IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 October 2022, accepted 24 October 2022, date of publication 31 October 2022, date of current version 8 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3218444

METHODS

Improvement of Urinary Stone Segmentation Using GAN-Based Urinary Stones Inpainting Augmentation

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, Midori-ku, 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, Bunkyo-ku, Tokyo 113-8519, Japan

Corresponding author: Wongsakorn Preedanan (preedanan.w.aa@m.titech.ac.jp)

This work was supported in part by the 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 A urinary stone is a type of abnormality that occurs frequently in the urinary system. An automated segmentation of urinary stones is important for assisting medical doctors in early diagnosis and further treatment. While deep learning techniques are effective for image segmentation, they require a large number of datasets to achieve high accuracy. We proposed a GAN-based augmentation technique for creating synthetic images based on stone and non-stone mask inputs in order to improve the segmentation network's performance by increasing the number and diversity of training data. The synthetic training images were generated from stone-contained images and stone-free images using existing stone ground truth and corresponding stone location maps, respectively. To segment urinary stones from full abdominal x-ray images, we trained the MultiResUnet model using both original stone-contained and our proposed synthetic samples. The proposed method obtained a 69.59% pixel-wise F_1 score and a 68.14% region-wise F_1 score, which achieved an improvement of 2.12% and 2.13%, respectively, over a model trained with only the original stone-contained dataset.

INDEX TERMS GANs, data augmentation, image inpainting, abdominal X-Ray imaging, urinary stone segmentation.

I. INTRODUCTION

Urinary stones are one of the most frequently encountered abnormalities in the urinary system [1]. Symptoms of a urinary stone include lower abdominal pain and gross hematuria; therefore, early diagnosis is necessary to treat patients before the disease becomes severe [2]. Urinary stones can be detected by using a plain x-ray image in a lower body region known as abdominal x-ray imaging, as the majority of stones are calcified, which are visible with this modality. Although abdominal x-ray images are not commonly used for stone detection, they are of less radiation exposure and less expensive than CT scanning, which is the standard medical imaging method on this task [3]. However, detecting urinary stones in a plain x-ray image is a time-consuming process and usually difficult for even an experienced urologists, as stones and other anatomic structures are projected in a 2D image in this modality. Some stones are difficult to detect due to their overlapping to other anatomical structures; and some types of stones, such as irregular ones, are barely visible. Therefore, a computer-aided diagnosis for urinary stone segmentation is

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

demanded to alleviate screening burden and assist medical doctors during diagnosis process.

Deep learning has been widely applied to various medical imaging tasks and has shown significant improvements over traditional feature engineering methods [4]. However, the performance of deep learning is typically dependent on the amount of training data. Medical image datasets are limited compared to other domains due to the high cost of data acquisition, privacy restrictions, and difficulties associated with image labeling, which require experts. Additionally, class imbalance is a prevalent problem in medical domains, where normal samples significantly outnumber samples with lesions.

In this work, we proposed an image inpainting framework to generate synthetic training images from stone and non-stone masks. To the best of our knowledge, this is the first study to focus on generating synthetic lesions in 2D radiography images based on the shape and contextual information from mask inputs and the surrounding region. Furthermore, we demonstrated in the experiments that training urinary stones segmentation network with real stone-contained images and additional synthetic images from the proposed inpainting framework can improve the performance of urinary stones segmentation and detection.

II. RELATED WORKS

Although basic data augmentation techniques such as image shifting, scaling, flipping, and rotations are frequently used to increase data diversity during the training stage, they cannot be used to increase diversity of lesion characteristics and locations. Accordingly, many investigators have proposed various methods for creating new positive training samples. For example, new lesions are simulated using a mathematical model and then superimposed on existing medical images, as demonstrated in the study in [5] for lung nodules, the one in [6] for mammography, and the one in [7] for digital breast tomosynthesis (DBT). In [8], [9], [10], [11], actual lesions are extracted from real CT scan images and then inserted at new locations in other images using various blending techniques. In our previous work [12], we proposed a method for superimposing a random urinary stone into normal x-ray images during the training stage, by blending the properties of the inserting stone and background. Additionally, our work demonstrated that a model trained with real and synthesized samples could improve the segmentation results.

Recently, generative adversarial networks (GANs) have been successfully used in medical imaging augmentation applications. For examples, a study in [13] used GANs to generate medical images of liver lesions in order to improve lesion classification performance. In skin lesion researches in [14], [15], [16], [17], GANs were used in synthesizing new skin lesion images. In recent studies, GANs were used in image inpainting to synthesize lesions in medical image patches to augment the training data in mammograms [18], and lung nodules in CT images [19], [20], [21], [22]. In these techniques, GANs were trained to fill objects of interests, such as lesions, in a cropped region. Deep learning methods trained on real and synthetic images generated by GANs [23], [24], [25] were shown to improve the performance in classification and segmentation tasks.

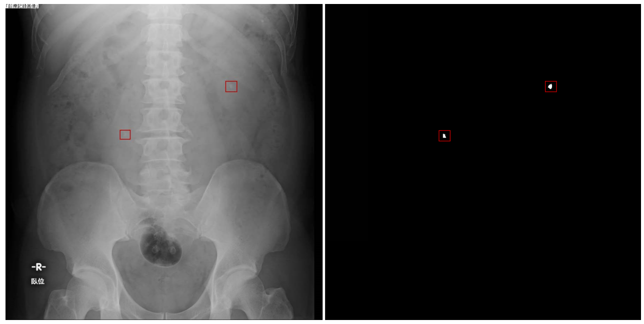


FIGURE 1. Illustration of an abdominal x-ray image with stones (left), along with the corresponding gold standard manual segmentation of the stones (right). The red box represents the cropped region of a urinary stone that was used to generate the dataset in stone inpainting process.

Image inpainting is a task of reconstructing a missing or distorted region in an image. Recently, GANs were used in this application instead of the traditional approaches. Context Encoder (CE) [26] is a framework for training an auto-encoder architecture with adversarial loss and reconstruction loss. The studies in [27], [28] enhance the CE framework by incorporating two discriminator networks: a local discriminator taking the completed region as input and a global discriminator taking the entire image as input. More recently, ip-MedGAN [29] has been developed as an inpainting framework for medical imaging. This method uses cascaded multiple U-Net networks as the generator trained with the combination loss of discriminator networks, reconstruction loss, perception loss, and style loss.

III. URINARY STONE INPAINTING FROM STONE MASK

A. CROPPED STONE AND NON-STONE MASKS

We created a dataset for training an image-to-image translation network by using abdominal x-ray images and their corresponding stone ground truth (Fig.1). For the stone mask dataset, the stone ground-truth images were cropped in a square shape around the stone region for every stone, where the width (w_m) and the top-left coordinates (x_m, y_m) of the stone mask M_s are defined in Eqs.(1) and (2), respectively.

$$w_m = \begin{cases} w_s + 0.2 \cdot w_s, & \text{if } w_s \geq h_s \\ h_s + 0.2 \cdot h_s, & \text{otherwise} \end{cases} \tag{1}$$

where w_s and h_s are the width and height of the urinary stone region, respectively.

$$(x_m, y_m) = \begin{cases} (x_s - 0.1 \cdot w_s, y_s - \dfrac{w_m - h_s}{2}), & \text{if } w_s \geq h_s \\ (x_s - 0.1 \cdot h_s, y_s - \dfrac{w_m - w_s}{2}), & \text{otherwise} \end{cases} \tag{2}$$

where x_s and y_s are top-left coordinates of a urinary stone region.

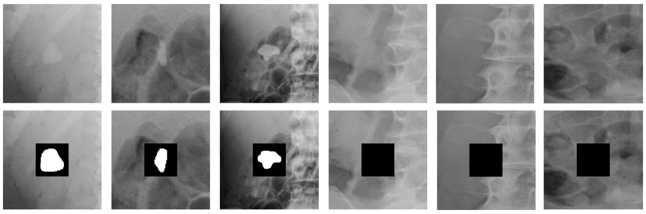


FIGURE 2. Illustration of cropped urinary stone images and their corresponding images with stone masks M_s in the image's center (columns 1-3), as well as cropped non-stone images and the corresponding images with non-stone masks M_{ns} in the image's center (columns 4-6).

For non-stone mask dataset, the top-left coordinates (x_m, y_m) of each non-stone mask M_{ns} were randomly chosen from non-stone region in stone-free (I_{sf}) images, and the width of each non-stone mask w_m was randomly chosen between [10, 50] pixels.

Then, full abdominal x-ray images were cropped as square regions with a width of $3 \cdot w_m$ at M_s or M_{ns}. Fig. 2 illustrates the original cropped stone-region images and their corresponding cropped images that center the binary stone or non-stone mask. We used these pairs, including 1,800 cropped stone-contained (I_{sc}) and 1,800 cropped stone-free (I_{sf}) samples for training and testing process for our image-to-image translation network.

B. NETWORK FOR INPAINTING STONE MASK REGIONS

1) CONDITIONAL INPAINTING GANs

Conditional GAN (cGAN) is a type of GAN that the network is conditioned during training by using some additional information. In this work, we used the image-to-image translation network to generate a missing region by using a stone mask input. This cGAN, learning the mapping from observed image x and random noise z to y, has two components including a generator and a discriminator. The generator G is trained to generate the output images, which are difficult to be distinguished from real images, while the discriminator D is trained to classify between the fake generated images and real images. The adversarial loss of a conditional GAN can be expressed as

$$\mathcal{L}_{cGAN} = E_{x,y}[logD(x, y)] + E_{x,y}[log(1 - D(x, G(x, z)))] \tag{3}$$

where G tries to minimize this objective, while an adversary D tries to maximize it.

2) GENERATOR ARCHITECTURE

The overall structure of this image-to-image translation network is illustrated in Fig.3. We used a stack of two U-Net models, as an inpainting generator, the input to the second network is the coarse inpainting result of the first network. Each model has two paths consisting of a contracting path and an expanding path. The generator takes 128×128 full images with a masked region as input. Each convolutional block consists of two 3×3 convolutional layers with LeakyReLu activation and Batch normalization, followed by a 3×3 convolutional layer with a stride of 2 to downsample the image resolution. At the mid-layers, we used the dilated convolutional layers with dilation rate (η) of 2, 4, 8, and 16. Dilated convolution increases the receptive field, while still using the same number of parameters and computational resources [31]. These layers at the low resolution are important for the image inpainting task because it needs a larger receptive field that can cover the contextual information and missing region. In the expanding path, the transposed convolutional layer was implemented to upsample the image resolution and concatenated with the encoder at the same spatial level. The output layer of each generator uses a 1×1 convolutional layer with Tanh activation.

3) DISCRIMINATOR ARCHITECTURE

An image inpainting task usually utilizes two discriminators with different receptive fields. The global discriminator D_g receives entire generated images and real images as the input, like other GANs do, while the local discriminator D_l receives only the masked region of generated and real images as input. The global discriminator network has the receptive field of 128×128 pixels and consists of 4 convolutional layers(convolution + LeakyReLU + Batch normalization) with 2 strides. By using the wide receptive field input, the network focuses on realistic details in the entire image and ensures that the inpainted region fits the contextual information surrounding the masked region. The local discriminator network consists of 3 convolutional blocks (convolution + LeakyReLU + Batch normalization) with 2 strides, and has a receptive field of 48 *times* 48 pixels cropped from the masked region. By using the smaller input receptive field at the masked region, this network only focuses on realistic details within the inpainted region. The last layer of both networks is a 1×1 convolutional layer with Sigmoid activation, which produce N $\times$ N output patches representing classification scores ('real' or 'fake'). The adversarial loss of cGAN $(\mathcal{L}_{adv})$ used in this work is the average between these two discriminators with different receptive fields, which can be expressed as

$$\mathcal{L}_{adv} = 0.5 \cdot \mathcal{L}_{adv}(G, D_g) + 0.5 \cdot \mathcal{L}_{adv}(G, D_l) \tag{4}$$

4) TRAINING METHODOLOGY

Recently, non-adversarial losses were usually used in an image-to-image translation task as it can obtain better consistent results [30]. In this work, we used a conventional pixel-wise reconstruction loss $(\mathcal{L}_{L1})$ as shown in Eq. (5) to minimize the mean absolute error (MAE) between the target and generated image.

$$\mathcal{L}_{L1} = E_{x,y,z}[\|x - G(y, z)\|_1] \tag{5}$$

We also utilized the content loss to enhance image details of an inpainted image. The feature maps of target $V_j(x)$ and

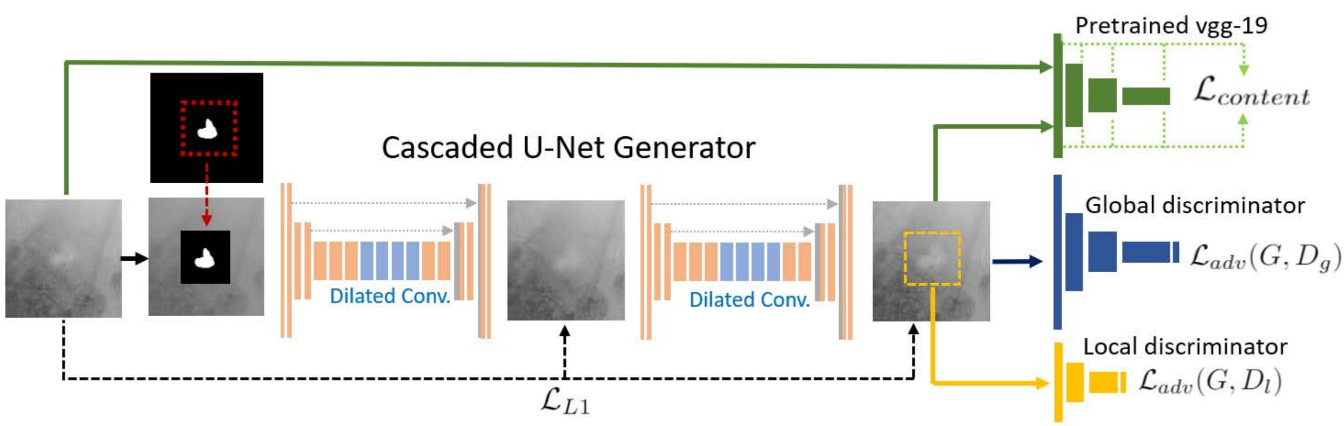


FIGURE 3. Overview of our framework for generative stone inpainting. A cascaded U-Net generator using dilated convolution is trained with reconstruction loss, content loss from the pre-trained VGG19, global adversarial loss, and local adversarial loss.

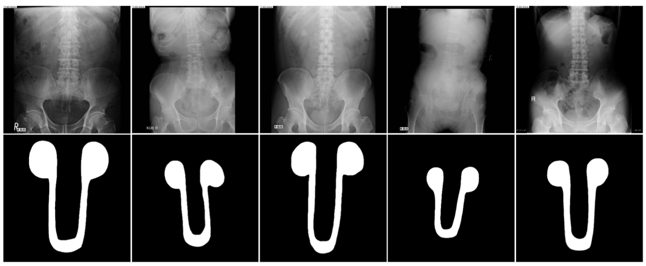


FIGURE 4. Examples of plain abdominal x-ray images (top), and their corresponding stone location maps (bottom).

generated image $V_j(y, z)$ were extracted from j^{th} convolutional layers of the pre-trained VGG-19 network trained on the ImageNet dataset in the classification task. Then, $\mathcal{L}_{content}$ can be computed by

$$\mathcal{L}_{content} = \sum_{j=1}^{4} \frac{1}{h_j w_j d_j} \|V_j(x) - V_j(G(y, z))\|_1 \tag{6}$$

where h_j, w_j, and d_j are the height, width, and depth of the extracted feature maps at the first layer of 1^{st} - 4^{th} blocks of VGG-19 network.

The first part of the cascaded U-Net model was trained with only $\mathcal{L}_{L1}$ loss to generate the coarse result, while the second network was optimized by using the combined objective functions of adversarial loss, L1 reconstruction loss, and content loss expressed as:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{L1} + \lambda_3 \mathcal{L}_{content} \tag{7}$$

where λ1, λ2, and λ3 represent the contributions of adversarial loss, L1 loss, and content loss, respectively. In this work, we used $\lambda 1 = \lambda 3 = 1$ and $\lambda 2 = 50$.

We used the ADAM optimizer [32] with a momentum value of 0.5 and a learning rate of 0.0002 to train the network for 15,000 iterations. The discriminator was trained once for every two iterations of training the generator. The dataset was split into 85% of training samples and 15% of testing samples.

IV. URINARY STONE SEGMENTATION

A. GAN-BASED STONE INPAINTING AUGMENTATION

1) STONE LOCATION MAP

According to medical domain knowledge, urinary stones are formed in kidneys and excreted via the ureters and bladder. Therefore, they are found only in these urinary organs. In this task, we created a map representing approximate locations of urinary stones in the urinary organs based on clinical data. The stone location maps were created by analyzing the characteristics of original full plain abdominal x-ray images of patients, as illustrated in Fig.4. These maps were used for stone synthesis process for stone-free samples (I_{sf}), as described in the following section.

2) SYNTHETIC IMAGES DATASET

The number of positive pixels (in a stone region) in an abdominal x-ray image is extremely small compared to that of negative pixels (in a non-stone region). The ratio of the stone to the non-stone area can be less than 0.1%. In this stage, we used the proposed urinary stone inpainting method described in the previous section to increase the number of positive data. The framework of image augmentation for stone-contained (I_{sc}) and stone-free (I_{sf}) training samples is shown in Fig.5.

For each real stone-free image (I_{sf}), 1 to 3 new target location(s) (x_t, y_t) were randomly selected from the corresponding stone location map to synthesize new stone(s) in the non-stone region. A cropped stone mask (M_s) was randomly selected from the cropped stone mask dataset and augmenting using image rotation [-10 °, $+10$ °], vertical flipping, and horizontal flipping to increase the diversity of the new stone's characteristics. The augmented stone mask M_s was then placed in the center of a selected location (x_t, y_t), and a full image (I_{sf}/I_{sc}) was cropped in a square shape around the placed stone mask M_s with a $3 \cdot w_m$ width to include the context region surrounding the stone mask, similar to the training data for the stone inpainting task. For each real

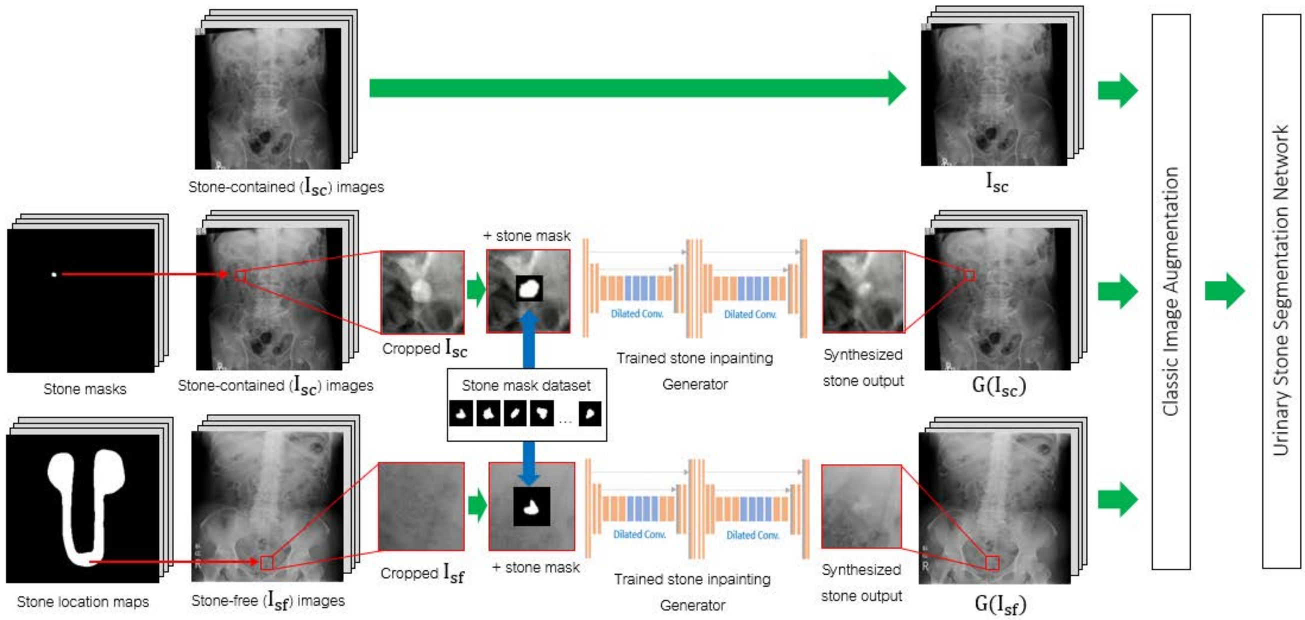


FIGURE 5. Proposed framework for image augmentation including GAN-based augmentation and classic augmentation techniques for urinary stone segmentation.

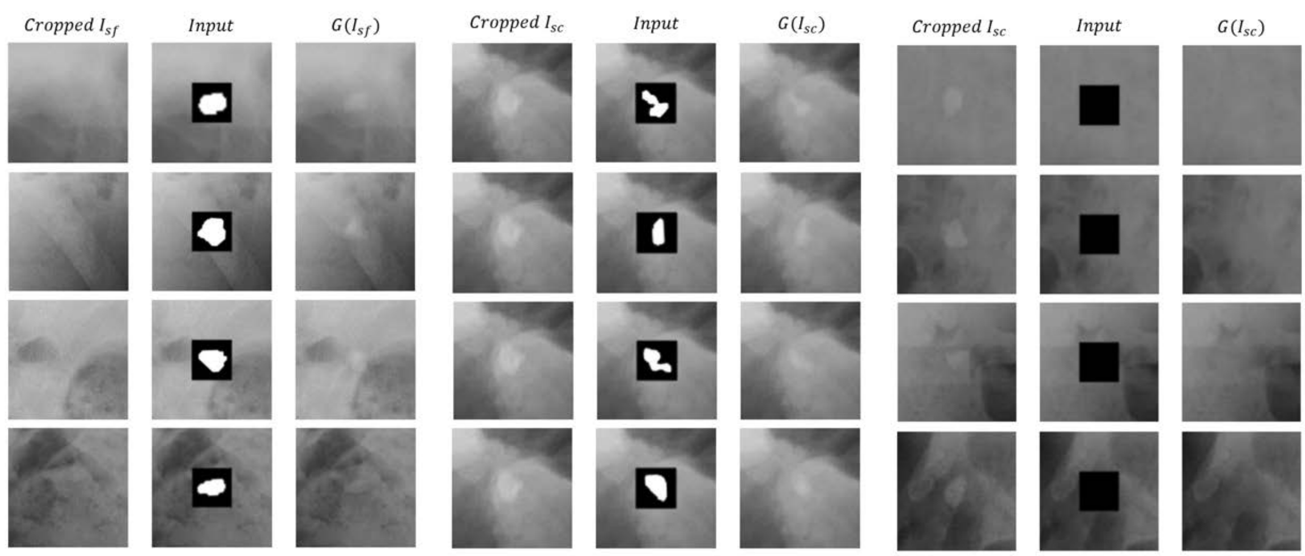


FIGURE 6. Illustrations in columns 1-3 show original cropped I_{sf} images, cropped I_{sf} with random stone masks, and $G(I_{sf})$ results from the stone-free augmentation. Illustrations of original cropped I_{sc} images, cropped I_{sc} with masks, and $G(I_{sc})$ results from the stone-synthesized and stone-removed augmentation are shown in columns 4-6, and 7-9, respectively.

stone-contained image (I_{sc}), the center coordinate of each stone (x_t, y_t) was randomly chosen to be replaced with either the stone mask to synthesize a new stone, or non-stone mask to remove the stone when there are multiple stones.

The input images were resized to 128×128 pixels and processed by the trained inpainting generator to generate a stone region based on the context pixels surrounding the missing region and an input stone or non-stone mask as illustrated in Fig. 6 (columns 1-3) for stone-free images, Fig. 6 (columns 4-6) for stone-contained images with stone mask inputs, and Fig. 6 (columns 7-9) for stone-contained images with non-stone mask inputs.

The cropped region in the full image was then replaced with the inpainted result, and the stone mask was placed at the same location in the full ground-truth image. This method was used to generate 10 additional samples for each I_{sc} and I_{sf}, as additional training samples for the segmentation network.

TABLE 1. Summary of our abdominal X-ray database for urinary stones segmentation.

	Real	Synthetic		Total
	I_{sc}	$G(I_{sc})$	$G(I_{sf})$	
Train	740	7400	7400	15540
Train per epoch	740	370	370	1480
Validate	185	-	-	185
Test	234	-	-	234

B. URINARY STONES SEGMENTATION NETWORK

The training images for the stone segmentation network were a combination of original full stone-contained images (I_{sc}), stone-synthesized stone-contained images ($G(I_{sc})$), and stone-synthesized stone-free images ($G(I_{sf})$). All training images were resized to 256×256 pixels and normalized to zero mean and unit variance. Then, during the training stage, all training images were randomly rotated [−5, 5] and horizontally flipped.

In this task, we used the MultiResUnet model [33] which is one of the state-of-the-art architecture that was designed to improve the classical U-Net architecture, and successfully used in medical image segmentation. It substitutes a MultiRes-Block for each convolution block in the original U-net model at each level. This block consists of three cascaded 3 × 3 convolutional layers interconnected together to extract various scales of spatial features. Then, a 1 × 1 convolution was added as a residual connection from the input to the output of the MultiRes-Block, in order to append the spatial information. Additionally, It replaces skip connections between encoder-decoder paired layers with a ResPaths block, which consists of 3 × 3 and 1 × 1 convolutional filters. The architecture of the MultiResUnet is illustrated in Fig. 7.

The model was optimized using the focal Tversky loss (FTL), a generalization of Dice loss (DL), which balances the contribution between FN and FP by α and β, respectively. Furthermore, it also has γ value for controlling non-linearity of Tversky index (TI) [34]. When $\gamma > 1$, this loss non-linearly focuses more on small TI samples, and suppresses the contribution of high TI samples to the loss function. TI and FTL are calculated as Eqs. (8) and (9), respectively.

$$TI = \frac{\sum_{i=1}^{N} p_{1i}g_{1i}}{\sum_{i=1}^{N} p_{1i}g_{1i} + \alpha \sum_{i=1}^{N} p_{0i}g_{1i} + \beta \sum_{i=1}^{N} p_{1i}g_{0i}}, \quad (8)$$

$$FTL = (1 - TI)^{1/\gamma} \quad (9)$$

where p_{1i} represents the probability that pixel i is a stone and p_{0i} represents the probability that pixel i is not a stone. While g_{1i} is 1 for stone pixels and 0 for non-stone pixels, and g_{0i} is the opposite. N denotes the total number of pixels in the current batch. This study used $\alpha = 0.7$, $\beta = 0.3$ to bias the model toward FN over FP values, and used $\gamma = 2.0$ to focus more on less accurate predictions.

In each epoch during the training stage, 5% of $G(I_{sc})$ and $G(I_{sf})$ datasets were randomly selected and combined with all I_{sc} training samples to train the network. The network was trained from scratch and used the Adam optimizer [32] to minimize FTL with an initial learning rate of 10^{-3}. During training, whenever validation loss did not decrease by at least 10^{-4} over 10 epochs, the learning rate was divided by 2, with the minimum learning rate set to 5×10^{-4}. For all experiments, the model was trained for 150 epochs with a batch size of 16 images.

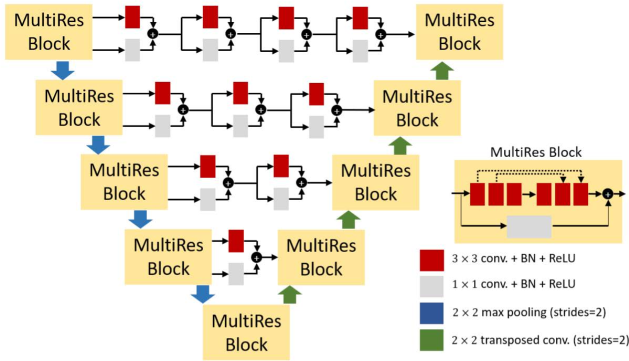


FIGURE 7. MultiResUnet architecture for urinary stones segmentation.

C. EXPERIMENTATION AND EVALUATION METHODS

For the urinary stone segmentation experiment, we used full abdominal x-ray images consisting of 1,159 I_{sc} and 740 I_{sf}. For each I_{sc}, experienced urology doctors manually drew the ground-truth masks of urinary stones. We used 5-fold cross-validation to evaluate segmentation performance. In each validation experiment, I_{sc} samples were divided into 64% training images, 16% validating images, and 20% testing images. $G(I_{sc})$ and $G(I_{sf})$ datasets were used only as additional training samples for the network. All dataset for urinary stones segmentation are summarized in Table 1. The experiments were conducted using TensorFlow 2.5.0 and all models were trained on an Nvidia GeForce 1080Ti (12GB) GPU.

The segmentation results were evaluated using pixel-level metrics such as recall, precision, and F_1 score. Although the conventional pixel-wise evaluation has been used in a wide variety of segmentation tasks, it has a disadvantage in the detection of multiple lesion because large lesions obscure the small ones. Therefore, we also evaluated the results using region-wise metrics, assessing the detection performance based on the ground-truth stones and predicted stones.

Each connected component [35] of stone-ground truth (G_i) was compared to the predicted stone connected component P that overlaps G_i in each testing image. The total number of region-wise true positives (TP_r), and false negatives (FN_r) can be defined in Eqs. (10), and (11), respectively.

$$TP_r = \sum_{i=1}^{N} G_i[\frac{G_i \cap P}{G_i} >= 0.5] \quad (10)$$

$$FN_r = \sum_{i=1}^{N} G_i[\frac{G_i \cap P}{G_i} < 0.5] \quad (11)$$

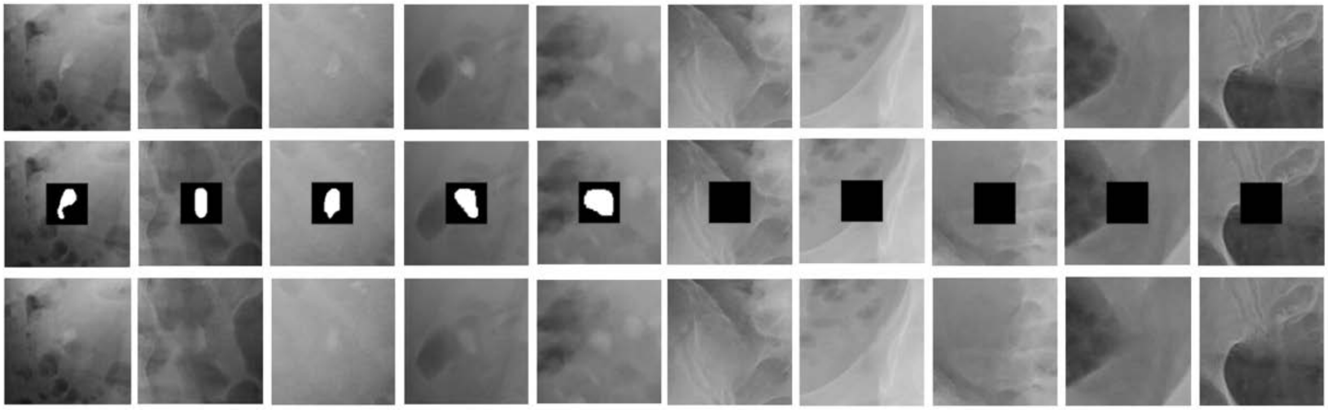


FIGURE 8. Illustration of the original cropped stone region images (1^{st} row), input images for the stone inpainting network(2^{nd} row), and synthesized urinary stone results generated by the stone inpainting network (3^{rd} row).

TABLE 2. Image quality assessment of our inpainted stone and non-stone results.

IQA methods	Testing samples		Average
	Stone mask	Non-stone mask	
MSE	0.00009	0.00007	0.00008
PSNR	42.54418	43.32517	42.93468
SSIM	0.99318	0.99342	0.99330

where the stone ground-truth have N connected components in total.

To calculate false positives (FP_r), each predicted connected component (P_j) was compared with the ground truth that overlaps P_j. Then, FP_r can be defined as Eq. (12).

$$FP_r = \sum_{j=1}^{M} P_j[\frac{P_j \cap G}{P_j} < 0.5] \quad (12)$$

where the predicted stones have M connected components in total.

Then, TP_r, FN_r, and FP_r were used to compute region-wise recall, precision, and F_B score, as shown in Eqs. (13), (14), and (15), respectively. By using the region-wise evaluation metric, the size of the lesion has no effect on these scores. Apart from frequently used F_1 score or dice coefficient, we also reported F_2 score results for region-wise evaluation. In our case, some false positive (FP_r) results are acceptable because all predictions must be confirmed by medical doctors in real-world clinical use. F_2 score, which weights FN_r more than FP_r, is also another suitable metric for our work, as we focused on the detection of urinary stones, and some increased false positives as a trade-off were acceptable.

$$Recall = \frac{TP_r}{TP_r + FN_r} \quad (13)$$

$$Precision = \frac{TP_r}{TP_r + FP_r} \quad (14)$$

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

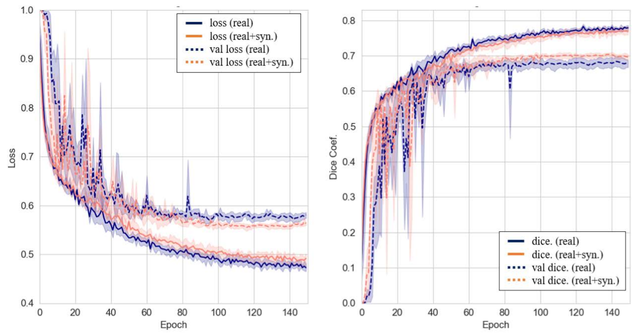


FIGURE 9. Comparisons of training and validation losses (left) and dice coefficients (right) in 5-fold cross validation for the MultiResUnet model trained with different training data.

V. RESULTS AND DISCUSSION

A. IMAGE INPAINTING RESULTS

We evaluated the quality of inpainted images using full-reference image quality assessment (FR-IQA) methods including MSE, PSNR, and SSIM [36], as shown in Table 2. Fig. 8 illustrates the results of an inpainting network implemented for testing samples. The input images in the stone region were trained to generate both a stone region and its surrounding region in the missing region as illustrated in Fig. 8 (columns 1-5), whereas the input images in non-stone regions were trained to fill the missing regions as illustrated in Fig. 8 (columns 6-10).

B. PIXEL-WISE AND REGION-WISE URINARY STONES SEGMENTATION RESULTS

In this experiment, we compared the pixel-wise and region-wise segmentation results of the MultiresUnet model trained with different training data, namely, real stone-contained (I_{sc}), real stone-free (I_{sf}), synthetic stone-contained ($G(I_{sc})$), and synthetic stone-free ($G(I_{sf})$). The model trained with only I_{sc} was selected as the baseline because its pixel-wise and region-wise F score was superior to that of the model trained with both I_{sc} and I_{sf}. The results in Table 3. demonstrate

TABLE 3. Pixel-wise and region-wise evaluation of segmentation results measured by recall, precision, and F_B score (average ± S.D.%) of the MultiResUnet model trained with different training data.

Training data		Pixel-wise evaluation			Region-wise evaluation			
Real	Synthetic	Recall (%)	Precision (%)	F1 score (%)	Recall (%)	Precision (%)	F1 score (%)	F2 score (%)
I_{sc}	-	72.05 (±2.01)	63.05 (±1.76)	67.47 (±0.54)	64.11 (±1.92)	68.02 (±3.36)	66.01 (±1.09)	64.86 (±1.19)
$I_{sc} + I_{sf}$	-	71.29 (±0.59)	63.46 (±2.60)	67.13 (±1.67)	61.99 (±1.31)	66.40 (±3.35)	64.12 (±1.90)	62.82 (±1.38)
I_{sc}	$G(I_{sc})$	72.18 (±1.14)	65.52 (±1.04)	68.68 (±0.60)	65.04 (±0.81)	68.93 (±0.62)	66.93 (±0.36)	65.73 (±0.61)
I_{sc}	$G(I_{sf})$	72.07 (±0.71)	66.07 (±0.38)	68.97 (±0.34)	65.42 (±1.33)	**70.57 (±0.93)**	67.90 (±0.82)	66.39 (±1.10)
I_{sc}	$G(I_{sc})$+$G(I_{sf})$	**72.84 (±1.65)**	**66.65 (±0.97)**	**69.59 (±0.45)**	**66.74 (±1.86)**	69.60 (±1.96)	**68.14 (±1.12)**	**67.29 (±1.45)**

TABLE 4. Pixel-wise and region-wise evaluation of segmentation results measured by recall, precision, and F_B score (average ± S.D. %) by state-of-the-art Unet-based models trained with different training data.

Model	Training data		Pixel-wise evaluation			Region-wise evaluation			
	Real	Syn.	Recall (%)	Precision (%)	F1 score (%)	Recall (%)	Precision (%)	F1 score (%)	F2 score (%)
U-Net	✓	-	71.13 (±1.95)	64.31 (±1.57)	67.51 (±0.43)	62.68 (±1.00)	68.83 (±0.82)	65.62 (±0.50)	63.83 (±0.77)
U-Net	✓	✓	71.28 (±1.08)	66.6 (±0.88)	68.86 (±0.73)	64.77 (±1.80)	69.61 (±1.64)	67.10 (±1.21)	65.68 (±1.50)
ResUnet	✓	-	68.13 (±2.37)	66.37 (±1.16)	67.21 (±1.11)	60.05 (±2.69)	71.08 (±1.35)	65.10 (±1.27)	61.98 (±2.16)
ResUnet	✓	✓	68.40 (±1.02)	68.02 (±1.18)	68.20 (±0.73)	61.21 (±1.00)	70.97 (±2.20)	65.73 (±0.93)	62.94 (±0.93)
Unet++	✓	-	66.86 (±1.20)	67.19 (±1.67)	67.02 (±1.05)	58.79 (±1.48)	**71.63 (±2.94)**	64.58 (±1.87)	60.98 (±1.57)
Unet++	✓	✓	68.02 (±1.48)	**68.74 (±1.58)**	68.35 (±0.09)	61.64 (±1.08)	70.05 (±2.36)	65.58 (±0.81)	63.16 (±0.73)
Attention Unet	✓	-	70.57 (±0.96)	63.20 (±1.12)	66.67 (±0.48)	62.85 (±0.46)	67.91 (±1.33)	65.28 (±0.47)	63.80 (±0.27)
Attention Unet	✓	✓	71.29 (±1.27)	64.24 (±0.80)	67.58 (±0.73)	65.32 (±1.09)	67.65 (±1.13)	66.46 (±0.20)	65.77 (±0.69)
MultiResUnet	✓	-	72.05 (±2.01)	66.07 (±1.76)	67.47 (±0.54)	64.11 (±1.92)	68.02 (±3.36)	66.01 (±1.09)	64.86 (±1.19)
MultiResUnet	✓	✓	**72.84 (±1.65)**	66.65 (±0.97)	**69.59 (±0.45)**	**66.74 (±1.86)**	69.60 (±1.96)	**68.14 (±1.12)**	**67.29 (±1.45)**
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	✓	✓	64.79 (±0.54)	69.02 (±1.32)	66.83 (±0.48)	57.21 (±0.68)	68.96 (±1.29)	62.53 (±0.68)	59.22 (±0.62)
UTNet	✓	-	65.21 (±3.23)	60.10 (±1.93)	62.49 (±1.29)	58.48 (±3.24)	64.26 (±1.69)	61.24 (±2.33)	59.55 (±2.88)
UTNet	✓	✓	66.46 (±1.97)	61.71 (±1.03)	64.00 (±1.44)	62.49 (±1.27)	64.70 (±3.40)	63.58 (±2.24)	62.92 (±1.63)

that our proposed synthetic training data could significantly improve segmentation results when compared to a baseline, and the model trained with I_{sc}, $G(I_{sc})$, and $G(I_{sf})$ could achieve the highest scores in all pixel-wise scores and region-wise recall, region-wise F_1, and region-wise F_2 scores. The proposed method outperformed the baseline 2.12% pixel-wise F_1 score (67.47 % to 69.59 %), and 2.13% region-wise F_1 score (66.01 % to 68.14 %). For region-wise evaluation, these synthetic training samples significantly improved recall scores in all experiments; thus, the improvement is obviously seen in region-wise F_2 score, in which FNs are weighted more than FPs. Fig. 9 shows the 5-fold cross validation training loss and dice coefficient for a baseline (real) and the proposed method (real+syn.), demonstrating that the proposed method's validation loss was lower than a baseline and its validation dice coefficient was also higher than a baseline.

Additionally, we performed statistical analysis on pixel-wise and region-wise F_1 score results using an independent two-sample t-test comparing the baseline method to those trained with real and synthetic training data, as shown in Fig. 10. For pixel-wise evaluation, I_{sc}+$G(I_{sc})$, I_{sc}+$G(I_{sf})$, and I_{sc}+$G(I_{sc})$+$G(I_{sf})$ training data all have a significantly higher F_1 score than the baseline ($p < 0.05$). For region-wise evaluation, MultiResUnet model trained with I_{sc}+$G(I_{sf})$, and I_{sc}+$G(I_{sc})$+$G(I_{sf})$ can improve F_1 score significantly ($p < 0.05$).

Overall, as illustrated in Fig. 11, both the baseline method and our proposed method are capable of detecting and segmenting large stones very well (columns 1-2). The example

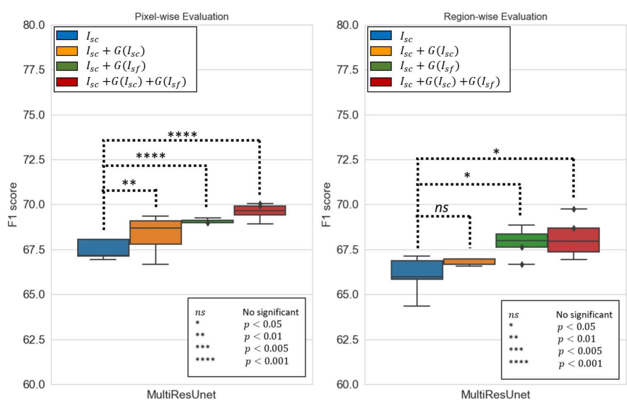


FIGURE 10. Comparison of pixel-wise (left) and region-wise (right) F_1 score of the MultiResUnet model trained with different training data.

results in Fig. 10 (columns 3-6) demonstrate that our proposed method is capable of detecting small stones that were missed by the baseline method. However, there are some cases, particularly small stones as illustrated in Fig. 10 (column 7), where both methods were unable to detect them.

C. COMPARATIVE EXPERIMENT BY STATE-OF-THE-ART UNET-BASED MODELS

In addition, we compared the state-of-the-art Unet-based models trained with only real stone-contained data (I_{sc}) to those trained with both real stone-contained and all synthetic data (I_{sc}+$G(I_{sc})$+$G(I_{sf})$). The Unet-based models used in this experiment are the original U-Net [37], ResUnet [38],

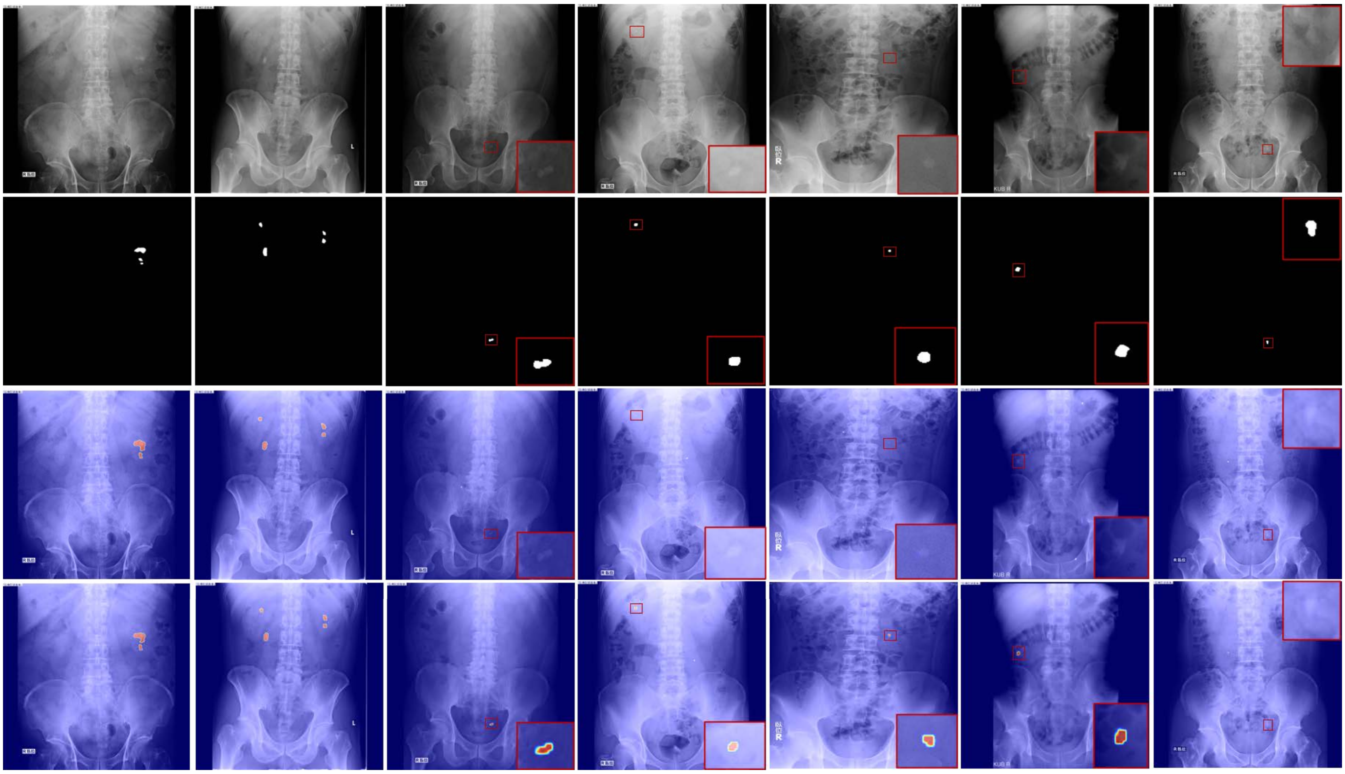


FIGURE 11. Comparisons between urinary stone segmentation results by a baseline MultiResUnet (3^{rd} Row) and the MultiResUnet trained with both real samples and our proposed synthetic samples (4^{th} Row). Red boxes show enlarged regions containing urinary stones.

Unet++ [39], Attention Unet [40], MultiResUnet [33], TransUnet [41], and UTNet [42]. In comparison to other Unet-based models, the MultiResUnet model has the highest recall and F_1 scores for pixel-wise results, and the highest recall, F_1, and F_2 scores for region-wise results. While Unet++ trained on real combined with synthetic samples has the best pixel-wise precision, and the one with only real data has the best region-wise precision. As shown in the pixel-wise and region-wise evaluation results in Table 4, all models trained on real data with additional synthetic training data ($G(I_{sc})$ and $G(I_{sf})$) achieved higher pixel-wise F_1 score, region-wise F_1, and F_2 scores than the baselines that was trained with only real data.

D. STONE SIZE VS. REGION-WISE RECALL

Furthermore, we investigated the effect of the stone size on the region-wise recall. All urinary stones were classified according to their size, including small-sized stones (0-200 pixels), medium-sized stones (201-500 pixels), and large-sized stones (> 500 pixels) based on the image's resolution of 1,024 × 1,024 pixels. The comparison of recalls across different stone size groups in Fig. 12 demonstrates that while all baseline models detected large stones well (recall > 0.8), their performance deteriorated significantly for medium and small stones. The addition of synthetic training samples ($G(I_{sc})$, and $G(I_{sc})$) significantly improved the region-wise recall for all models, particularly for small stones, but had a slight effect on recall scores for medium and large stones.

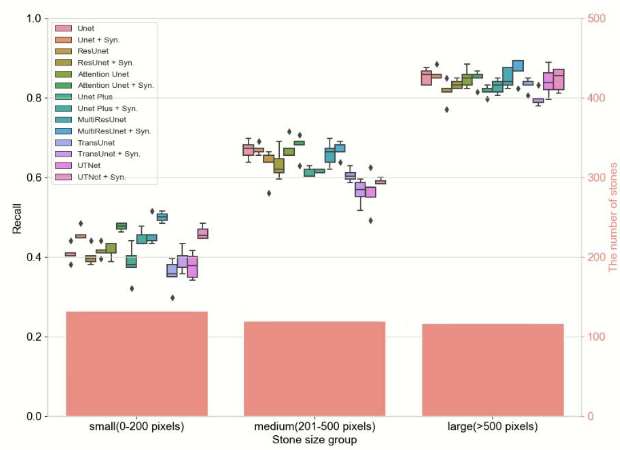


FIGURE 12. Comparison of region-wise recalls of state-of-the-art deep methods trained with and without synthetic training samples in different stone size groups.

This method, by increasing the number of positive training data $G(I_{sc})$ and $G(I_{sf})$, can support the network in learning to segment urinary stones using a wider variety of images. This method augments the number and variety of positive training samples, which is important when training deep learning to detect urinary stones with irregular shapes, locations, or background properties. Although lower region-wise precision in some models means the model is more likely to predict more FPs when trained with synthetic images, the model also detects more TPs as a trade-off, as evidenced by a significant increase in region-wise F_2 score.

This augmentation method is important for medical imaging applications, where the number of positive cases is typically less than the negative cases. This method is important for medical imaging applications in which the number of positive cases is typically lower than the number of negative cases. By utilizing existing medical images of healthy samples, this method can also be used to reduce the number of actual positive samples required and also improve the segmentation performance of deep learning models.

VI. CONCLUSION

We proposed a GAN-based inpainting augmentation technique for generating the synthetic images based on the input masks and their surrounding regions. The proposed inpainting model was used to generate the synthetic training samples from original stone-contained images and stone-free images to increase the number and variety of positive training samples for the lesion segmentation model. The experimental results indicated that our proposed method was able to achieve higher pixel-wise and region-wise F-score than the baseline methods. In overall, this method could significantly improve the segmentation performance, especially for small stones and stones located in less common locations.

ACKNOWLEDGMENT

The authors are thankful to medical doctors in Department of Urology at 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, p. 16008, 2016.

[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] 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, 2018, doi: 10.1002/jmri.26534.

[5] 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.

[6] 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.

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

[8] R. D. Ambrosini and W. G. O'Dell, ''Realistic simulated lung nodule dataset for testing CAD detection and sizing,'' in *Proc. SPIE*, vol. 7624, Mar. 2010, p. 76242.

[9] 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.

[10] M. T. Madsen, K. S. Berbaum, K. M. Schartz, and R. T. Caldwell, ''Improved implementation of the abnormality manipulation software tools,'' in *Proc. SPIE*, Mar. 2011, pp. 7966121–7966127.

[11] 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.

[12] 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.

[13] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, ''GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification,'' *Neurocomputing*, vol. 321, pp. 321–331, Dec. 2018.

[14] C. Baur, S. Albarqouni, and N. Navab, ''Generating highly realistic images of skin lesions with GANs,'' in *OR 2.0, Context-Aware Operating Theaters Computer Assisted Robotic Endoscopy Clinical Image-Based Procedures and Skin Image Analysis*. Cham, Switzerland: Springer, 2018.

[15] A. Bissoto, F. Perez, E. Valle, and S. Avila, ''Skin lesion synthesis with generative adversarial networks,'' in *OR 2.0, Context-Aware Operating Theaters Computer Assisted Robotic Endoscopy Clinical Image-Based Procedures and Skin Image Analysis*. Cham, Switzerland: Springer, 2018, pp. 294–302.

[16] K. Abhishek and G. Hamarneh, ''Mask2Lesion: Mask-constrained adversarial skin lesion image synthesis,'' in *Proc. Int. Workshop Simul. Synth. Med. Imag.* Cham, Switzerland: Springer, 2019, pp. 71–80.

[17] F. Pollastri, F. Bolelli, R. Paredes, and C. Grana, ''Augmenting data with GANs to segment melanoma skin lesions,'' *Multimedia Tools Appl.*, vol. 79, nos. 21–22, pp. 15575–15592, Jun. 2020.

[18] E. Wu, K. Wu, D. Cox, and W. Lotter, ''Conditional infilling GANs for data augmentation in mammogram classification,'' in *Image Analysis for Moving Organ, Breast, and Thoracic Images*. Cham, Switzerland: Springer, 2018, pp. 98–106.

[19] D. Jin, Z. Xu, Y. Tang, A. P. Harrison, and D. J. Mollura, ''CT realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation,'' in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent (MICCAI)*, Granada, Spain, 2018, pp. 732–740.

[20] S. Liu, E. Gibson, S. Grbic, Z. Xu, A. A. A. Setio, J. Yang, B. Georgescu, and D. Comaniciu, ''Decompose to manipulate: Manipulable object synthesis in 3D medical images with structured image decomposition,'' 2018, *arXiv:1812.01737*.

[21] J. Yang, S. Liu, S. Grbic, A. A. A. Setio, Z. Xu, E. Gibson, G. Chabin, B. Georgescu, A. F. Laine, and D. Comaniciu, ''Class-aware adversarial lung nodule synthesis in CT images,'' in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 1348–1352.

[22] Z. Xu, X. Wang, H.-C. Shin, D. Yang, H. Roth, F. Milletari, L. Zhang, and D. Xu, ''Correlation via synthesis: End-to-end nodule image generation and radiogenomic map learning based on generative adversarial network,'' 2019, *arXiv:1907.03728*.

[23] L. Lindner, D. Narnhofer, M. Weber, C. Gsaxner, M. Kolodziej, and J. Egger, ''Using synthetic training data for deep learning-based GBM segmentation,'' in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 6724–6729.

[24] Y. Shin, H. A. Qadir, and I. Balasingham, ''Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance,'' *IEEE Access*, vol. 6, pp. 56007–56017, 2018.

[25] V. Thambawita, P. Salehi, S. A. Sheshkal, S. A. Hicks, H. L. Hammer, S. Parasa, T. D. Lange, P. Halvorsen, and M. A. Riegler, ''SinGAN-Seg: Synthetic training data generation for medical image segmentation,'' *PLoS One*, vol. 17, no. 5, May 2022, Art. no. e0267976, doi: 10.1371/journal.pone.0267976.

[26] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, ''Context encoders: Feature learning by inpainting,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[27] S. Iizuka, E. Simo-Serra, and H. Ishikawa, ''Globally and locally consistent image completion,'' *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, 2017.

[28] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, ''Generative image inpainting with contextual attention,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.

[29] K. Armanious, Y. Mecky, S. Gatidis, and B. Yang, ''Adversarial inpainting of medical image modalities,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3267–3271.

[30] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, ''Image-to-image translation with conditional adversarial networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2016, pp. 5967–5976.

[31] F. Yu and V. Koltun, ''Multi-scale context aggregation by dilated convolutions,'' in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2016, pp. 1–13.

[32] 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.

[33] 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.

[34] N. Abraham and N. M. Khan, ''A novel focal Tversky loss function with improved attention U-Net for lesion segmentation,'' 2018, *arXiv:1810.07842*.

[35] 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.

[36] Z. Wang, E. P. Simoncelli, and A. C. Bovik, ''Multiscale structural similarity for image quality assessment,'' in *Proc. Asilomar Conf. Signals, Syst., Comput. (ASILOMAR)*, Pacific Grove, CA, USA, Nov. 2003, pp. 1398–1402.

[37] O. Ronneberger, P. Fischer, and T. Brox, ''U-Net: Convolutional networks for biomedical image segmentation,'' in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[38] 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.

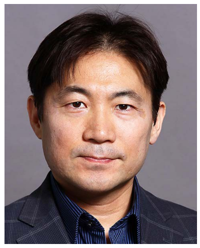
[39] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, ''UNet++: A nested U-Net architecture for medical image segmentation,'' 2018, *arXiv:1807.10165*.

[40] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, ''Attention U-Net: Learning where to look for the pancreas,'' 2018, *arXiv:1804.03999*.

[41] 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*.

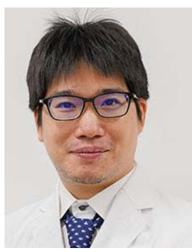
[42] Y. Gao, M. Zhou, and D. Metaxas, ''UTNet: A hybrid transformer architecture for medical image segmentation,'' in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2021, pp. 61–71.

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 interests include developing deep learning techniques for urinary stone segmentation in radiography and kidney tumor classification in multi-modal medical images.

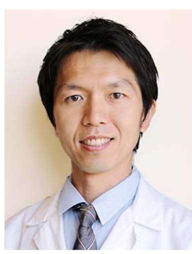


KENJI SUZUKI received the Ph.D. degree from Nagoya University. He was worked at Hitachi Medical Corporation, Aichi Prefectural University, Japan, as a Faculty Member at the Department of Radiology, University of Chicago, as an Assistant Professor at the Medical Imaging Research Center, Illinois Institute of Technology, and as an Associate Professor (Tenured). He is currently a Professor (Tenured) and the Founding Director of Biomedical Artificial Intelligence Research Unit, Tokyo Institute of Technology, Japan. He published more than 350 papers (including 116 peer-reviewed journal papers). He has been actively researching on deep learning in medical imaging and AI-aided diagnosis in the past 25 years. He received 23 awards, including three Best Paper Awards in leading journals. He serves as an Editor of 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 worked as a Research Engineer at Canon Inc., Japan. He is currently 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, and computer vision, such as, depth estimation and motion estimation and pattern recognition.



MASAKI KOBAYASHI received the degree from Tokyo Medical and Dental University, Tokyo, Japan. He joined the urology division of Tokyo Medical and Dental University. His research interests include urolithiasis and oncology.



HAJIME TANAKA received the degree from Tokyo Medical and Dental University, Tokyo, Japan. He is currently an Assistant Professor 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 currently a Lecturer with the Graduate School of Medical and Dental Sciences Medical and Dental Sciences, Division of Public Health, 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 Tokyo Medical and Dental University Graduate School, 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. He is currently a Professor with the Department of Urology, Tokyo Medical and Dental University Graduate School. 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. He is currently a Professor with the Institute of Innovative Research, Tokyo Institute of Technology. He has published a number of papers and books on user interfaces, image processing, neural computing, and pattern recognition. His research interests include pattern recognition, image processing, deep learning, and user interface. He received Grants from JSPS, JST, and a number of private funds. He is a member of IEICE, IPSJ, and ITE. He received awards from these academic societies in the IEEE Virtual Reality Conference, in 2013.