S. Yoshida, M. Yokoyama, R. Suda, R. Saiki, K. Suzuki, I. Kumazawa, and Y. Fujii, "Computer-aided diagnosis with a convolutional neural network algorithm for automated detection of urinary tract stones on plain X-ray," *BMC Urol.*, vol. 21, no. 1, pp. 1–10, Dec. 2021, doi: 10.1186/s12894-021-00874-9.

[9] M. A. Mazurowski, M. Buda, A. Saha, and M. R. Bashir, "Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI," *J. Magn. Reson. Imag.*, vol. 49, no. 4, pp. 939–954, 2019, doi: 10.1002/jmri.26534.

[10] X. Li, E. Samei, D. M. Delong, R. P. Jones, A. M. Gaca, C. L. Hollingsworth, C. M. Maxfield, C. W. T. Carrico, and D. P. Frush, "Three-dimensional simulation of lung nodules for paediatric multidetector array CT," *Brit. J. Radiol.*, vol. 82, no. 977, pp. 401–411, May 2009.

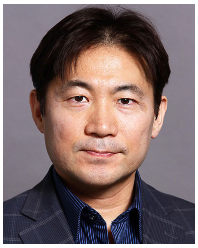
[11] A. Rashidnasab, P. Elangovan, M. Yip, O. Diaz, D. R. Dance, K. C. Young, and K. Wells, "Simulation and assessment of realistic breast lesions using fractal growth models," *Phys. Med. Biol.*, vol. 58, no. 16, 2013, pp. 5613–5627.

[12] M. S. Vaz, Q. Besnehard, and C. Marchessoux, "3D lesion insertion in digital breast tomosynthesis images," *Proc. SPIE*, vol. 7961, Mar. 2011, Art. no. 79615Z.

[13] R. D. Ambrosini and W. G. O'Dell, "Realistic simulated lung nodule dataset for testing CAD detection and sizing," *Proc. SPIE*, vol. 7624, Mar. 2010, Art. no. 76242.

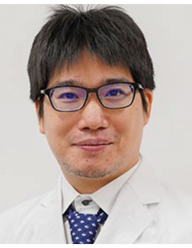
[14] A. P. Peskin and A. A. Dima, "Modeling clinical tumors to create reference data for tumor volume measurement," in *Advances in Visual Computing* (Lecture Notes in Computer Science), vol. 6454. Berlin, Germany: Springer, 2010.
[15] M. T. Madsen, K. S. Berbaum, K. M. Schartz, and R. T. Caldwell, "Improved implementation of the abnormality manipulation software tools," *Proc. SPIE*, vol. 7966, Mar. 2011, Art. no. 7966121.
[16] A. Pezeshk, N. Petrick, W. Chen, and B. Sahiner, "Seamless lesion insertion for data augmentation in CAD training," *IEEE Trans. Med. Imag.*, vol. 36, no. 4, pp. 1005–1015, Apr. 2017, doi: 10.1109/TMI.2016.2640180.
[17] W. Preedanan, I. Kumazawa, T. Kondo, and I. Junichiro, "Urinary stones segmentation in abdominal X-ray images based on U-Net deep learning model and data augmentation techniques," in *Proc. IEEE 5th Int. Conf. Signal Image Process. (ICSIP)*, Oct. 2020, pp. 118–123, doi: 10.1109/ICSIP49896.2020.9339452.
[18] W. Preedanan, K. Suzuki, T. Kondo, M. Kobayashi, H. Tanaka, J. Ishioka, Y. Matsuoka, Y. Fujii, and I. Kumazawa, "Improvement of urinary stone segmentation using GAN-based urinary stones inpainting augmentation," *IEEE Access*, vol. 10, pp. 115131–115142, 2022, doi: 10.1109/ACCESS.2022.3218444.
[19] Y. Zhang, Y. Wang, F. Hou, J. Yang, G. Xiong, J. Tian, and C. Zhong, "Cascaded volumetric convolutional network for kidney tumor segmentation from CT volumes," 2019, *arXiv:1910.02235*.
[20] L. Chaonan, F. Rongda, and Z. Shaohua, "Kidney and kidney tumor segmentation using a two-stage cascade framework," in *International Challenge on Kidney and Kidney Tumor Segmentation*, vol. 13168. Cham, Switzerland: Springer, 2022.
[21] P. Sun, Z. Mo, F. Hu, F. Liu, T. Mo, Y. Zhang, and Z. Chen, "Kidney tumor segmentation based on FR2PAttU-net model," *Frontiers Oncol.*, vol. 12, Mar. 2022, doi: 10.3389/fonc.2022.853281.
[22] G. Wang, W. Li, S. Ourselin, and T. Vercauteren, "Automatic brain tumor segmentation based on cascaded convolutional neural networks with uncertainty estimation," *Frontiers Comput. Neurosci.*, vol. 13, Aug. 2019, doi: 10.3389/fncom.2019.00056.
[23] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2015, pp. 234–241.
[24] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.
[25] C. H Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," 2017, *arXiv:1707.03237*.
[26] B. Shirokikh, "Universal loss reweighting to balance lesion size inequality in 3D medical image segmentation," in *Proc. MICCAI*, vol. 12264. Cham, Switzerland: Springer, 2022, pp. 523–532.
[27] K. Suzuki, I. Horiba, and N. Sugie, "Linear-time connected-component labeling based on sequential local operations," *Comput. Vis. Image Understand.*, vol. 89, no. 1, pp. 1–23, Jan. 2003.
[28] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," 2017, *arXiv:1706.05721*.
[29] N. Abraham and N. M. Khan, "A novel focal Tversky loss function with improved attention U-Net for lesion segmentation," 2018, *arXiv:1810.07842*.
[30] T. D. Perrot, J. Hofmeister, S. Burgermeister, S. P. Martin, G. Feutry, J. Klein, and X. Montet, "Differentiating kidney stones from phleboliths in unenhanced low-dose computed tomography using radiomics and machine learning," *Eur. Radiol.*, vol. 29, no. 9, pp. 4776–4782, Sep. 2019, doi: 10.1007/s00330-019-6004-7.
[31] J. Jendeberg, P. Thunberg, and M. Lidén, "Differentiation of distal ureteral stones and pelvic phleboliths using a convolutional neural network," *Urolithiasis*, vol. 49, no. 1, pp. 41–49, Feb. 2021, doi: 10.1007/s00240-020-01180-z.
[32] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.
[33] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," 2018, *arXiv:1807.10165*.
[34] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," 2018, *arXiv:1807.10165*.
[35] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, Jan. 2020.
[36] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

WONGSAKORN PREEDANAN received the B.Eng. degree in biomedical engineering from Srinakharinwirot University, Thailand, and the M.Eng. degree in information and communications for embedded systems from the Sirindhorn International Institute of Technology, Thammasat University, Thailand. He is currently pursuing the Ph.D. degree in information and communications engineering with the Tokyo Institute of Technology, Japan. His research interests include deep learning, medical image processing, and computer vision. His current research is aimed at developing deep-learning techniques for urinary stone segmentation in radiography, and kidney tumor classification in multi-modal medical images.



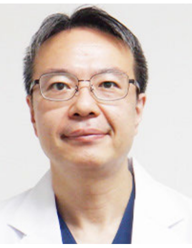
KENJI SUZUKI received the Ph.D. degree from Nagoya University. He was with Hitachi Medical Corp, Aichi Prefectural University, Japan, as a Faculty Member; the Department of Radiology, University of Chicago, as an Assistant Professor; and the Medical Imaging Research Center, Illinois Institute of Technology, as an Associate Professor (Tenured). He is currently a Professor (Tenured) and the Founding Director of the Biomedical Artificial Intelligence Research Unit, Tokyo Institute of Technology, Japan. He has published more than 350 articles (including, 116 peer-reviewed journal articles). He has been actively researching deep learning in medical imaging and AI-aided diagnosis for the past 25 years. He has received 23 awards, including three Best Paper Awards in leading journals. He serves as an Editor for 34 leading international journals, including *Pattern Recognition*.

TOSHIAKI KONDO received the B.Eng. degree in mechanical engineering and the M.Eng. degree in information processing from the Tokyo Institute of Technology, Japan, the M.Eng. degree in image processing from The University of Sydney, Australia, and the Ph.D. degree in image processing from the National University of Singapore, Singapore. He was a Research Engineer with Canon Inc., Japan. Currently, he is an Associate Professor with the School of Information, Computer, and Communication Technology (ICT), Sirindhorn International Institute of Technology, Thammasat University, Thailand. His research interests include digital image processing, such as feature detection and image segmentation, computer vision, such as depth estimation and motion estimation, and pattern recognition.



MASAKI KOBAYASHI received the graduate degree from Tokyo Medical and Dental University, Tokyo, Japan. He joined the Urology Division, Tokyo Medical and Dental University. His research interests include urolithiasis and oncology.

HAJIME TANAKA received the graduate degree from Tokyo Medical and Dental University, Tokyo, Japan. He is an Assistant Professor with Tokyo Medical and Dental University. His research interests include urolithiasis and oncology.



JUNICHIRO ISHIOKA received the Ph.D. degree in medicine from Tokyo Medical and Dental University, Tokyo, Japan. He is a Lecturer with the Division of Public Health, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University. His research interests include computer-aided diagnosis in urology, diagnosis of urothelial cancer, and diagnosis of renal tumors by a multi-modal deep learning model.

YOH MATSUOKA received the Ph.D. degree in medicine from the Graduate School, Tokyo Medical and Dental University, Japan. He is qualified as a medical Doctor and certified as a Urological Specialist and a General Clinical Oncologist in Japan. His current research interests include the development of image-guided, individualized diagnostic procedures and treatments for urological cancer, including the investigation of prostate MRI and pathology to provide less-invasive focal therapy using brachytherapy and other modalities.

YASUHISA FUJII received the M.D. and Ph.D. degrees from the School of Medicine, Tokyo Medical and Dental University. Currently, he is a Professor with the Department of Urology, Graduate School, Tokyo Medical and Dental University. His research interest includes computer-aided diagnosis in medical imaging.

ITSUO KUMAZAWA (Member, IEEE) received the B.Eng. degree in electrical and electronic engineering and the M.Eng. and Ph.D. degrees in computer science from the Tokyo Institute of Technology. Currently, he is a Professor with the Institute of Innovative Research, Tokyo Institute of Technology. He has published a number of articles and books on user interfaces, image processing, neural computing, and pattern recognition. He received grants from JSPS and JST, and a number of private funds. His research interests include pattern recognition, image processing, deep learning, and user interface. He is a member of IEICE, IPSJ, and ITE. He received awards from these academic societies at the IEEE Virtual Reality Conference, in 2013.