Original Article | Gastrointestinal Imaging

Korean Journal of Radiology KJR

eISSN 2005-8330
https://doi.org/10.3348/kjr.2022.0588
Korean J Radiol 2023;24(4):294-304

Deep Learning-Based Computed Tomography Image Standardization to Improve Generalizability of Deep Learning-Based Hepatic Segmentation

Seul Bi Lee[1,2]*, Youngtaek Hong[3]*, Yeon Jin Cho[1,2], Dawun Jeong[3,4], Jina Lee[3,4], Soon Ho Yoon[1,2,5], Seunghyun Lee[1,2], Young Hun Choi[1,2], Jung-Eun Cheon[1,2,6]

[1]Department of Radiology, Seoul National University Hospital, Seoul, Korea
[2]Department of Radiology, Seoul National University College of Medicine, Seoul, Korea
[3]CONNECT-AI R&D Center, Yonsei University College of Medicine, Seoul, Korea
[4]Brain Korea 21 PLUS Project for Medical Science, Yonsei University, Seoul, Korea
[5]MEDICALIP Co. Ltd., Seoul, Korea
[6]Institute of Radiation Medicine, Seoul National University Medical Research Center, Seoul, Korea

Objective: We aimed to investigate whether image standardization using deep learning-based computed tomography (CT) image conversion would improve the performance of deep learning-based automated hepatic segmentation across various reconstruction methods.

Materials and Methods: We collected contrast-enhanced dual-energy CT of the abdomen that was obtained using various reconstruction methods, including filtered back projection, iterative reconstruction, optimum contrast, and monoenergetic images with 40, 60, and 80 keV. A deep learning based image conversion algorithm was developed to standardize the CT images using 142 CT examinations (128 for training and 14 for tuning). A separate set of 43 CT examinations from 42 patients (mean age, 10.1 years) was used as the test data. A commercial software program (MEDIP PRO v2.0.0.0, MEDICALIP Co. Ltd.) based on 2D U-NET was used to create liver segmentation masks with liver volume. The original 80 keV images were used as the ground truth. We used the paired *t*-test to compare the segmentation performance in the Dice similarity coefficient (DSC) and difference ratio of the liver volume relative to the ground truth volume before and after image standardization. The concordance correlation coefficient (CCC) was used to assess the agreement between the segmented liver volume and ground-truth volume.

Results: The original CT images showed variable and poor segmentation performances. The standardized images achieved significantly higher DSCs for liver segmentation than the original images (DSC [original, 5.40%–91.27%] vs. [standardized, 93.16%–96.74%], all $P < 0.001$). The difference ratio of liver volume also decreased significantly after image conversion (original, 9.84%–91.37% vs. standardized, 1.99%–4.41%). In all protocols, CCCs improved after image conversion (original, -0.006–0.964 vs. standardized, 0.990–0.998).

Conclusion: Deep learning-based CT image standardization can improve the performance of automated hepatic segmentation using CT images reconstructed using various methods. Deep learning-based CT image conversion may have the potential to improve the generalizability of the segmentation network.

Keywords: Artificial intelligence; Automated segmentation; Image conversion; Quality control; Reproducibility

INTRODUCTION

The development and application of artificial intelligence (AI) in the field of radiology have made significant progress in automated image analysis [1]. In radiology, automated tools using deep learning have been developed for image

Received: August 17, 2022 **Revised:** December 19, 2022 **Accepted:** January 24, 2023

*These authors contributed equally to this work.

Corresponding author: Yeon Jin Cho, MD, PhD, Department of Radiology, Seoul National University Hospital, 101 Daehak-ro, Jongno-gu, Seoul 03080, Korea.
• E-mail: blue1010c@gmail.com

classification, lesion detection, and image segmentation [2]. Automatic image analysis with a deep learning algorithm can improve workflows in radiology by eliminating time-consuming workloads. In addition, the reported accuracies of many deep learning algorithms are beginning to match or even exceed those of radiologists [3-5].

One substantial barrier to the development of deep learning algorithms is securing large-scale annotated data [6,7]. Large and heterogeneous datasets with high-quality images from multiple institutions and different geographic areas are essential for training and developing high-quality deep learning algorithms with general applications [8]. However, curating large datasets is challenging owing to the limited availability of radiologists and tedious annotation processes [9]. Therefore, in most cases, the development of deep learning algorithms has been achieved using a limited dataset that represents the characteristics of certain distributions in the research population, and the developed algorithms may suffer from overfitting, which results in poor generalizability [9-11].

According to a recent review of the external validation of deep learning algorithms for radiologic diagnosis, the vast majority of algorithms demonstrated diminished performance on the external dataset, with some reporting a substantial performance decrease [12]. A few methods have been suggested to improve the generalizability of deep learning algorithms, including transfer learning [10]; however, transfer learning has the disadvantage of additional data augmentation, which is required whenever the characteristics of the input images are changed.

A recent study reported a deep learning algorithm capable of converting various computed tomography (CT) images derived from diverse CT protocols into target CT images [13]. This study used a generative adversarial network (GAN) to standardize CT images to improve the reproducibility of radiomics features. The developed deep learning algorithm shows potential for the standardization of medical image data. Another study reported that data normalization and augmentation improved the generalizability of neural network-based cardiac magnetic resonance image segmentation methods [14]. Therefore, we aimed to investigate whether image standardization using deep learning-based CT image conversion improves the performance of deep learning-based automated hepatic segmentation across various reconstruction methods.

MATERIALS AND METHODS

The institutional review boards of the Seoul National University Hospital (IRB No. 2202-096-1301) approved this retrospective study and waived the requirement for informed patient consent.

Study Population

For the training and tuning datasets, we collected data from 117 patients who underwent 142 contrast-enhanced abdominal CT examinations with dual-energy (DE) scans at a single tertiary hospital between March 2021 and July 2021 (Table 1). The collected data were divided into training and tuning datasets at a ratio of 9:1 (i.e., 128 and 14 examinations, respectively). For the test dataset, we separately collected 43 contrast-enhanced abdominal

Table 1. Characteristics of the Datasets

	Training & Tunning Dataset	Test Dataset
Number of patients (male:female)	117 (57:60)	42 (19:23)
Age* (range)	8.7 ± 5.5 years (2 months–19 years)	10.1 ± 8.7 years (7 months–49 years)
BMI* (range)	18.2 ± 4.6 (10.5–33.4)	18.5 ± 5.0 (12.5–35.2)
Reason for examination	Abdominal pain (n = 15, 12.8%) Tumor follow-up (n = 69, 59.0%), Others (n = 33, 28.2%)	Abdominal pain (n = 3, 7.1%), Tumor follow-up (n = 33, 78.6%), Others (n = 6, 14.3%)
Number of CT scans	142	43
CT machines	SOMATOM Force (Siemens)	SOMATOM Force (Siemens)
Tube voltage	70 kVp and 150 kVp	70 kVp and 150 kVp
Reference tube current	370 mAs for the 70 kVp tube 93 mAs for the 150 kVp tube	370 mAs for the 70 kVp tube 93 mAs for the 150 kVp tube
CT slice thickness	3 mm	3 mm
Scan timing	Portal phase†	Portal phase†

*Values represent mean±standard deviation, †65 s after the initiation of contrast agent injection. BMI = body mass index, CT = computed tomography

https://doi.org/10.3348/kjr.2022.0588

CT examinations in 42 patients without focal hepatic lesions obtained using DE scans at the same tertiary center between December 2021 and January 2022 (Table 1).

The images were acquired using a SOMATOM Force CT scanner (Siemens) in DE mode. The scanned data were reconstructed using filtered back projection (FBP), iterative reconstruction (IR) with a strength of 3 (SAFIRE), and virtual mono-energetic images with 40 keV (M40), 60 keV (M60), 80 keV (M80), and optimum contrast (OPT). The DECT parameters are listed in Table 1.

Architecture of the Deep Neural Network for Image Standardization

Lee et al. [13] proposed a GAN to improve the reproducibility of CT-based radiomics features. We employed a generator network (G) architecture, including a hierarchical feature synthesis module from a prior study [13]. We modified the first layer of the generator network to have a pixel-unshuffle layer instead of a spatial average pooling layer, to reduce the dimensions of the input image to a quarter. The pixel unshuffle is the reverse operation of the pixel-shuffle [15], which helps computational efficiency by reducing the input dimension without pixel loss. We employed a U-Net discriminator (D) for spectral normalization [16]. The U-Net discriminator network was proposed for the image super-resolution task and can provide per-pixel feedback to the generator. The proposed GAN architecture is illustrated in

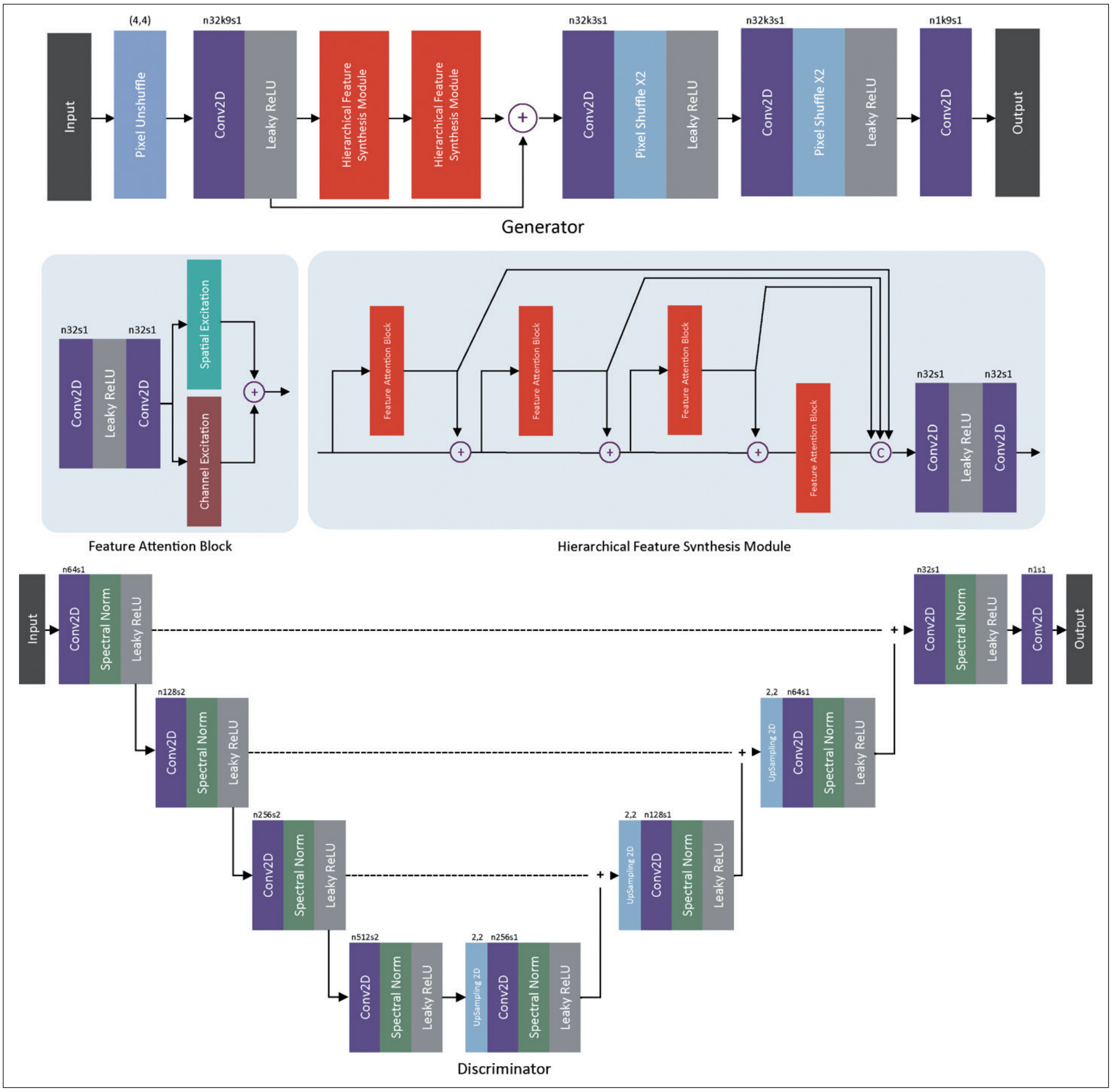

Fig. 1. Architecture of the generator and discriminator. In the generator, the pixel unshuffle is the reverse operation of the pixel shuffle to reduce the input dimension without pixel loss. The discriminator is a U-Net architecture with spectral normalization. Where n is the number of output feature maps of the convolution, k is the convolutional kernel size, and s is the stride of the convolution along the height and width. Conv = convolution, Norm = normalization

Figure 1. We employed a Real-ESRGAN training strategy [16], which is divided into two parts. First, we train a generator network without a discriminator network. Therefore, the generator network was trained with an L1 loss between the generated image and the ground-truth image and exhibited a peak signal-to-noise ratio (PSNR)-oriented performance. The equation for the L1 loss is

$$L_{l1} = \frac{1}{hw} \|G(x)-x\|_F^1 \qquad \text{Eq. (1)}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and h and w denote the height and width of the 2D image, respectively. x is the given source image and G(x) is the generated image. Second, the PSNR-oriented model was used for initialization and then trained using a discriminator network. A combination of L1 loss, perceptual loss [17], and GAN loss [18,19] was used to train the generator network. The perceptual loss equation is as follows:

$$L_{perceptual} = \frac{1}{hwd} \|\phi(G(x)) - \phi(x)\|_F^2 \qquad \text{Eq. (2)}$$

where ϕ is the feature extractor, and h, w, and d represent the height, width, and depth of the feature space, respectively. The feature extractor is the 16th convolution layer of the well-known pre-trained VGG-19 network [20]. The GAN loss is defined as follows:

$$L_{GAN} = \mathbb{E}_{x\sim p_{data}(x)}[-\log D(G(x))] \qquad \text{Eq. (3)}$$

where $\mathbb{E}(\cdot)$ is the expectation operator and p_{data} denotes the probability distributions of the source image. Combining equations Eqs. (1-3), the total loss function is obtained as follows:

$$L_{total} = \lambda_1 \times L_{l1} + \lambda_2 \times L_{perceptual} + \lambda_3 \times L_{GAN} \qquad \text{Eq. (4)}$$

where λ_1, λ_2, and λ_3 are the parameters used to adjust the balance of L_{l1}, $L_{perceptual}$,and L_{GAN} in L_{total}. In this study, we set $\lambda_1 = 10$, $\lambda_2 = 1$, and $\lambda_3 = 1$ in equation Eq. (4).

Training of the Deep Neural Network for Image Standardization

In this study, the network input comprised six different reconstructed images: FBP, IR, M40, M60, M80, and OPT, and the target was M80. We randomly sampled a 256 x 256 local patch from the same location in all the reconstructed images. Min-max normalization was applied to rescale the CT values to "0, 1". A local patch was used as an input to the generator network, and the generated image had the same size. The discriminator network provides pixel-wise feedback over the local patch. All parameters of the generator and discriminator networks were optimized using an adaptive moment estimation optimizer (hyper-parameters α = 1 x [10]^(-4), β_1 = 0.9, and β_2 = 0.99) [21]. The networks were trained for 200 epochs, and the learning rate decayed by 95% of the decay rate after 100 epochs. We implemented this deep learning network using PyTorch 1.10. All experiments were performed on a personal computer (Intel i7 9770 [Intel] with 32 GB of memory) and accelerated using an NVIDIA RTX 2080 Ti GPU (NVIDIA) with 11 GB of memory.

Liver Segmentation

We used a previously trained 2D U-NET in a commercially available segmentation software program (MEDIP PRO v2.0.0.0, MEDICALIP Co. Ltd.) to create liver segmentation masks and calculated the liver volume for the test dataset. A pre-test was performed with five CT examinations from the training dataset to assess the performance of segmentation in various protocols. According to the pre-test results, we selected the protocol with the highest segmentation performance among the different images as the ground truth for image conversion. Two radiologists assessed the liver segmentation masks for all protocols, and the original M80 protocol, which showed nearly perfect segmentation, was selected as the ground truth image (Supplementary Fig. 1).

Evaluation of Liver Segmentation Performance

To evaluate the performance of the automated segmentation, the masks that were generated on each protocol image were compared with the reference segmentation mask of the ground truth image (i.e., originalM80 images). We calculated the Dice similarity coefficients (DSCs) by comparing two segmentation masks. Segmentation performance was obtained for the original CT images and standardized CT images that were synthesized using the deep learning algorithm for image standardization. Moreover, we calculated the absolute difference between the segmented volume and ground truth liver volume (i.e., the original M80 images) as well as the ratio of the liver volume difference to the ground truth volume.

Statistical Analyses

We used a paired *t*-test to compare the segmentation performance in DSC before and after the image conversion using the deep learning-based standardization algorithm. We also compared the absolute volume difference from the ground truth volume and the ratio of the liver volume difference to the ground truth volume between the original and standardized images using a paired *t*-test. Agreement between the segmented liver volume and ground truth volume was analyzed using concordance correlation coefficients (CCCs) [22] and Bland–Altman analysis.

RESULTS

Patient Demographics

For training and tuning, we used 142 CT examinations (128 and 14, respectively) of 117 patients. For testing, 43 CT scans from 42 patients were included in the test dataset. The mean age of the test population was 10.1 years (range, 7–49 years). Detailed demographics of the test datasets are summarized in Table 1.

Table 2. Comparison of Dice Similarity Coefficients between the Original and Standardized Images in Various Protocols

	DSC (%)		
	Original	Standardized	*P*
FBP	90.79 ± 5.57	96.74 ± 2.43	< 0.001
IR	91.27 ± 5.11	96.57 ± 2.45	< 0.001
M40	5.40 ± 18.06	93.16 ± 4.19	< 0.001
M60	87.55 ± 7.91	95.88 ± 2.61	< 0.001
OPT	83.63 ± 12.50	96.27 ± 2.33	< 0.001

Values represent the mean ± standard deviation. DSC = Dice similarity coefficient, FBP = filtered back projection, IR = iterative reconstruction, M40 = mono-energy 40 keV, M60 = mono-energy 60 keV, OPT = optimal contrast

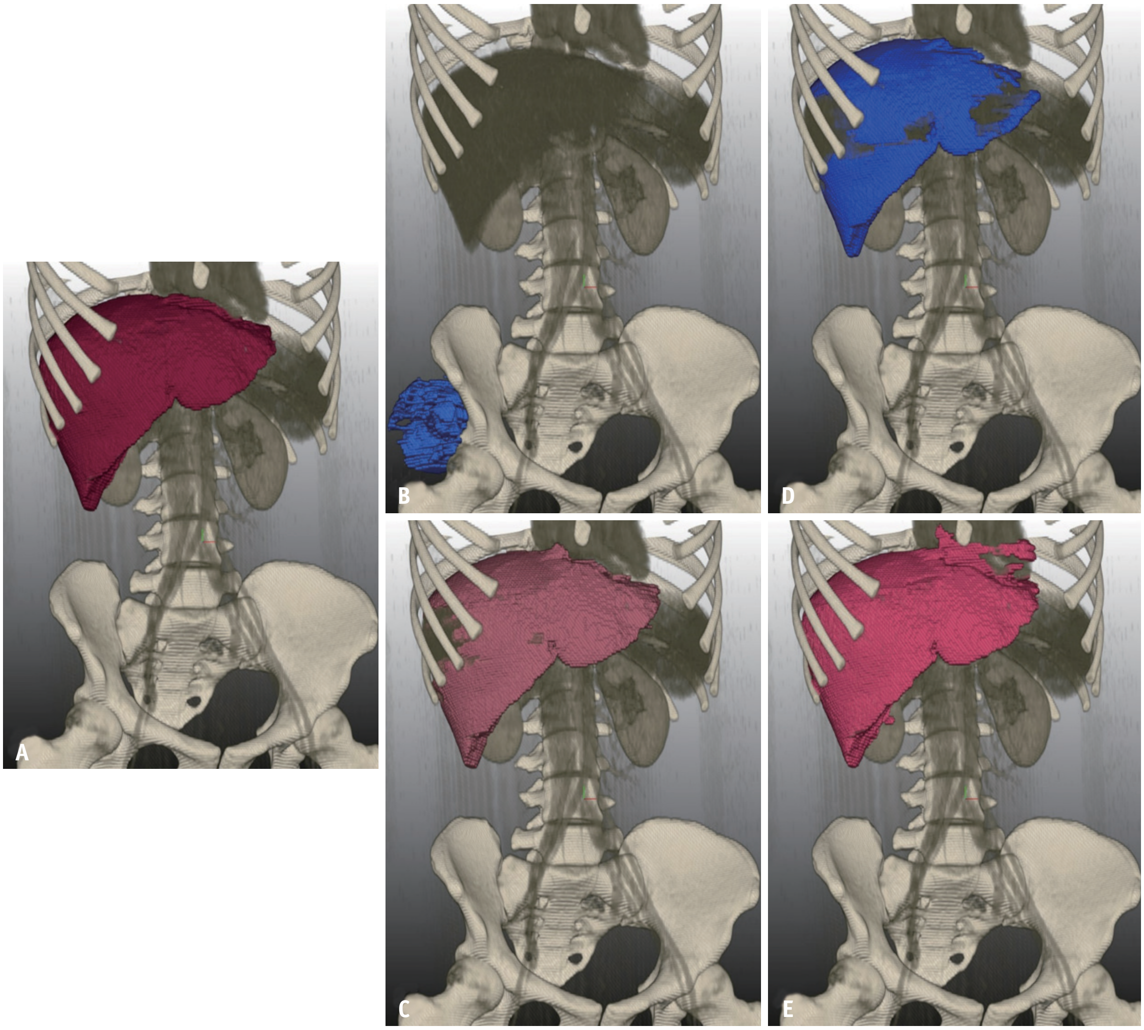

Fig. 2. A representative case of automated liver segmentation of a 14-year-old girl with various protocols. The images show the three-dimensional volume rendering of the liver segmentation. **A:** Liver segmentation at ground truth. **B:** Liver segmentation at a mono-energy of 40 keV (M40). **C:** Liver segmentation using standardized M40. **D:** Liver segmentation at a mono-energy of 60 keV (M60). **E:** Liver segmentation using standardized M60. All standardized images exhibited better segmentation performance than the original images.

Comparison of Segmentation Performance

The DSCs of the automated liver segmentation in the original and standardized images are summarized in Table 2. According to the protocols, the original images showed variable and poor segmentation performances for the liver. The standardized images that were obtained using deep neural networks achieved significantly higher DSCs for liver segmentation than the original images (all $P < 0.001$). M40 had the lowest DSC in the original images, and the largest performance increase with the image standardization. Representative images are shown in Figures 2 and 3.

Comparison of Liver Volume

Compared with the ground truth liver volume, the absolute difference and difference ratio of the liver volume in the original and standardized images are presented in Table 3. The absolute difference in liver volume decreased after image conversion (all $P < 0.001$). The difference ratio of the liver volume also decreased significantly after the image conversion (all $P < 0.001$).

Agreement of Segmented Liver Volume with Ground Truth Volume

The agreement of the segmented liver volume results with the ground truth images, in terms of CCC, for the original images and standardized CT images in various protocols is summarized in Table 4. In all protocols, CCCs improved after the image conversion. In the original images, M40, M60,

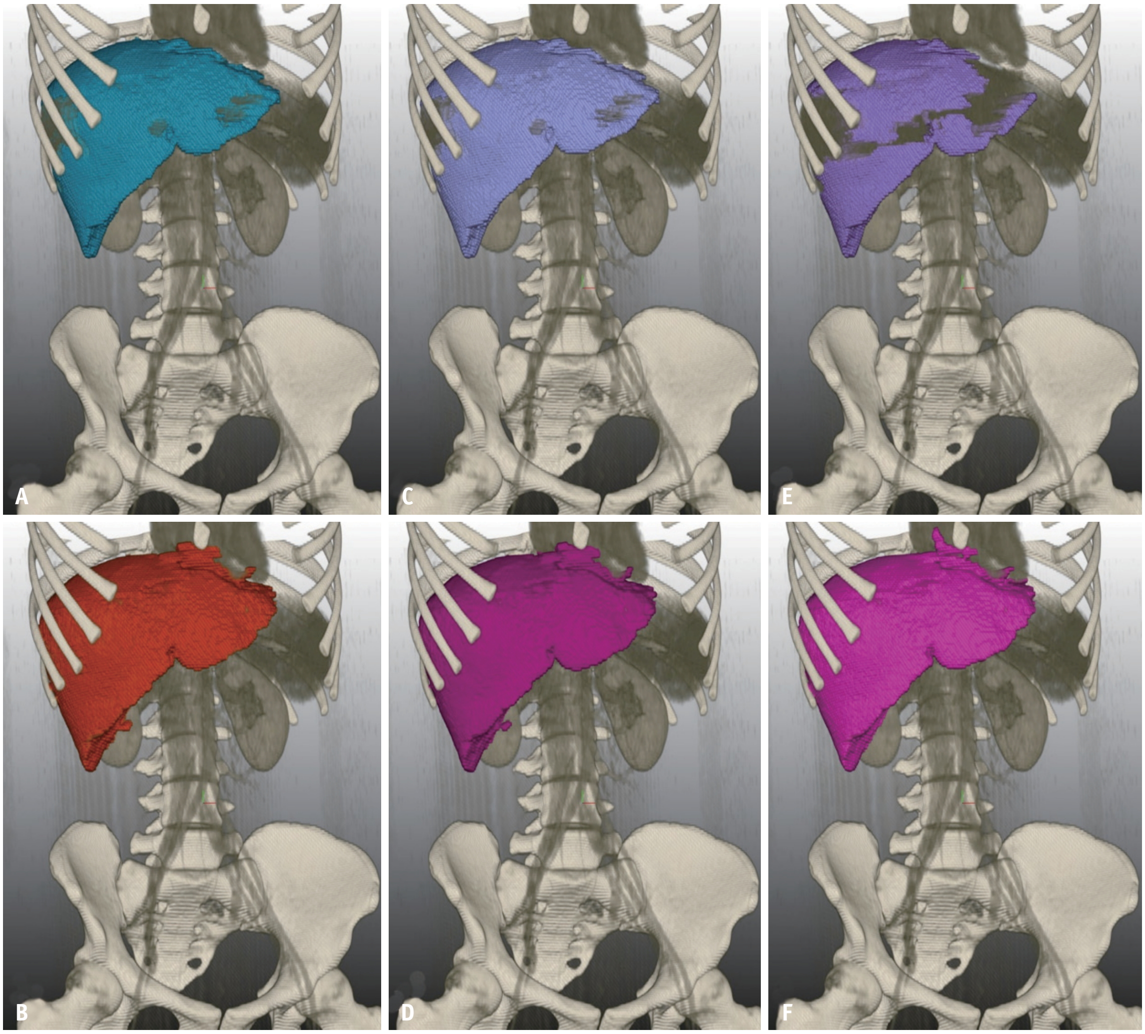

Fig. 3. A representative case of automated liver segmentation of a 14-year-old girl with various protocols. The images show the three-dimensional volume rendering of the liver segmentation. **A:** Liver segmentation using filtered back projection (FBP). **B:** Liver segmentation using standardized FBP. **C:** Liver segmentation using hybrid iterative reconstruction (IR). **D:** Liver segmentation using standardized IR. **E:** Liver segmentation with optimal contrast (OPT). **F:** Liver segmentation using standardized OPT. All standardized images exhibited better segmentation performance than the original images.

and OPT showed poor agreement. After image conversion, all protocols showed almost perfect agreement.

Figures 4 and 5 show the Bland–Altman plot of the liver volume difference ratio against the ground truth volume. For FBP, IR, and OPT, prior to applying image conversion, the plot exhibited a negative bias (mean difference: FBP -11.5%, IR -9.8%, OPT -21.5%) with wide 95% limits of agreement (LOAs) (FBP -29.2%–6.2%, IR -26.3%–6.7%, OPT -58.0%–15.0%), and there was no systematic trend in bias with liver volume at ground truth. After image conversion,

Table 4. Concordance Correlation Coefficients of Liver Volume in the Original and Standardized Images with the Ground-Truth Images in Various Protocols

	Original	Standardized
FBP	0.955 (0.923–0.973)	0.998 (0.996–0.999)
IR	0.964 (0.939–0.979)	0.998 (0.996–0.999)
M40	-0.006 (-0.046–0.035)	0.990 (0.983–0.995)
M60	0.031 (0.018–0.044)	0.995 (0.991–0.997)
OPT	0.809 (0.696–0.883)	0.996 (0.993–0.998)

Values in parentheses are 95% confidence interval. FBP = filtered back projection, IR = iterative reconstruction, M40 = monoenergy 40 keV, M60 = mono-energy 60 keV, OPT = optimal contrast

Table 3. Comparison of Liver Volume between Original and Standardized Images Using Various Protocols

	Absolute Liver Volume Difference (mL)			Difference Ratio of Liver Volume (%)		
	Original vs. GT	Standardized vs. GT	*P*	Original vs. GT	Standardized vs. GT	*P*
FBP	949.7 ± 784.1	154.0 ± 223.1	< 0.001	11.5 ± 9.0	2.0 ± 2.9	< 0.001
IR	808.2 ± 732.7	159.3 ± 208.0	< 0.001	9.8 ± 8.4	2.0 ± 2.8	< 0.001
M40	8684.8 ± 4371.7	356.8 ± 414.1	< 0.001	91.0 ± 19.3	4.4 ± 5.1	< 0.001
M60	8587.8 ± 3652.5	324.7 ± 222.4	< 0.001	91.4 ± 1.4	4.3 ± 3.8	< 0.001
OPT	1917.9 ± 1789.5	192.8 ± 299.8	< 0.001	21.6 ± 18.4	2.4 ± 3.3	< 0.001

Values represent the mean ± standard deviation. FBP = filtered back projection, IR = iterative reconstruction, M40 = monoenergy 40 keV, M60 = mono-energy 60 keV, OPT = optimal contrast, GT = ground truth

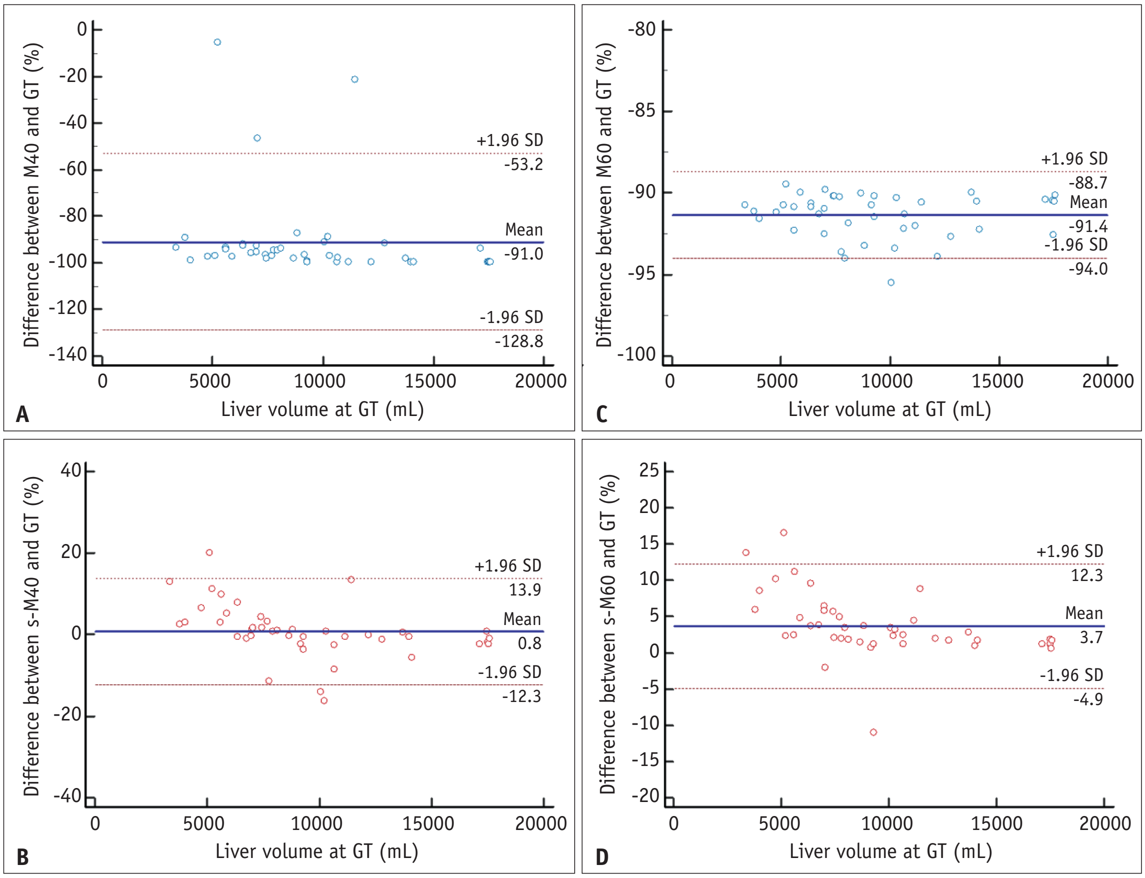

Fig. 4. Bland–Altman curve of liver volume between various protocols and ground truth. **A:** Bland–Altman plot between mono-energy 40 keV (M40) and ground truth (GT) (mean difference = -91.0%, 95% limits of agreement [LOA] = -128.8%–-53.2%). **B:** Bland–Altman plot between standardized M40 and GT (mean difference = 0.8%, 95% LOA = -12.83%–13.9%). **C:** Bland–Altman plot between mono-energy 60 keV (M60) and GT (mean difference = -91.4%, 95% LOA = -94.0%–-88.7%). **D:** Bland–Altman plot between standardized M60 and GT (mean difference = 3.7%, 95% LOA = -4.9%–12.3%). s = standardized, SD = standard deviation

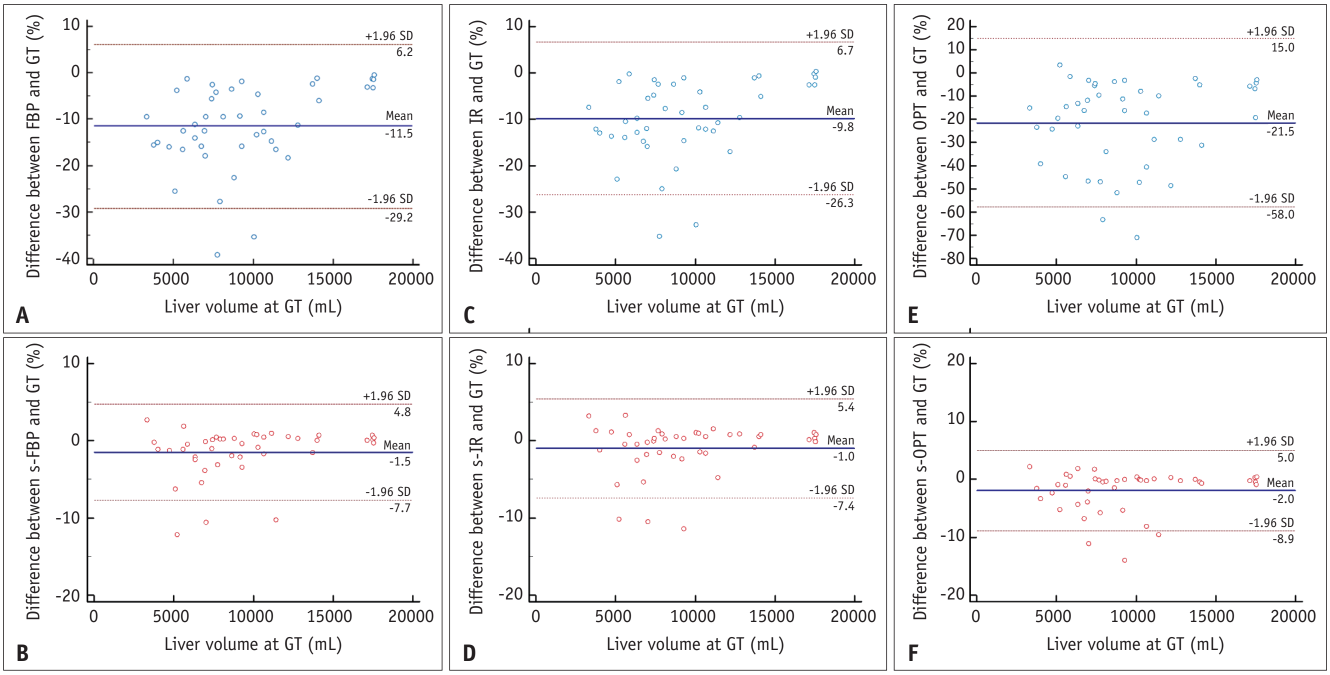

Fig. 5. Bland–Altman curve of liver volume between various protocols and ground truth. **A:** Bland–Altman plot between filtered back projection (FBP) and ground truth (GT) (mean difference = -11.5%, 95% limits of agreement [LOA] = -29.2%–6.2%). **B:** Bland–Altman plot between standardized FBP and GT (mean difference = -1.5%, 95% LOA = -7.7%–4.8%). **C:** Bland–Altman plot between hybrid iterative reconstruction (IR) and GT (mean difference = -9.8%, 95% LOA = -26.3%–6.7%). **D:** Bland–Altman plot between standardized IR and GT (mean difference = -1.0%, 95% LOA = -7.4%–5.4%). **E:** Bland–Altman plot between optimal contrast (OPT) and GT (mean difference = -21.5%, 95% LOA = -58.0%–15.0%). **F:** Bland–Altman plot between standardized OPT and GT (mean difference = -2.0%, 95% LOA = -8.9%–-5.0%). s- = standardized, SD = standard deviation

the plot showed a small negative bias (mean difference: FBP -1.5%, IR -1.0%, OPT -2.0%) with narrow LOAs (FBP -7.7% – 4.8%, IR -7.4%–5.4%, OPT -8.9%–5.0%). For M40, the original image exhibited a large negative bias (mean difference, -91.0%) with a wide LOA (-128.8% – -53.2%). In contrast, the standardized image showed a small positive bias (0.8%) with a narrow LOA (-12.3%–13.9%). The original M60 images showed a narrow LOA (-94.0% – -88.7%) but had the largest negative bias (mean difference, -91.4%). After image conversion, the plot exhibited a small positive bias (mean difference, 3.7%) with a narrow LOA (-4.9%–12.3%).

DISCUSSION

This study demonstrated that deep learning-based image analysis tools without training data augmentation showed poor performance when medical images that were different from the training dataset were used as input data. Liver segmentation performance showed variable and poor performance according to the image protocols, despite the data being obtained from the same patient simultaneously. After image conversion using the developed deep neural network algorithms, organ segmentation performance was significantly improved.

AI is a rapidly growing field in medical imaging. A considerable number of articles submitted and published in the field of radiology are related to AI, and many AI-based medical image analysis software programs are being developed [23]. As the medical imaging volume increased, radiologist loading increased. The application of AI to medical imaging helps various tasks of radiologists [1]. Image analysis using AI extends beyond lesion detection and enables lesion classification, treatment response prediction, image conversion, synthetic image generation, and various quantitative analyses. AI helps improve accuracy and efficiency in the detection of lesions and reduces measurement and perceptual errors [24-26].

AI is a promising technology for medical imaging; however, there are several obstacles to its application in daily clinical practice. The biggest limitation in AI models that can be applied to clinical practice is securing generalizability [11,27]. Developed and tested deep learning-based algorithms, showing excellent performance in external tests, may demonstrate degraded performance in real clinical practice [10,28-30]. In this study, liver segmentation performance using automated deep learning-based algorithms showed variable results across different reconstruction methods. The lower the similarity with the ground truth image, the lower the segmentation performance. In the training of the deep learning-based algorithm, image characteristics, including image noise, contrast, artifacts, and texture, were trained. In CT images, various CT parameters (e.g., tube voltage or tube current), patient body size, reconstruction method, machine, and vendor type affect the image characteristics. Deploying deep learning-based algorithms to individual institutions in real clinical practice encounters new datasets that are not identical to the training dataset. This heterogeneity of medical images from individual institutions degrades the performance of deep learning-based algorithms. A few studies have attempted to improve the generalizability of deep learning-based algorithms with data augmentation [10,11,31,32]. Rauschecker et al. [10] used a limited additional local training dataset to overcome generalization issues. The major advantage of transfer learning is that it can improve the performance of the developed algorithm using only a modest amount of data that is acquired by the institution that uses the developed algorithm. Transfer learning may be a solution for applying the algorithm to external institutions. However, this method has several limitations. Additional data augmentation is required for each institution to use this algorithm. Additional transfer learning is required whenever the environment for the acquisition of medical images is changed (e.g., changing the imaging protocols or CT machine). Moreover, there are no specific guidelines for the number of local datasets that are required to restore the internally tested performance of the algorithm.

Eche et al. [11] attempted to secure generalizability using computational stress testing. Generating various datasets by artificially modifying images can improve the generalizability of stress tests. In radiology, various image datasets can be generated by adjusting image noise, contrast, section thickness, or artifacts. Training and validation using a large heterogeneous dataset can improve generalizability. In the case of a model undergoing stress testing to obtain generalizability, the average performance of several test datasets may be high, but the performance of a specific dataset that is evenly distributed with the training set can be rather low [11].

However, image conversion and standardization have several advantages over the other methods. First, it can be a useful means of securing the generalizability of deep learning models regardless of the application environment. If the converted image data are close to the training set, they can show ideal performance regardless of the CT machine, vendor, and reconstruction method that is used in individual institutions. Second, there is no need to collect the dataset for transfer learning or have a stress test dataset, which saves time and effort for the deployment of the algorithm.

Our study had some limitations. First, the data used as the ground truth for image conversion were not used to train the segmentation algorithm, which may be the reason for the slightly low segmentation performance. However, when we evaluated each segmentation mask from the ground truth, no substantial errors were observed (e.g., the segmentation of other organs). Second, we used a single CT machine for the training and validation. Other CT machines and vendors should be evaluated to expand the use of the developed model in clinical practice.

In conclusion, we suggest that deep learning-based CT image standardization can improve the performance of automated segmentation of the liver using CT images that are reconstructed with various methods. Deep learning-based CT image conversion may have the potential to improve the generalizability of the segmentation network.

Supplement

The Supplement is available with this article at https://doi.org/10.3348/kjr.2022.0588.

Availability of Data and Material

The datasets generated or analyzed during the study are available from the corresponding author on reasonable request.

Conflicts of Interest

Soon Ho Yoon works as a chief medical officer in MEDICAL IP and has a stock option in the firm, outside the present study. Young Hun Choi and Jung-Eun Cheon, contributing editors of the *Korean Journal of Radiology*, was not involved in the editorial evaluation or decision to publish this article. Other remaining authors have declared no conflicts of interest.

Author Contributions

Conceptualization: Yeon Jin Cho. Data curation: Seul

Bi Lee. Formal analysis: Seul Bi Lee, Youngtaek Hong, Dawun Jeong, Jina Lee. Funding acquisition: Yeon Jin Cho. Investigation: Seul Bi Lee, Youngtaek Hong, Dawun Jeong, Jina Lee. Methodology: Yeon Jin Cho. Project administration: Youngtaek Hong. Software: Soon Ho Yoon. Supervision: Yeon Jin Cho, Seunghyun Lee, Young Hun Choi, Jung-Eun Cheon. Validation: Youngtaek Hong, Dawun Jeong, Jina Lee. Visualization: Seul Bi Lee, Dawun Jeong, Jina Lee. Writing—original draft: Seul Bi Lee, Youngtaek Hong. Writing—review & editing: Yeon Jin Cho, Soon Ho Yoon, Seunghyun Lee, Young Hun Choi, Jung-Eun Cheon.

ORCID iDs

Seul Bi Lee
https://orcid.org/0000-0002-5163-3911
Youngtaek Hong
https://orcid.org/0000-0003-2104-5905
Yeon Jin Cho
https://orcid.org/0000-0001-9820-3030
Dawun Jeong
https://orcid.org/0000-0001-9791-9555
Jina Lee
https://orcid.org/0000-0003-1395-5474
Soon Ho Yoon
https://orcid.org/0000-0002-3700-0165
Seunghyun Lee
https://orcid.org/0000-0003-1858-0640
Young Hun Choi
https://orcid.org/0000-0002-1842-9062
Jung-Eun Cheon
https://orcid.org/0000-0003-1479-2064

Funding Statement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (the Ministry of Science and ICT) (No. NRF-2020R1F1A1077320).

REFERENCES

1. Cheng PM, Montagnon E, Yamashita R, Pan I, Cadrin-Chênevert A, Perdigón Romero F, et al. Deep learning: an update for radiologists. *Radiographics* 2021;41:1427-1445
2. McBee MP, Awan OA, Colucci AT, Ghobadi CW, Kadom N, Kansagra AP, et al. Deep learning in radiology. *Acad Radiol* 2018;25:1472-1480
3. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med* 2018;15:e1002686
4. Joo B, Ahn SS, Yoon PH, Bae S, Sohn B, Lee YE, et al. A deep learning algorithm may automate intracranial aneurysm detection on MR angiography with high diagnostic performance. *Eur Radiol* 2020;30:5785-5793
5. Kuo W, Häne C, Mukherjee P, Malik J, Yuh EL. Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning. *Proc Natl Acad Sci U S A* 2019;116:22737-22745
6. Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing medical imaging data for machine learning. *Radiology* 2020;295:4-15
7. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. *Nat Med* 2019;25:24-29
8. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 2018;286:800-809
9. Candemir S, Nguyen XV, Folio LR, Prevedello LM. Training strategies for radiology deep learning models in data-limited scenarios. *Radiol Artif Intell* 2021;3:e210014
10. Rauschecker AM, Gleason TJ, Nedelec P, Duong MT, Weiss DA, Calabrese E, et al. Interinstitutional portability of a deep learning brain MRI lesion segmentation algorithm. *Radiol Artif Intell* 2021;4:e200152
11. Eche T, Schwartz LH, Mokrane FZ, Dercle L. Toward generalizability in the deployment of artificial intelligence in radiology: role of computation stress testing to overcome underspecification. *Radiol Artif Intell* 2021;3:e210097
12. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. *Radiol Artif Intell* 2022;4:e210064
13. Lee SB, Cho YJ, Hong Y, Jeong D, Lee J, Kim SH, et al. Deep learning-based image conversion improves the reproducibility of computed tomography radiomics features: a phantom study. *Invest Radiol* 2022;57:308-317
14. Chen C, Bai W, Davies RH, Bhuva AN, Manisty CH, Augusto JB, et al. Improving the generalizability of convolutional neural network-based segmentation on CMR images. *Front Cardiovasc Med* 2020;7:105
15. Shi WZ, Caballero J, Huszar F, Totz J, Aitken AP, Bishop R, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network.com Web site. https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Shi_Real-Time_Single_Image_CVPR_2016_paper.pdf. Published September 23, 2016. Accessed December 20, 2021
16. Wang X, Xie L, Dong C, Shan Y. Real-ESRGAN: training real-world blind super-resolution with pure synthetic data.com Web site. https://openaccess.thecvf.com/content/ICCV2021W/AIM/papers/Wang_Real-ESRGAN_Training_Real-World_Blind_Super-Resolution_With_Pure_Synthetic_Data_ICCVW_2021_paper.pdf. Published August 17, 2021. Accessed December 20,

2021

17. Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution.com Web site. https://doi.org/10.1007/978-3-319-46475-6_43. Published March 27, 2016. Accessed December 20, 2021
18. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets.com Web site. https://papers.nips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf. Published June 10, 2014. Accessed December 20, 2021
19. Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, et al. Photo-realistic single image super-resolution using a generative adversarial network.com Web site. https://openaccess.thecvf.com/content_cvpr_2017/papers/Ledig_Photo-Realistic_Single_Image_CVPR_2017_paper.pdf. Published May 25, 2017. Accessed December 20, 2021
20. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 [Preprint]. [posted September 4, 2014; revised April 10, 2015; cited December 22, 2021] https://arxiv.org/abs/1409.1556
21. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv:1412.6980 [Preprint]. [posted December 22, 2014; revised January 30, 2017; cited December 22, 2021] https://doi.org/10.48550/arXiv.1412.6980
22. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989;45:255-268
23. Bluemke DA, Moy L, Bredella MA, Ertl-Wagner BB, Fowler KJ, Goh VJ, et al. Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers-from the radiology editorial board. *Radiology* 2020;294:487-489
24. van Winkel SL, Rodriguez-Ruiz A, Appelman L, Gubern-Mérida A, Karssemeijer N, Teuwen J, et al. Impact of artificial intelligence support on accuracy and reading time in breast tomosynthesis image interpretation: a multi-reader multi-case study. *Eur Radiol* 2021;31:8682-8691
25. Quon JL, Han M, Kim LH, Koran ME, Chen LC, Lee EH, et al. Artificial intelligence for automatic cerebral ventricle segmentation and volume calculation: a clinical tool for the evaluation of pediatric hydrocephalus. *J Neurosurg Pediatr* 2020;27:131-138
26. Winkel DJ, Wetterauer C, Matthias MO, Lou B, Shi B, Kamen A, et al. Autonomous detection and classification of PI-RADS lesions in an MRI screening population incorporating multicenter-labeled deep learning and biparametric imaging: proof of concept. *Diagnostics (Basel)* 2020;10:951
27. Kawaguchi K, Kaelbling LP, Bengio Y. Generalization in deep learning. arXiv:1710.05468 [Preprint]. [posted October 16, 2017; revised December 11, 2017; cited December 22, 2021] https://doi.org/10.48550/arXiv.1710.05468
28. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health* 2020;2:e489-e492
29. Bhuva AN, Bai W, Lau C, Davies RH, Ye Y, Bulluck H, et al. A multicenter, scan-rescan, human and machine learning CMR study to test generalizability and precision in imaging biomarker analysis. *Circ Cardiovasc Imaging* 2019;12:e009214
30. Onofrey JA, Casetti-Dinescu DI, Lauritzen AD, Sarkar S, Venkataraman R, Fan RE, et al. Generalizable multi-site training and testing of deep neural networks using image normalization. *Proc IEEE Int Symp Biomed Imaging* 2019;2019:348-351
31. Sanford TH, Zhang L, Harmon SA, Sackett J, Yang D, Roth H, et al. Data augmentation and transfer learning to improve generalizability of an automated prostate segmentation model. *AJR Am J Roentgenol* 2020;215:1403-1410
32. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data* 2019;6:1-48