Computers in Biology and Medicine 121 (2020) 103767

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

Computers in Biology and Medicine

An International Journal

Bone segmentation on whole-body CT using convolutional neural network with novel data augmentation techniques

Shunjiro Noguchi [a,*], Mizuho Nishio [a], Masahiro Yakami [a,b], Keita Nakagomi [c], Kaori Togashi [a]

[a] Department of Diagnostic Imaging and Nuclear Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan
[b] Preemptive Medicine and Lifestyle-Related Disease Research Center, Kyoto University Hospital, Kyoto, Japan
[c] Medical Products Technology Development Center, R&D Headquarters, Canon Inc, Tokyo, Japan

ARTICLE INFO

Keywords:
Bone
Segmentation
CT
CNN
U-net
Data augmentation
Mixup
RICAP

ABSTRACT

Background: The purpose of this study was to develop and evaluate an algorithm for bone segmentation on whole-body CT using a convolutional neural network (CNN).
Methods: Bone segmentation was performed using a network based on U-Net architecture. To evaluate its performance and robustness, we prepared three different datasets: (1) an in-house dataset comprising 16,218 slices of CT images from 32 scans in 16 patients; (2) a secondary dataset comprising 12,529 slices of CT images from 20 scans in 20 patients, which were collected from The Cancer Imaging Archive; and (3) a publicly available labelled dataset comprising 270 slices of CT images from 27 scans in 20 patients. To improve the network's performance and robustness, we evaluated the efficacy of three types of data augmentation technique: conventional method, mixup, and random image cropping and patching (RICAP).
Results: The network trained on the in-house dataset achieved a mean Dice coefficient of 0.983 ± 0.005 on cross validation with the in-house dataset, and 0.943 ± 0.007 with the secondary dataset. The network trained on the public dataset achieved a mean Dice coefficient of 0.947 ± 0.013 on 10 randomly generated 15-3-9 splits of the public dataset. These results outperform those reported previously. Regarding augmentation technique, the conventional method, RICAP, and a combination of these were effective.
Conclusions: The CNN-based model achieved accurate bone segmentation on whole-body CT, with generalizability to various scan conditions. Data augmentation techniques enabled construction of an accurate and robust model even with a small dataset.

1. Introduction

Computed tomography (CT) images play a crucial role in evaluation of bony structures and abnormalities. Therefore, accurate segmentation of bone on CT images is important in many medical disciplines. The results of bone segmentation could facilitate bone disease diagnosis and post-treatment assessment, and support planning and image guidance for many treatment modalities including surgery, interventional radiology, and radiation therapy. They also provide stable structural reference for localization of other organs. However, manual segmentation of bone is time consuming and labor intensive. Further development of automated bone segmentation systems is highly desirable.

Thresholding is the most commonly used technique in automatic bone segmentation [1]. Because bone has high density, bone segmentation on CT images could be thought an easier task compared with segmentation of the internal organs; however, there are some specific difficulties to consider. Bones are found in every part of the body, and each has a different shape and density. Even within a single bone, the densities of cortical bone, cancellous bone, and bone marrow differ markedly. In addition, there are very small and irregular shaped structures such as numerous joints throughout the body, paranasal sinus, and temporal bones. Hence it is difficult to provide accurate segmentation labels for all structures at the same time. Furthermore, segmentation errors can easily be produced by the presence of various artificial high-density materials such as contrast media, dental work, and surgical implants.

It is widely known that recent advances in deep-learning-based methods have achieved success in many different fields, including medical image segmentation [2–4]. Several reports have investigated convolutional neural network (CNN)-based architecture for bone

[*] Corresponding author.
E-mail address: noguchi.shunjiro.c95@kyoto-u.jp (S. Noguchi).

https://doi.org/10.1016/j.compbiomed.2020.103767
Received 25 January 2020; Received in revised form 15 April 2020; Accepted 15 April 2020
Available online 20 April 2020

segmentation on CT [5–7]. However, these studies have several limitations, including their evaluation methodologies, low image quality [5, 6], and limited scan coverage [7].

The purpose of this study was to develop a CNN-based model to perform bone segmentation on whole-body CT images, and to evaluate the accuracy and generalizability of the model by testing in three separate datasets. With the aim of further improving the segmentation performance, we adapted and assessed two novel data augmentation methods, *mixup* and *random image cropping and patching (RICAP)*.

2. Materials and methods

This retrospective study was approved by the ethical committee of our institute. For this type of study formal consent was not required.

2.1. Datasets

The current study included the following three datasets for developing and validating the deep learning model. Details of the scan conditions, including scanner manufacturer, slice thickness, scan area and use of contrast materials, are summarized in Table 1.

1. In-house Dataset

The in-house dataset included 16,218 slices of CT images comprising 32 scans that were obtained in 16 patients and performed at our institute. All scans had been performed as follow-up assessment for malignancy (lung cancer, n = 3, breast cancer, n = 2, other, n = 11). Among the 16 patients, 9 patients had known sites of bone metastases. The CT images were acquired by three helical CT scanners (Aquilion 64, Aquilion ONE, Aquilion PRIME; Canon Medical Systems, Otawara, Japan). Slice thickness was 0.5, 1.0, or 5.0 mm, and axial in-plane image resolution was 0.41–0.68 mm. Ground truth labels were established using PLUTO: a software platform for analysis and visualization of medical images [36]. First, the base labels were created with PLUTO's segmentation tools such as thresholding, and then manual correction were performed by seven medical experts using PLUTO's manual segmentation tool. Finally, the labels were verified and modified by a radiologist (18-years' experience in diagnostic imaging) for each scan. Three-fold cross validation was used for training and testing. The split of three-fold was randomly generated and scans from the same patient were not used in both the training and validation datasets.

2. Secondary Dataset

To assess the ability of our model to generalize to other datasets, we tested the model trained with our in-house dataset on a secondary dataset collected from The Cancer Imaging Archive (TCIA) [8–11]. The secondary dataset included 12,529 slices of CT images, comprising 20 scans obtained in 20 patients. Slice thickness was 1 or 1.25 mm, and axial in-plane image resolution was 0.63–0.97 mm. The images were acquired at a single institute in the United States. No information about the CT scanner used for acquisition is publicly available. All examinations were obtained with intravenous and oral contrast material, and patients were scanned from the chest to the pelvis. Ground truth segmentation was performed by a medical expert and verified by a radiologist (6-years' experience in diagnostic imaging) for each scan.

3. Public Dataset

To compare our model with those of previous studies, we trained and validated our network on a publicly available labelled dataset [12]. The dataset comprised 270 slices of CT images from 27 scans, obtained in 20 patients. These images were acquired with a helical CT scanner (Philips, Amsterdam, The Netherlands), with a slice thickness of 5 mm and axial in-plane image resolution of 0.78 mm. Of the 27 examinations, 15 were used for training, 3 for validation, and 9 for testing, as described previously [5]. As the dataset is rather small, the test score naturally depends on the combination of cases upon splitting [13]. We repeated training and testing across 10 randomly generated 15-3-9 splits and calculated the average of 10 test scores as the final score.

2.2. Data augmentation

To prevent overfitting in training our model and to improve the performance and robustness of the network, we utilized the following three types of augmentation algorithm.

1. Conventional Method

Conventional augmentation methods include rotation (−10°–10°), zooming (60%–140%), horizontal flip, as well as shear transformation (shear angle = 5°). These are typical operations used in medical image segmentation [5].

2. Mixup

Mixup is a recently proposed augmentation technique that generates new training samples from linear combination of existing images and their labels: (x_i, y_i) and (x_j, y_j) [14]. Generated samples are given by: $x_{mixup} = \lambda x_i + (1-\lambda)x_j, y_{mixup} = \lambda y_i + (1-\lambda)y_j$. The parameter λ ranges from 0 to 1 and is distributed according to a beta distribution: $\lambda\tilde{}$ $Beta(\beta, \beta)$ for $\beta \in (0, \infty)$. The samples to be combined are chosen randomly from all available images (in our case, batches). We set hyperparameter β as 0.2, empirically, and with reference to original paper [14]. A more specific explanation of the implementation is shown in Fig. 1.

Table 1
Scan conditions of the three datasets.

	Scanner manufacturer /(model)		Slice thickness		Scan coverage		Contrast material	
In-house dataset	Canon Medical Systems	32	0.5 mm	4	Head	4	Uncontrasted	20
	(Aquilion 64)	8	1 mm	20	Chest	5	Intravenous	12
	(Aquilion ONE)	11	5 mm	8	Abdomen	2		
	(Aquilion PRIME)	13			Neck, chest	2		
					Chest, abdomen	7		
					Neck, chest, abdomen	12		
Secondary dataset	Unknown	20	1 mm	15	Chest, abdomen	20	Intravenous	20
			1.25 mm	5			Oral	20
Public dataset	Philips	27	5 mm	27	Head	4	Uncontrasted	5
					Chest	9	Intravenous	22
					Abdomen	6		
					Limbs	8		

Each value represents the number of scans.

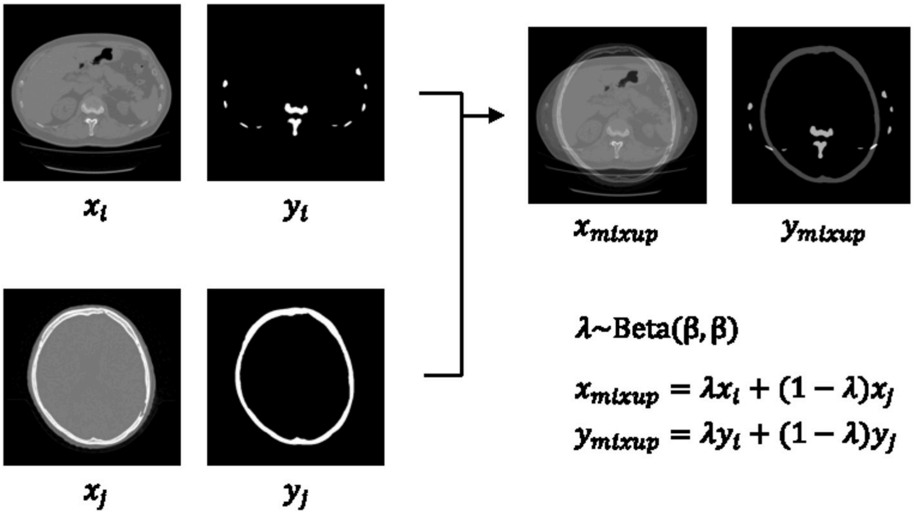
x_i
y_i
x_j
y_j
x_{mixup}
y_{mixup}
$\lambda \sim \mathrm{Beta}(\beta, \beta)$
$x_{mixup} = \lambda x_i + (1-\lambda)x_j$
$y_{mixup} = \lambda y_i + (1-\lambda)y_j$

Fig. 1. Implementation of mixup.

3. RICAP

Random image cropping and patching (RICAP) is another recently proposed augmentation technique in which a new training image is constructed from four randomly selected images [15]. The four images are randomly cropped and patched according to the boundary position (w, h), which is generated from a beta distribution: $Beta(\beta, \beta)$. We set hyperparameter β as 0.3 empirically, and with reference to original paper [15]. The coordinates $(x_k, y_k)(k = 1, 2, 3,\ and\ 4)$ of the upper left corners of the cropped areas are randomly selected based on the value of (w, h), such that it does not increase the image size. A more specific explanation of the implementation is shown in Fig. 2.

Although the mixup and RICAP algorithms were initially proposed for image classification tasks, we utilized them for the task of image segmentation by cropping and patching labels in same way as is done for original images.

2.3. Architecture

A convolutional deep neural network, termed U-Net [16], was adapted to perform segmentation (Fig. 3). Our configuration consisted of four down-sampling and four up-sampling steps, which reduced the 512 × 512 × 1 input image to a 32 × 32 × 64 representation and then up-sampled it into a 512 × 512 × 1 output.

The images were fed into the network after dividing the original CT value by 1000. Adjustment of window setting was not performed. No special preprocessing was performed to accommodate differences in slice thickness and axial in-plane resolution, or differences in scanner models and facilities.

Each down-sampling step consisted of the repeated application of two 3 × 3 convolutions (padded), followed by a rectified linear unit (ReLU) and a 2 × 2 max pooling operation with stride 2 for down-sampling. The number of feature channels in the first convolutional channel was set to 16 and doubled in each of the next two down-sampling steps. In the second half of the network, up-sampling

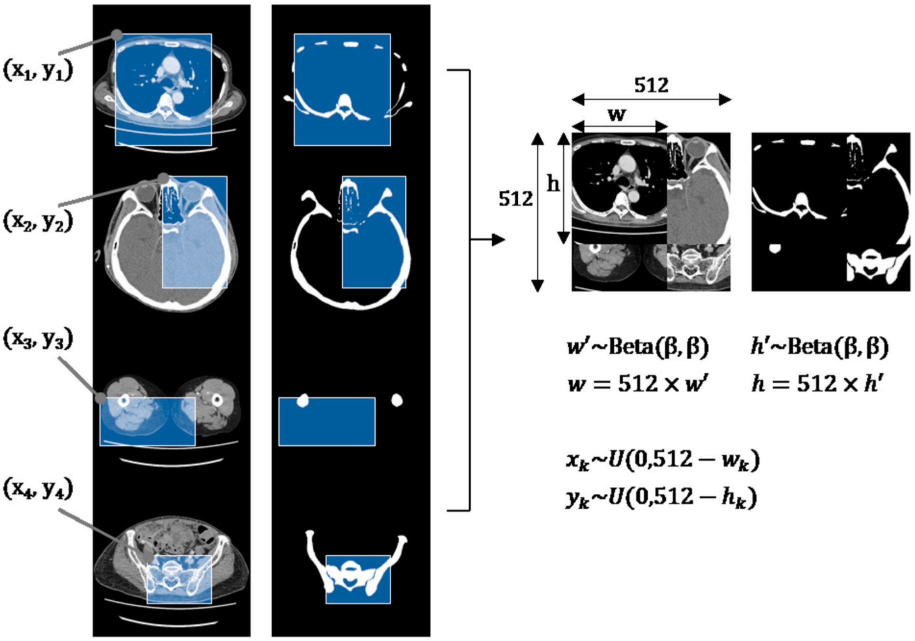
(x_1, y_1)
(x_2, y_2)
(x_3, y_3)
(x_4, y_4)
512
w
512
h
$w' \sim \mathrm{Beta}(\beta, \beta)$
$w = 512 \times w'$
$h' \sim \mathrm{Beta}(\beta, \beta)$
$h = 512 \times h'$
$x_k \sim U(0{,}512 - w_k)$
$y_k \sim U(0{,}512 - h_k)$

Fig. 2. Implementation of RICAP

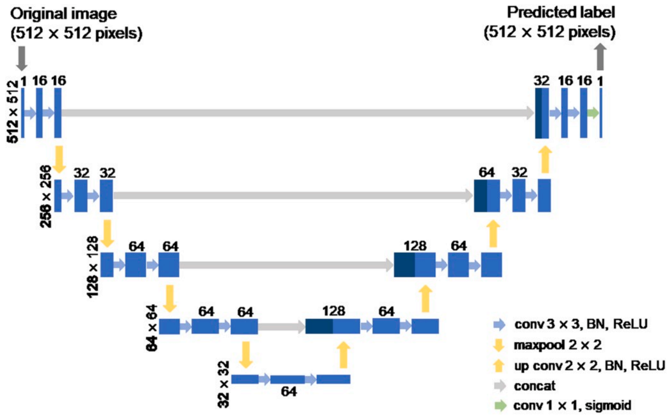

Fig. 3. Schematic illustration of the U-Net architecture used in the present study.

operations were performed followed by two 3 × 3 convolutional layers with ReLU activation. A key feature of the U-Net architecture is that the convolutional kernel outputs from the encoding half of the network are concatenated with each of the corresponding decoding steps, which helps preserve the detail of the original images. In the final layer, 1 × 1 convolutions followed by sigmoid activation functions were used to output scores for the segmentation. Pixels that scored greater than 0.5 were labelled as bone.

We applied batch normalization operation [17] after each convolutional layer to normalize the batch data with its mean and variance. This operation can accelerate the learning convergence of the network during the training stage. The optimal hyperparameters were determined experimentally. Weight initialization of the network was performed using He uniform initialization. Batch size was 8 sections.

The network was written in Python 3.6.5 (Python Software Foundation, Beaverton, Oregon) using Keras 2.2.4 (open source) with Tensorflow 1.10.1 (open source, Google, Mountain View, Calif) backend.

2.4. Training

For the in-house dataset, the network was trained using 20 epochs. RMSprop optimizer was used for parameter optimization, with a learning rate of 0.0005 for the first 10 epochs, followed by 0.00005 for the next 10 epochs. Performance on the training and validation sets were assessed using Dice coefficient loss.

The secondary dataset was employed for testing only, not for training.

For the public dataset, the network was initially trained and evaluated separately from the in-house dataset. The number of training epoch was set to 500, because this dataset is much smaller than the in-house dataset. The learning rates of the public and in-house datasets decayed in the same manner. Additionally, we also employed fine-tuning; a model pre-trained on the in-house dataset without data augmentation was re-trained and fine-tuned on the public dataset. In the process of fine-tuning, all layers in the network were set to be trainable (in other word, no layer was frozen). The number of training epoch was set to 200, because employment of pre-trained model was expected to accelerate convergence.

2.5. Thresholding-based method

For comparison with the proposed model, a conventional thresholding-based segmentation method was implemented. Thresholding was performed at a threshold of 200 HU, followed by morphological closing and opening. Morphological operations were implemented using SciPy 1.1.0 (open source).

2.6. Quality measurements

Dice coefficient, Jaccard index, sensitivity, and positive predictive value were used to evaluate segmentation accuracy in the experiments. The Dice coefficient was used as the major metric. Each of these metrics highlights a different aspect of the quality of the segmentation [18].

2.7. Visual assessment

In addition to quantitative evaluation described above, two series of visual assessment were performed for difficult cases; i.e. visual assessment for (1) bone metastasis lesions and (2) limbs and head scans.

Bone metastasis lesions were collected from the in-house dataset. Among 16 patients included in the in-house dataset, 9 patients had known sites of bone metastases, and 16 lesions were selected for evaluation (small lesions less than 1 cm were excluded). These lesions were divided into three types depending on their imaging patterns: sclerotic bone metastases (7 lesions), lytic bone metastases (7 lesions), and mixed sclerotic and lytic bone metastases (2 lesions).

Limbs and head scans for visual assessment consisted of 18 scans obtained in our institute; shoulder, elbow, hand, hip, knee, and foot: 2 scans each, brain and jaw: 3 scans each. 6 of 18 scans include metallic implants in the scan coverage. None of them were included in the in-house dataset. Since our in-house dataset include no limbs scan and only four head scans, segmentation of these cases was expected to be difficult.

Visual assessment of segmentation results was performed by two board certificated radiologists (14- and 6-years' experience in diagnostic imaging, respectively) using a 5-point scale; 1: unacceptable, 2: slightly unacceptable, 3: acceptable, 4: good, 5: excellent. The final score was decided by a consensus of two radiologists.

3. Results

1. In-house Dataset

Table 2 lists the results for the in-house dataset. The statistics for test

Table 2
Results for the in-house dataset.

	DC	JI	SE	PPV	Training time[a]
Thresholding	0.782 ± 0.102	0.654 ± 0.144	0.773 ± 0.116	0.806 ± 0.127	
Proposed model					
+ no augmentation	0.983 ± 0.005	0.968 ± 0.009	0.984 ± 0.004	0.983 ± 0.007	8173 ± 194
+ conv	0.981 ± 0.004	0.962 ± 0.008	0.976 ± 0.006	0.985 ± 0.006	11,212 ± 348
+ mixup	0.981 ± 0.005	0.963 ± 0.009	0.983 ± 0.004	0.980 ± 0.008	8444 ± 207
+ RICAP	0.983 ± 0.005	0.967 ± 0.010	0.985 ± 0.004	0.981 ± 0.008	8237 ± 190
+ conv + mixup	0.980 ± 0.003	0.961 ± 0.006	0.979 ± 0.006	0.981 ± 0.005	22,044 ± 892
+ conv + RICAP	0.982 ± 0.005	0.964 ± 0.009	0.981 ± 0.005	0.982 ± 0.006	12,229 ± 1152
+ mixup + RICAP	0.980 ± 0.004	0.962 ± 0.008	0.984 ± 0.005	0.977 ± 0.008	7433 ± 169
+ conv + mixup + RICAP	0.979 ± 0.004	0.960 ± 0.007	0.981 ± 0.006	0.978 ± 0.007	23,591 ± 1029

DC, Dice coefficient; JI, Jaccard index; SE, sensitivity; PPV, positive predictive value.

[a] Training time for each fold (seconds).

predictions were calculated for each fold, and the average was presented as the final score. The proposed segmentation algorithm achieved a Dice coefficient of 0.983 ± 0.005 and Jaccard index of 0.968 ± 0.009 (the confidence bounds show the standard deviation over three folds), without using data augmentation. None of the three data augmentation techniques improved the results. Samples of the predicted images are shown in Figs. 4 and 5.

Obvious segmentation errors were observed most frequently for costal cartilage, intervertebral discs, metal from dental work, and surgical implants, as shown in Fig. 6a–c. However, segmentation errors on metal were minimized with proposed model, compared to segmentation with thresholding-based method (Fig. 6c). Otherwise, the predicted labels were mostly perfect (Figs. 4 and 5).

The training time was ~2 h per fold. The segmentation of a whole-body CT scan (~600 slices) took ~12 s using a GeForce GTX 1080 Ti (NVIDIA, Santa Clara, CA, USA) graphics processing unit (GPU).

2. Secondary Dataset

When the network trained on the in-house dataset was tested on the secondary dataset (without using data augmentation) it achieved a Dice coefficient of 0.943 ± 0.007 and Jaccard index of 0.898 ± 0.010. The results are shown in detail in Table 3. Conventional augmentation and RICAP improved the prediction accuracy in the secondary dataset, although the improvement was relatively small (increase in the Dice coefficient was 0.002–0.004). In contrast, mixup worsen the results.

Unlike the in-house dataset, the most significant source of segmentation error in the secondary dataset was contrast material in the gastrointestinal tract (Fig. 6d). We consider that this error occurred because none of the scans in the in-house dataset were acquired with oral contrast material, although high-density debris was observed incidentally in the gastrointestinal tract of several patients.

3. Public Dataset

Table 4 lists the results for the public dataset. Initially, our model was trained and evaluated for this dataset separately from the in-house dataset and the secondary dataset. Then we also employed fine-tuning of the model pre-trained on in-house dataset.

The network that was trained with a combination of conventional augmentation and RICAP achieved the best results (Dice coefficient, 0.947 ± 0.013; Jaccard index, 0.899 ± 0.023; both without fine-tuning), and these results are superior to those reported previously (Dice

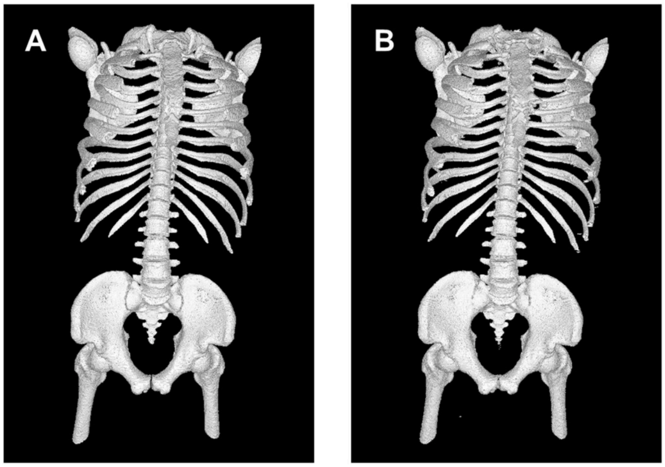

Fig. 4. Reconstructed 3D image of (A) ground truth bone label from in-house dataset, and (B) segmentation result of proposed model, trained on in-house dataset without data augmentation. Segmentation result was near perfect, except for small errors observed on calcified costal cartilage.

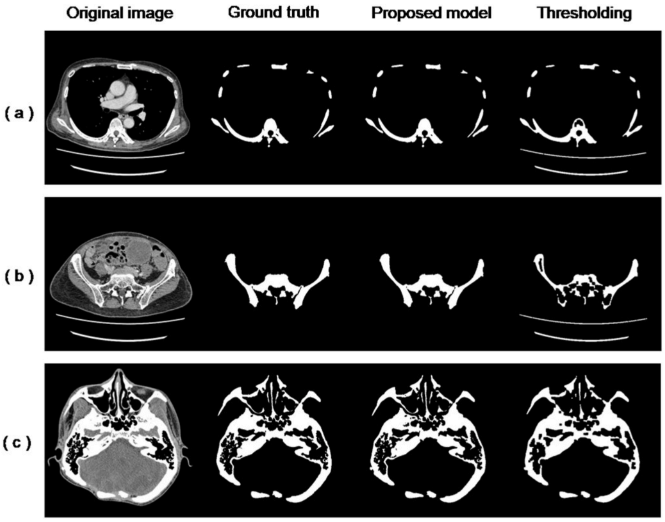

Fig. 5. Representative segmentation results for the in-house dataset. (a), (b) and (c) are samples of test images of chest, abdominal and head CT, respectively. From left to right, the four images in each row are the original image, ground truth label, segmentation result of proposed model (trained without data augmentation), and segmentation result of thresholding-based method. Segmentation results of proposed model are near perfect, whereas obvious segmentation errors were observed in thresholding-based method (on scanner couch, fatty bone marrow, and small structures such as mastoid air cells, for example).

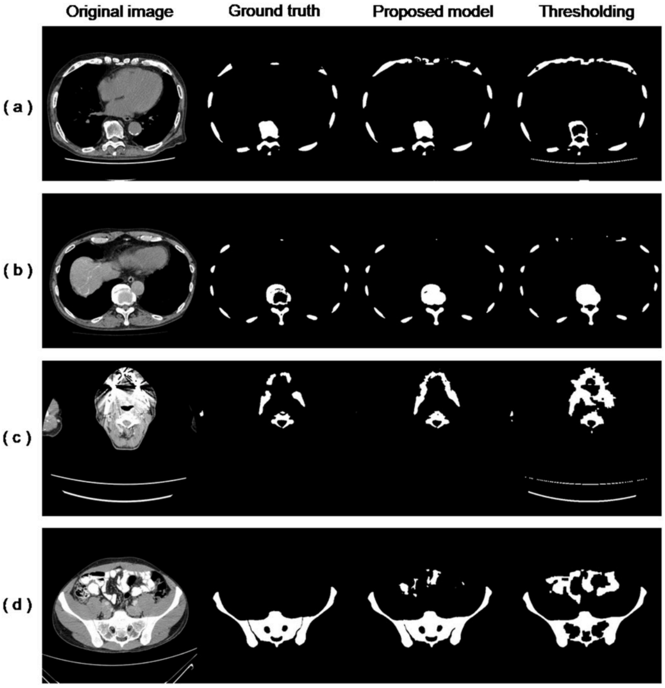

Fig. 6. Representative cases of segmentation errors obtained with a model trained without data augmentation. For the in-house dataset, segmentation errors were most frequently observed in costal cartilage (a), intervertebral discs (b), dental work (c), and surgical implants. For the secondary dataset, the most significant source of segmentation error was contrast material in the gastrointestinal tract (d). However, segmentation errors on artificial materials were minimized with proposed model, compared to segmentation with thresholding-based method (c, d).

Table 3
Results for the secondary dataset.

	DC	JI	SE	PPV
Thresholding	0.740 ± 0.073	0.592 ± 0.090	0.851 ± 0.071	0.656 ± 0.079
Proposed model (trained without data augmentation)				
+ no augmentation	0.943 ± 0.007	0.898 ± 0.010	0.976 ± 0.001	0.918 ± 0.010
+ conv	0.947 ± 0.010	0.904 ± 0.015	0.976 ± 0.001	0.925 ± 0.015
+ mixup	0.906 ± 0.045	0.846 ± 0.058	0.974 ± 0.002	0.865 ± 0.062
+ RICAP	0.946 ± 0.008	0.902 ± 0.012	0.979 ± 0.002	0.920 ± 0.013
+ conv + mixup	0.927 ± 0.010	0.871 ± 0.015	0.976 ± 0.002	0.890 ± 0.018
+ conv + RICAP	0.945 ± 0.014	0.902 ± 0.021	0.977 ± 0.001	0.922 ± 0.021
+ mixup + RICAP	0.934 ± 0.010	0.879 ± 0.018	0.979 ± 0.003	0.895 ± 0.019
+ conv + mixup + RICAP	0.946 ± 0.009	0.900 ± 0.015	0.979 ± 0.002	0.918 ± 0.016

Table 4
Results for the public dataset.

	DC	JI	SE	PPV	Training time[a]
Thresholding	0.707 ± 0.221	0.586 ± 0.240	0.815 ± 0.120	0.713 ± 0.305	
Proposed model without fine-tuning (trained using 500 epochs)					
+ no augmentation	0.908 ± 0.018	0.833 ± 0.030	0.901 ± 0.032	0.918 ± 0.029	2556 ± 19
+ conv	0.942 ± 0.014	0.892 ± 0.025	0.935 ± 0.021	0.952 ± 0.025	4533 ± 109
+ mixup	0.892 ± 0.037	0.809 ± 0.058	0.858 ± 0.062	0.937 ± 0.026	2645 ± 15
+ RICAP	0.943 ± 0.014	0.893 ± 0.024	0.942 ± 0.021	0.946 ± 0.021	2635 ± 23
+ conv + mixup	0.943 ± 0.014	0.894 ± 0.024	0.929 ± 0.024	0.959 ± 0.018	8636 ± 293
+ conv + RICAP	0.947 ± 0.013	0.899 ± 0.023	0.943 ± 0.020	0.952 ± 0.023	5115 ± 123
+ mixup + RICAP	0.933 ± 0.020	0.877 ± 0.034	0.916 ± 0.035	0.954 ± 0.023	3459 ± 107
+ conv + mixup + RICAP	0.945 ± 0.009	0.896 ± 0.017	0.937 ± 0.023	0.954 ± 0.023	9714 ± 64
Proposed model with fine-tuning (trained using 200 epochs)					
+ no augmentation	0.950 ± 0.012	0.906 ± 0.022	0.937 ± 0.024	0.965 ± 0.008	1033 ± 5
+ conv	0.942 ± 0.019	0.890 ± 0.031	0.952 ± 0.026	0.933 ± 0.036	1737 ± 63
+ mixup	0.952 ± 0.006	0.909 ± 0.010	0.944 ± 0.016	0.962 ± 0.017	1073 ± 4
+ RICAP	0.951 ± 0.007	0.908 ± 0.014	0.939 ± 0.020	0.966 ± 0.009	1064 ± 5
+ conv + mixup	0.942 ± 0.026	0.890 ± 0.044	0.955 ± 0.019	0.930 ± 0.044	3333 ± 88
+ conv + RICAP	0.961 ± 0.007	0.926 ± 0.013	0.957 ± 0.016	0.966 ± 0.012	1912 ± 42
+ mixup + RICAP	0.939 ± 0.026	0.887 ± 0.022	0.937 ± 0.024	0.965 ± 0.008	1033 ± 5
+ conv + mixup + RICAP	0.954 ± 0.007	0.914 ± 0.012	0.952 ± 0.020	0.959 ± 0.013	3637 ± 244

[a] Training time for each split (seconds).

coefficient, 0.92 ± 0.05; Jaccard index, 0.85 ± 0.08) [5]. For the present task, conventional augmentation, RICAP, and a combination of these methods were the most effective. We observed a significant increase in Dice coefficient (from 0.034 to 0.039) over the baseline model for the test dataset. In contrast, mixup did not appear to be suitable for the task. Training time was approximately 40 min per split, using a GeForce GTX 1080 Ti GPU (NVIDIA). The conventional augmentation method prolonged training time significantly, whereas mixup and RICAP did not.

With fine-tuning of the model pre-trained on in-house dataset, the scores generally improved, achieving highest Dice coefficient of 0.961 ± 0.007 with conventional augmentation and RICAP. Compared to the model trained without fine-tuning, the increasement in Dice coefficient was largest when no augmentation was used (increased from 0.908 to 0.950).

4. Visual Assessment

Table 5 shows results of the visual assessment for bone metastasis lesions and for limbs and head scans. Fig. 7 shows representative images.

Average scores for sclerotic bone metastases, mixed sclerotic and lytic bone metastases, and lytic bone metastases were 5.00, 4.00, and 2.43, respectively, and the overall score was 3.75. Segmentation results for sclerotic lesions and mixed lesions were generally good (Fig. 7A and B), whereas osteolytic lesions were relatively poor (Fig. 7C). However, segmentation results for small lytic lesions were acceptable.

As for limbs and head scans, average scores for limbs scans and head scans were 3.42 and 4.5, respectively, and the overall score was 3.78. Segmentation results of limbs were good or acceptable in most cases (Fig. 7D and E). However, segmentation at the tip of the finger or toe were often defected (Fig. 7F), and segmentation around the artificial joints were rather poor (Fig. 7G). Segmentation for head scans were generally good (Fig. 7H).

4. Discussion

The results of our study show insight into application of deep neural networks to bone segmentation on whole-body CT. We evaluated the performance of our network for three different datasets and achieved sufficient accuracy. The efficacy of data augmentation methods was demonstrated for a public dataset.

For decades, various methodologies have been proposed for bone segmentation on CT [1]. Of these, thresholding has been most widely applied [19–24]. As bony structures have high density levels on CT, they can usually be separated from soft tissue by thresholding. Other techniques applied for bone segmentation include region growing [22,23], edge detection [24,25], atlas, statistical shape modeling [26,27,29–32], contextual and topological modeling [34], anisotoropic diffusion filtering and morphological reconstruction [35], and supervoxel based approach [33]. In practice, these methods are seldom used independently, and are commonly combined. However, the majority of these methodologies focus on particular bone types and are not applicable to all bones on whole-body CT; in addition, most require some degree of manual input, which is laborious in clinical practice.

CNN-based methods have recently achieved strong results in various fields, including medical image segmentation [2–4]. Several studies have investigated CNN-based architecture for bone segmentation on whole-body CT. Klein et al. proposed a bone segmentation model for CT images based on CNN, with the main purpose of assessing patients

Table 5
Results of visual assessment for bone metastasis lesions and for limbs and head scans.

	Number of score 1	Number of score 2	Number of score 3	Number of score 4	Number of score 5	Average score
Sclerotic lesions					7	5.00
Mixed lesions				2		4.00
Lytic lesions	2	2	1	2		2.43
Lesions total	2	2	1	4	7	3.75
Limbs scans		2	4	5	1	3.42
Head scans				3	3	4.50
Limbs and head total		2	4	8	4	3.78

Score 1: unacceptable, 2: slightly unacceptable, 3: acceptable, 4: good, 5: excellent.

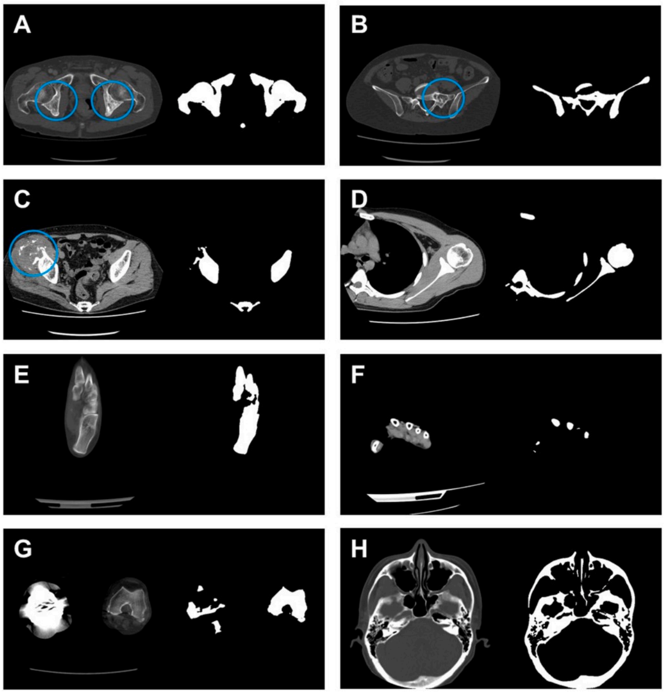

Fig. 7. Representative segmentation results used for visual assessment of bone metastasis lesions (A–C: lesions are annotated with circles) and limbs and head scans (D–H). The left images are original images, and the right are segmentation results of proposed model trained on in-house dataset without data augmentation. Note that original images are displayed with different window setting, for better visualization of bones and lesions. Segmentation results for sclerotic lesions and mixed lesions were generally good (A: sclerotic lesion, scored 5; B: mixed lesion, scored 4), whereas osteolytic lesions were poor (C: lytic lesion, scored 1). As for limb scans, segmentation results were good or acceptable in most cases (D: shoulder scan, scored 4; E: foot scan, scored 4). Segmentation at the tip of the finger or toe were often defected (F: hand scan, scored 3), and segmentation around the artificial joints were rather poor (G: knee scan with right artificial joint, scored 2; note that segmentation of contralateral knee was good). Segmentation for head scan was generally good (H: a slice from mandible scan, scored 4; small structures such as mastoid air cells and paranasal sinus were well segmented).

suffering from multiple myeloma [5]. Their network, which was trained on a dataset comprising 18 low-dose CT scans that were obtained as part of positron emission tomography/computed tomography (PET/CT) studies, achieved a Dice coefficient of 0.95 and Jaccard index of 0.91. However, the model was trained and tested on a dataset acquired with identical scan conditions. Belal et al. proposed a CNN-based model for bone segmentation with the aim of detecting 49 bones of the axial skeleton, for the development of a quantifiable measure of bony changes assessed on PET/CT [6]. However, not all bones in the axial skeleton were included as targets for segmentation (cervical vertebrae were excluded). In addition, evaluation of the model's accuracy was performed for only five selected bones.

Our model showed outstanding results in all metrics (Dice coefficient, Jaccard index, sensitivity, and positive predictive value) of model evaluation for the in-house dataset; which is considered to be attributable to the high image quality of the dataset. Image processing was rapid with a currently available GPU, requiring ~12 s to segment a whole-body CT of ~600 slices. This was partially because our model was comparatively simple; our model has at most 128 feature channels, whereas 1024 in the original paper of U-Net [16] and previous study using U-Net for bone segmentation [5]. In our experience, too much number of channels was not beneficial for U-Net-based bone segmentation model, resulting in overfitting and increased computation time. We proved the generalizability of our model by testing on a secondary dataset that had been obtained under different scan conditions. Segmentation accuracy with the secondary dataset was sufficiently high, aside from oral contrast material observed in the gastrointestinal tract. We consider that our network would be generalizable to CT images obtained under various conditions; e.g., to images from different institutes and to subjects of different racial backgrounds.

Although the accuracy of segmentation for the in-house dataset was near-perfect, some errors were found, most common sources being costal cartilage, intervertebral discs, metal from dental work, and surgical implants (Fig. 6a–c). Costal cartilage often becomes calcified, especially in elderly patients, and thus presents with a similar shape and density to ribs. Because it is cartilage rather than bone, costal cartilage was not labelled as bone in the ground truth data. As well as being difficult for our network, it is also difficult for human reviewers to visually distinguish calcified costal cartilages from ribs. Intervertebral discs were also difficult to segment because we used only 2D axial images for training and testing. That is, when an intervertebral disc and the upper and/or lower adjacent vertebras are included in a single axial slice, the intervertebral disc tends to be labelled as part of the vertebra. Although we could have used 3D volume images to assist with this problem, more computational resources would have been needed. The segmentation of images that include metal from dental work and surgical implants was another challenging task. These items cause strong artifacts on CT images, and thus it is difficult to delineate a clear tissue boundary around them. In our experiments, we observed over-segmentation as well as under-segmentation due to metallic implants. Increasing the number of samples including these metallic implants is considered to be beneficial for reducing this type of error. In the secondary dataset, the most significant source of segmentation error was oral contrast material in the gastrointestinal tract (Fig. 6d). As none of the scans included in the in-house dataset was acquired with oral contrast material, it is not unexpected that it was difficult to assess features that were not included in the training data.

In addition to quantitative evaluation for the three datasets, we also performed visual assessment of segmentation results for bone metastasis lesions (collected from the in-house dataset) and limbs and head scans (collected separately from the three datasets). In the clinical setting, bone segmentation is mainly applied to pathological cases. Therefore, it was important to validate the performance of our model on these cases. According to the visual assessment, segmentation for large osteolytic bone metastases were insufficient. This is attributed to the fact that segmentation of bones on CT is highly dependent on density (i.e., Hounsfield units), and density of osteolytic bone metastases is much different from normal bone. For better segmentation of these lesions, more osteolytic lesions might be needed in the training data. However, segmentation for sclerotic lesions, mixed lesions, and small lytic lesions were good or acceptable. Segmentation for limbs and head scans were also good or acceptable (except for limbs scans with surgical implants), despite that these parts of the body were not included or only a few cases in the training data. It is speculated that our model captures not only features specific to individual bones, but also general characteristics of whole-body bones. Overall, these results indicated that our model was applicable, at least to some extent, to pathological cases and to other parts of body which were not included in the training data.

To improve the performance and robustness of the network, we applied three types of data augmentation techniques. Medical image segmentation is often constrained by availability of labelled training data. Data augmentation artificially increases the size of the training dataset, helps to prevent memorization of the training data and improves the network's performance on data from outside the training set. As such, it is necessary in building robust deep learning models. Augmentation in medical imaging typically involves applying small transformations such as rotations, resizes, reflections, and elastic deformations. In the segmentation task, these transformations are applied to both the images and labels equally, creating warped versions of the training data. Although several new augmentation methods have been proposed, most have been used on image classification tasks. Eaton-Rosen et al. proposed the application of mixup to medical image segmentation [28]. Although the images generated by this method are noticeably different to the training images (appearing as two super-posed images; see Fig. 1), this augmentation technique improved the segmentation performance for brain tumor for a multi-modal MRI dataset [28].

In our study, mixup did not improve the results on any of three datasets. In contrast, the conventional method and RICAP achieved an increase in the Dice coefficient over the baseline model. It is likely that these disparities occurred because distinction of bony structure on CT is highly dependent on density (as mentioned above). The gray levels in the images generated by mixup are different from original images, whereas conventional method and RICAP never alter gray level of images. We consider that these characteristics explain the differences in the results and learning efficiency. It is of note that RICAP completely alters the shape and localization of each organ. Because distinction of the internal organs is dependent on these factors, the efficacy of RICAP for the segmentation of other organs should be discussed separately.

Generative adversarial networks (GAN) is another method to generate synthetic images for training, and several reports have demonstrated its efficiency [38]. However, using GAN for data augmentation is a complicated approach and its cost-effectiveness need some consideration. In this regard, mixup and RICAP is rather simple and easy to implement.

Our results demonstrated that implementing data augmentation methods enables bone segmentation networks to be built with high accuracy, even if the dataset is rather small, such as for 150 training images. In contrast, data augmentation had little efficacy for the in-house dataset. We presume that the impact of data augmentation lessens to some degree when the dataset is large.

Fine-tuning a network which has been trained on a large dataset is an alternative to full training in order to overcome the problem of limited data availability [37]. In our experiment on public dataset, the segmentation accuracy was generally improved with fine-tuning of the model pre-trained on in-house dataset. Fine-tuning is worth trying on the task of medical image segmentation, if appropriate pre-trained model is available.

There were several limitations to the present study. First, the variety of pathological conditions included in the visual assessment was limited. We validated segmentation performance on bone metastasis lesions; however, other types of bone disease such as multiple myeloma, massive bone deformation, or congenital malformation were not included. To fully assess generalizability of the model to other pathological conditions, further investigation is required. Second, observed segmentation accuracy might be biased by frequency of artificial materials in the training dataset and test dataset. According to our experiments, the most significant source of error on bone segmentation was artificial materials such as contrast media and surgical implants. Though this type of errors could be lessened by increasing the numbers of samples including such materials, performance of segmentation might be affected by distribution of these samples in the datasets. Third, segmentation accuracy of limbs scans was evaluated only by visual assessment relies on radiologists' subjective judgement, and no objective evaluation was performed. This was because we did not have manually annotated ground truth label for these parts of the body.

5. Conclusion

We developed a fully automated bone segmentation system based on CNN and evaluated its robustness using a secondary dataset obtained under different scan conditions, obtained from patients with different racial background. After being trained on a dataset of sufficient variability, the deep CNN was able to detect and segment bony structure with high accuracy. The experimental results demonstrated that implementation of RICAP in addition to the conventional augmentation method could improve segmentation accuracy, by increasing the variety of training images and preventing overfitting.

Declaration of competing interest

K. Togashi received a research grant from Canon Medical Systems Corporation. K. Nakagomi is an employee of Canon Inc. However, these companies had no role on design, preparation, review or approval of this paper. The other authors declare that they have no conflict of interest related to this paper.

Acknowledgments

The present study was partially supported by JSPS KAKENHI (Grant Number JP19K17232).

References

[1] M. van Eijnatten, R. van Dijk, J. Dobbe, G. Streekstra, J. Koivisto, J. Wolff, CT image segmentation methods for bone used in medical additive manufacturing, Med. Eng. Phys. 51 (2018) 6–16, https://doi.org/10.1016/j.medengphy.2017.10.008.

[2] Z. Zhang, E. Sejdić, Radiological images and machine learning: trends, perspectives, and prospects, Comput. Biol. Med. 108 (2019) 354–370, https://doi.org/10.1016/j.compbiomed.2019.02.017.

[3] C.S. Perone, J. Cohen-Adad, Promises and limitations of deep learning for medical image segmentation, J Med Artif Intell 2 (1) (2019), https://doi.org/10.21037/jmai.2019.01.01.

[4] M.H. Hesamian, W. Jia, X. He, P. Kennedy, Deep learning techniques for medical image segmentation: achievements and challenges, J. Digit. Imag. 32 (4) (2019) 582–596, https://doi.org/10.1007/s10278-019-00227-x.

[5] A. Klein, J. Warszawski, J. Hillengaß, K.H. Maier-Hein, Automatic bone segmentation in whole-body CT images, Int J Comput Assist Radiol Surg 14 (1) (2019) 21–29, https://doi.org/10.1007/s11548-018-1883-7.

[6] S. Lindgren Belal, M. Sadik, R. Kaboteh, O. Enqvist, J. Ulén, M.H. Poulsen, J. Simonsen, P.F. Høilund-Carlsen, L. Edenbrandt, E. Trägårdh, Deep learning for segmentation of 49 selected bones in CT scans: first step in automated PET/CT-based 3D quantification of skeletal metastases, Eur. J. Radiol. 113 (2019) 89–95, https://doi.org/10.1016/j.ejrad.2019.01.028.

[7] J. Minnema, M. van Eijnatten, W. Kouw, F. Diblen, A. Mendrik, J. Wolff, CT image segmentation of bone for medical additive manufacturing using a convolutional neural network, Comput. Biol. Med. 103 (2018) 130–139, https://doi.org/10.1016/j.compbiomed.2018.10.012.

[8] H.R. Roth, L. Lu, A. Seff, K.M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, Summers RM: a new 2.5D representation for lymph node detection in CT, Canc. Imag. Archive (2015), https://doi.org/10.7937/K9/TCIA.2015.AQIIDCNM.

[9] H.R. Roth, L. Lu, A. Seff, K.M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, R. M. Summers, A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations, Med. Image Compu.

Comput. Assisted Intervent. (2014) p520–527, https://doi.org/10.1007/978-3-319-10404-1_65.

[10] A. Seff, L. Lu, K.M. Cherry, H.R. Roth, J. Liu, S. Wang, J. Hoffman, E. Turkbey, R. M. Summers, 2D view aggregation for lymph node detection using a shallow hierarchy of linear classifiers, Med. Image Compu. Comput. Assisted Intervent. (2014) p544–552, https://doi.org/10.1007/978-3-319-10404-1_68.

[11] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, F. Prior, The cancer imaging archive (TCIA): maintaining and operating a public information repository, J. Digit. Imag. 26 (6) (2013) 1045–1057, https://doi.org/10.1007/s10278-013-9622-7.

[12] J.A. Pérez-Carrasco, B. Acha, C. Suárez-Mejías, J.L. López-Guerra, C. Serrano, Joint segmentation of bones and muscles using an intensity and histogram-based energy minimization approach, Comput. Methods Progr. Biomed. 156 (2018) 85–95, https://doi.org/10.1016/j.cmpb.2017.12.027.

[13] K. Gorman, B. Steven, We need to talk about standard splits. https://wellformedness.com/papers/gorman-bedrick-2019.pdf. (Accessed 24 July 2019).

[14] H. Zhang, M. Cisse, Y.N. Dauphin, Lopez-Paz D: mixup: beyond empirical risk minimization, arXiv (2017) arXiv:1710.09412.

[15] R. Takahashi, T. Matsubara, K. Uehara, RICAP: random image cropping and patching data augmentation for deep CNNs, in: Asian Conference on Machine Learning, 2018, pp. p786–798.

[16] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015, pp. p234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

[17] S. Ioffe, C. Szegedy, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, 2015 arXiv preprint arXiv:1502.03167.

[18] H.H. Chang, A.H. Zhuang, D.J. Valentino, W.C. Chu, Performance measure characterization for evaluating neuroimage segmentation algorithms, Neuroimage 47 (1) (2009) 122–135, https://doi.org/10.1016/j.neuroimage.2009.03.068.

[19] D.L. Pham, C. Xu, J.L. Prince, Current methods in medical image segmentation, Annu. Rev. Biomed. Eng. 2 (1) (2000) 315–337, https://doi.org/10.1146/annurev.bioeng.2.1.315.

[20] H.R. Buie, H.R. Buie, G.M. Campbell, R.J. Klinck, J.A. MacNeil, S.K. Boyd, Automatic segmentation of cortical and trabecular compartments based on a dual threshold technique for in vivo micro-CT bone analysis, Bone 41 (4) (2007) 505–515, https://doi.org/10.1016/j.bone.2007.07.007.

[21] T.N. Hangartner, Thresholding technique for accurate analysis of density and geometry in QCT, pQCT and muCT images, J. Musculoskelet. Neuronal Interact. 7 (1) (2007) 9.

[22] J. Zhang, C.H. Yan, C.K. Chui, S.H. Ong, Fast segmentation of bone in CT images using 3D adaptive thresholding, Comput. Biol. Med. 40 (2) (2010) 231–236, https://doi.org/10.1016/j.compbiomed.2009.11.020.

[23] M. Fiebich, C.M. Straus, V. Sehgal, B.C. Renger, K. Doi, K.R. Hoffmann, Automatic bone segmentation technique for CT angiographic studies, J. Comput. Assist. Tomogr. 23 (1) (1999) 155–161, https://doi.org/10.1097/00004728-199901000-00031.

[24] K. Rathnayaka, T. Sahama, M.A. Schuetz, B. Schmutz, Effects of CT image segmentation methods on the accuracy of long bone 3D reconstructions, Med. Eng. Phys. 33 (2) (2011) 226–233, https://doi.org/10.1016/j.medengphy.2010.10.002.

[25] M. Krčah, G. Székely, R. Blanc, Fully automatic and fast segmentation of the femur bone from 3D-CT images with no shape prior, IEEE international symposium on biomedical imaging: from nano to macro (2011) p2087–2090, https://doi.org/10.1109/ISBI.2011.5872823.

[26] T. Heimann, H.P. Meinzer, Statistical shape models for 3D medical image segmentation: a review, Med. Image Anal. 13 (4) (2009) 543–563, https://doi.org/10.1016/j.media.2009.05.004.

[27] H. Seim, D. Kainmueller, M. Heller, H. Lamecker, S. Zachow, H.C. Hege, Automatic segmentation of the pelvic bones from CT data based on a statistical shape model, VCBM 8 (2008) 93–100, https://doi.org/10.2312/VCBM/VCBM08/093-100.

[28] Z. Eaton-Rosen, F. Bragman, S. Ourselin, M.J. Cardoso, Improving data augmentation for medical image segmentation, in: International Conference on Medical Imaging with Deep Learning, 2018.

[29] J. Wang, Y. Cheng, C. Guo, Y. Wang, S. Tamura, Shape-intensity prior level set combining probabilistic atlas and probability map constrains for automatic liver segmentation from abdominal CT images, J. Digit. Imag. 11 (5) (2016) 817–826, https://doi.org/10.1007/s11548-015-1332-9.

[30] C. Shi, Y. Cheng, F. Liu, Y. Wang, J. Bai, S. Tamura, A hierarchical local region-based sparse shape composition for liver segmentation in CT scans, Pattern Recogn. 50 (2016) 88–106, https://doi.org/10.1016/j.patcog.2015.09.001.

[31] C. Shi, Y. Cheng, J. Wang, Y. Wang, K. Mori, S. Tamura, Low-rank and sparse decomposition based shape model and probabilistic atlas for automatic pathological organ segmentation, Med. Image Anal. 38 (2017) 30–49, https://doi.org/10.1016/j.media.2017.02.008.

[32] J. Schmid, J. Kim, N. Magnenat-Thalmann, Robust statistical shape models for MRI bone segmentation in presence of small field of view, Med. Image Anal. 15 (1) (2011) 155–168, https://doi.org/10.1016/j.media.2010.09.001.

[33] N. Lay, D. Liu, I. Nogues, R.M. Summers, Accurate 3D bone segmentation in challenging CT images: bottom-up parsing and contextualized optimization, in: Winter Conference on Applications of Computer Vision, 2016, https://doi.org/10.1109/WACV.2016.7477606.

[34] C. Li, D. Jin, C. Chen, E.M. Letuchy, K.F. Janz, T.L. Burns, J.C. Torner, S.M. Levy, P. K. Saha, Automated cortical bone segmentation for multirow-detector CT imaging with validation and application to human studies, Med. Phys. 42 (8) (2015) 4553–4565, https://doi.org/10.1118/1.4923753.

[35] C. Chen, D. Jin, X. Zhang, S.M. Levy, P.K. Saha, Robust segmentation of trabecular bone for in vivo CT imaging using anisotropic diffusion and multi-scale morphological reconstruction, SPIE Medical Imaging, 2017, https://doi.org/10.1117/12.2254546.

[36] Y. Nimura, D. Deguchi, T. Kitasaka, K. Mori, Y. Suenaga, PLUTO: a common platform for computer-aided diagnosis, Med. imaging Technol. 26 (3) (2008) 187–191, https://doi.org/10.11409/mit.26.187 (in Japanese).

[37] N. Tajbakhsh, J.Y. Shin, S.R. Gurudu, R.T. Hurst, C.B. Kendall, M.B. Gotway, J. Liang, Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imag. 35 (5) (2016) 1299–1312, https://doi.org/10.1109/TMI.2016.2535302.

[38] D. Jin, Z. Xu, Y. Tang, A.P. Harrison, D.J. Mollura, CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation, Int. Conf. Med. Image Comput. Comput. Assisted Intervent. (2018) p732–740, https://doi.org/10.1007/978-3-030-00934-2_81.

Shunjiro Noguchi, the corresponding author of this manuscript, is a board-certified radiologist. He is also specialized in medical image analysis. Currently, he works at Kyoto University Graduate School of Medicine as a researcher of medical image analysis.