RESEARCH ARTICLE

Improving the Reproducibility of Computed Tomography Radiomic Features Using an Enhanced Hierarchical Feature Synthesis Network

DAWUN JEONG[1,2], **YOUNGTAEK HONG**[2,3], **(Member, IEEE), JINA LEE**[1,2], **SEUL BI LEE**[4], **YEON JIN CHO**[4], **HACKJOON SHIM**[2], **AND HYUK-JAE CHANG**[2,3,5]

[1]Brain Korea 21 Project, Graduate School of Medical Science, Yonsei University College of Medicine, Seoul 03722, South Korea
[2]CONNECT-AI Research Center, Yonsei University College of Medicine, Seoul 03722, South Korea
[3]Ontact Health Inc., Seoul 07332, South Korea
[4]Department of Radiology, Seoul National University Hospital, Jongno-gu, Seoul 03080, Republic of Korea
[5]Division of Cardiology, Severance Cardiovascular Hospital, Yonsei University College of Medicine, Yonsei University Health System, Seoul 03722, South Korea

Corresponding author: Youngtaek Hong (hyt0205@gmail.com)

ABSTRACT Radiomics has gained popularity as a quantitative analysis method for medical images. However, computed tomography (CT) scans are performed using various parameters, such as X-ray dose and reconstruction kernels, which is a fundamental reason for the lack of reproducibility of radiomic features. This study evaluated whether the proposed network improves the reproducibility of radiomic features across various CT protocols and reconstruction kernels. We set five CT scan protocols and two reconstruction kernels to create various noise settings for the obtained CT images with an abdominal phantom. We developed an enhanced hierarchical feature synthesis (EHFS) network to improve the reproducibility of radiomic features across various CT protocols and reconstruction kernels. Eight hundred and nineteen radiomic features were extracted, including first-order, second-order, and wavelet features. Reproducibility was assessed using Lin's concordance correlation coefficient (CCC) on internal and external testing. We considered a radiomic feature with CCC $\geq$ 0.85 as a high-agreement feature. As a result, the average number of reproducible features increased in all protocols, from 241 ± 38 to 565 ± 11 in internal testing. In external testing, consisting of a new phantom and unseen protocol, 239 ± 74 reproducible features were in source images and 324 ± 16 were in generated images. The EHFS network is a novel approach to improving the reproducibility of radiomic features. It outperforms existing methods in reproducibility and generalization, as demonstrated by comprehensive experiments on both internal and external datasets. Our deep-learning-based CT image conversion could be a solution for standardization in ongoing radiomics research.

INDEX TERMS Artificial intelligence, computed tomography, radiomics, reproducibility.

I. INTRODUCTION

Classical medical image analysis relies on radiologists' perceptual and qualitative assessment; however, quantitative image analysis has recently gained popularity. Radiomics is

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

a quantitative image analysis method that extracts numerous features from radiological images [1]. Representative radiomic features can provide a tissue phenotype that can be utilized for clinical decision-making [1], [2]; however, its reproducibility remains a significant challenge. Berenguer et al. [3] showed that many radiomic features might be non-reproducible and redundant. Computed

tomography (CT) scans are performed using various parameters, such as X-ray dose and reconstruction kernels. The scan parameters are determined based on the imaging purpose and patient characteristics, which is a fundamental reason for the lack of reproducibility. The full potential of radiomic features can be realized if harmonization methods minimize unwanted variations induced by differences in scan parameters and enhance their reproducibility. The ComBat method [4], [5], a widely used harmonization approach, directly harmonizes radiomic features that have already been extracted considering different distributions of multi-center, multi-protocol datasets. While ComBat is effective in removing cohort bias and preserving differences in phantom radiomic features, it requires known batch covariates and struggles with computational complexity demands when applied to images or high-dimensional signals [6]. Considering this, we focused on harmonizing raw data sets instead of data-derived features.

Factors such as noise, blur, and artifacts can negatively affect radiomic features. Therefore, it is imperative that images used for radiomic analysis are high-quality and have minimal artifacts [7], [8]. Many studies have reported on improving radiomics reproducibility by utilizing architectures such as full convolutional networks (FCNs) [9] and generative adversarial networks (GANs) [10], which are widely used for removing noise in medical images. Choe et al. [11] proposed a CT image conversion method between two different kernels using a convolutional neural network (CNN) to improve radiomic reproducibility. De Farias et al. [12] confirmed the robustness of radiomic features through super-resolution images generated using the GAN-CIRCLE model and bicubic interpolation. Marcadent et al. [13] utilized Cycle-GAN to perform texture conversion to improve radiomic features reproducibility between manufacturers. Chen et al. [14] developed the CT denoising method to convert low-dose to high-dose CT images by using cycle GANs, improving the reproducibility of radiomic features. Selim et al. [15] used a pre-trained U-net as a generator and applied the window-based training approach to standardize images obtained from three non-standard reconstruction kernels. In contrast, this study focused on harmonizing CT scans, which converted nine CT protocols with five distinct CT scan protocols and two reconstruction kernels into one target protocol.

In our previous work [16], a hierarchical feature synthesis (HFS) network was proposed and demonstrated to improve the reproducibility of radiomic features. The HFS network was inspired by the residual feature aggregation (RFA) network [17]. Because the chain of the residual blocks creates complex fused features, the main idea behind RFA is to use a clean residual feature to reconstruct the target image. The RFA network has only a spatial attention module to rescale the features according to the spatial context, but convolutional features are naturally spatial and channel-wise. In HFS, channel and spatial attention modules were employed in parallel to improve the feature attention performance; however,

Woo et al. [18] demonstrated that the sequential application of channel and spatial attention modules has a better feature attention performance than parallel applications.

In this study, we developed an enhanced HFS (EHFS) network with sequential attention modules and performed comprehensive experiments to demonstrate that the improved reproducibility of radiomic features across various CT protocols and reconstruction kernels.

II. METHODS

This study aimed to explore a function G that converts source images x, scanned through various CT protocols and reconstruction kernels, into target CT images y.

$$G(x) \approx y, \tag{1}$$

The mapping function G cannot perfectly define the relationship between x and y, which can only output an image $G(x)$ close to y, similar to most end-to-end deep learning methods. A GAN consists of two neural networks: a generator (G) and a discriminator (D) [19]. In this study, G is an EHFS network, and the architecture of D is based on a photorealistic single-image super-resolution GAN [20].

The EHFS is an extension of the HFS network that adopts the sequential application of advanced channel and spatial attention modules, emphasizing feature representation for image conversion. The architectures of G and D are illustrated in Fig. 1.

A. GENERATOR NETWORK ARCHITECTURE

G is a fully convolutional network consisting of two-dimensional (2D) convolutional layers. The encoded features allow G to learn high-level features from a given image x using two RFA modules. The EHFS was enhanced by replacing the parallel application of the HFS channel and spatial attention modules with an advanced sequential application channel and spatial attention modules.

We adopted a convolutional block attention module (CBAM) [18] for the sequential application channel and spatial attention modules. CBAM has an adaptive enhancement effect by sequentially inferring attention maps by applying channel and spatial attention. In the feature attention block, the convolutional feature map $F \in \mathbb{R}^{C \times H \times W}$ sequentially passes a 1D channel attention $A_c \in \mathbb{R}^{C \times 1 \times 1}$ and a 2D spatial attention $A_s \in \mathbb{R}^{1 \times H \times W}$. The channel attention is composed of two different spatial context descriptors average pooling and max pooling: F_{avg}^C and F_{max}^C. The channel attention is computed as:

$$\begin{aligned} A_c(F) &= \varphi(MLP(AvgPool(F)) \\ &\quad + MLP(MaxPool(F))) \\ &= \varphi(W_1 * W_0 * F_{avg}^C \\ &\quad + W_1 * W_0 * F_{max}^C), \end{aligned} \tag{2}$$

where φ denotes the sigmoid function. The multi-layer perceptron (MLP) weights $W_0 \in \mathbb{R}^{C/r \times C}$, and $W_1 \in \mathbb{R}^{C \times C/r}$
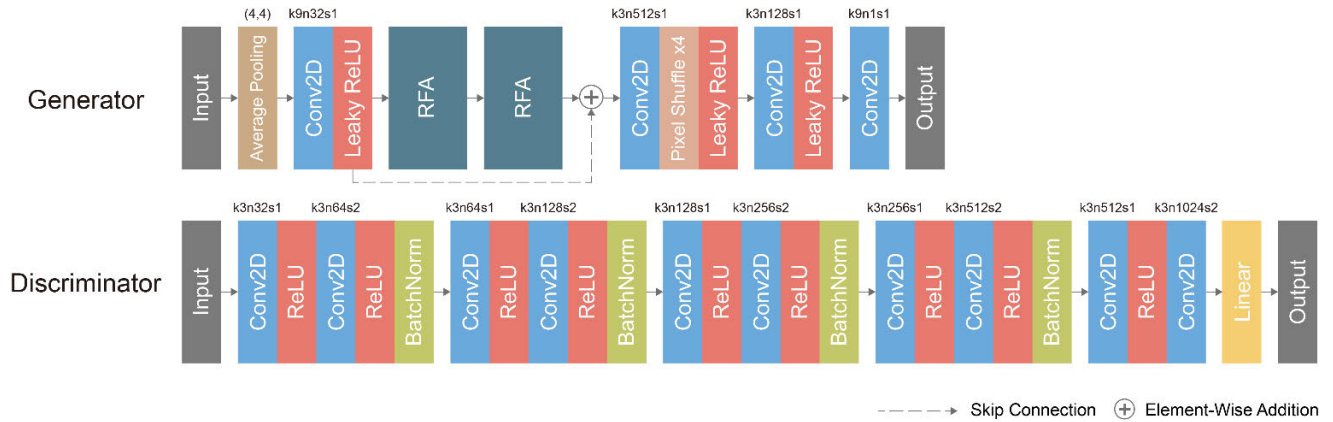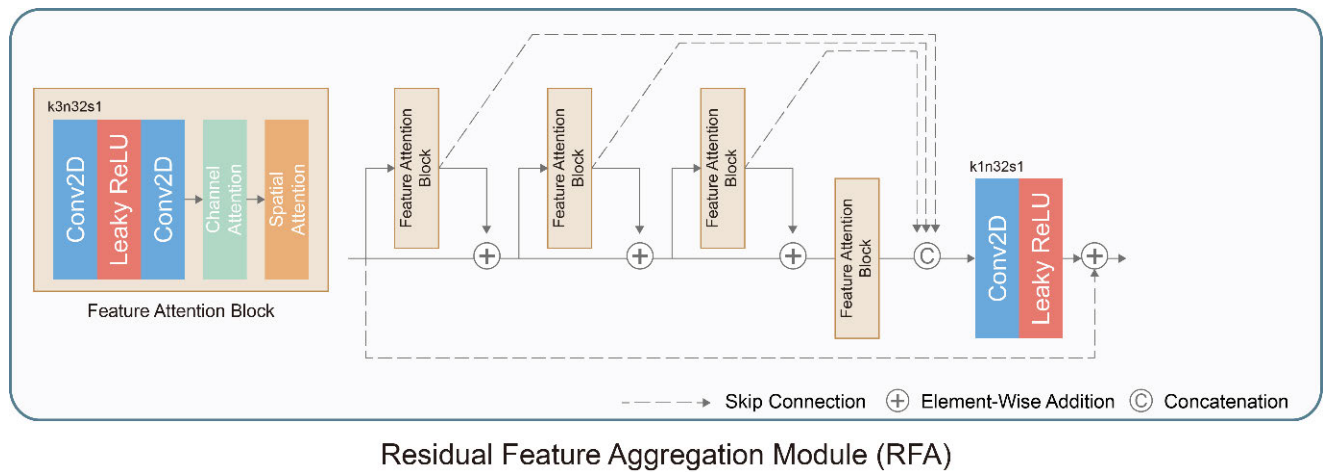
FIGURE 1. Architecture of the proposed network with corresponding kernel size (k), number of feature maps (n), and stride (s) indicated for the convolutional layer.

FIGURE 2. Structure of residual feature aggregation module and feature attention block.

are shared for both F_{avg}^C and F_{max}^C, where r is the feature dimension reduction ratio. In this study, we set $c = 32$ and $r = 16$. The spatial attention is composed of two pooling operations average pooling and max pooling: F_{avg}^S and F_{max}^S. The spatial attention is computed as follows:

$$A_s(F) = \varphi\left(K^{7\times7} * ([AvgPool(F); MaxPool(F)])\right)$$
$$= \varphi\left(K^{7\times7} * \left(\left[F_{avg}^S; F_{max}^S\right]\right)\right), \quad (3)$$

where φ denotes the sigmoid function. $K^{7\times7}$ denotes a 7×7 convolutional kernel, and $*$ denotes the convolutional operation. The overall attention process can be summarized as follows:

$$F' = A_c(F) \otimes F,$$
$$F'' = A_s(F') \otimes F', \quad (4)$$

where $\otimes$ denotes element-wise multiplication. The channel and spatial attention block and RFA module are shown in Fig. 2.

To reduce the dimensions of the input image to a quarter, we placed a spatial average pooling layer as the first layer

of G. The downsampled image was then passed through a convolutional layer with a size of 9×9 kernel and activated through the Leaky ReLU [21]. The 2D convolutional layer had 3×3 kernels, and zero padding was applied to prevent the dimension reduction of the convolutional feature maps. The weights and biases of the convolutional layers were initialized using the He normal initialization [22].

B. DISCRIMINATOR NETWORK ARCHITECTURE
D consists of five consecutive convolutional blocks. The convolutional block has two 2D convolutional layers with 3×3 kernels and a ReLU activation layer. The second 2D convolutional layer was operated with a stride of 2 to reduce the dimensions of the feature map. A batch normalization layer [23] was attached to the end of the convolutional block. The last convolutional block has no ReLU activation layer and batch normalization after the second 2D convolutional layer. The fully connected layer is attached to the end of the convolutional block and activated with a sigmoid function. All parameters of network D were initialized using He normal initialization.

C. LOSS FUNCTION

We employed three loss functions highly related to image quality: perceptual loss, joint sparsifying transformation (JST) loss, and adversarial loss.

1) PERCEPTUAL LOSS

Minimizing the pixel-wise mean squared error (MSE) cost function, widely used in image generation, is prone to an overly smooth and poor perceptual quality [20]. Recent studies have been based on the feature space [20], [24], [25], [26] for good perceptual similarity. We used a perceptually convincing loss function based on a pre-trained VGG19 network [27]. We defined ϕ as the feature map obtained by the ninth layer of VGG19. The perceptual loss L_{PL} is defined as follows:

$$L_{PL} = \frac{1}{WH} \| \phi(y) - \phi(G(x)) \|_F^2, \quad (5)$$

where $\| \cdot \|_F$ denotes the Frobenius norm, and W and H denote the width and height of the 2D output feature dimensions, respectively.

2) JST LOSS

You et al. [28] proposed a JST loss function, a nonlinear total variation (TV) loss function with joint constraints to express image sparsity. We employed the JST loss to mitigate conspicuous noise and artifacts and preserve the anatomical characteristics by minimizing the difference from the target image. The JST loss L_{JST} is defined as follows:

$$L_{JST} = \tau \| G(x) \|_{TV} + (1 - \tau) \| y - G(x) \|_{TV}, \quad (6)$$

where τ is a scaling factor set to 0.5.

3) ADVERSARIAL LOSS

The adversarial loss encourages the generator network G to learn the local patterns of the target images to convert images from various protocols into images with indistinguishable local patterns. The adversarial loss L_{adv} is defined as follows:

$$L_{adv} = E_{z \sim p_{data}(x)} \left[-\log D(G(x)) \right], \quad (7)$$

where $E(\cdot)$ is the expectation operator and $D(G(x))$ denotes the logistic probability of the generated image.

4) OBJECTIVE FUNCTIONS

The proposed network G was jointly optimized using three loss functions: perceptual loss, JST loss, and adversarial loss.

$$L_{total} = \lambda_1 \times L_{PL} + \lambda_2 \times L_{JST} + \lambda_3 \times L_{adv}, \quad (8)$$

where λ_1, λ_2, and λ_3 are the parameters used to weight the loss components of L_{total}. We experimentally found that $\lambda_1 = 10$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.1$ showed the best optimization performance.

III. EXPERIMENTAL SETUP

A. CT ACQUISITION

The CT images of the PH-5 CT Abdomen Phantom (Kyoto Kagaku, Japan) for training and internal test sets were obtained using the SOMATOM Definition Flash CT scanner (Siemens Medical Systems, Erlangen, Germany) at Seoul National University Hospital. Five CT scan protocols and two reconstruction kernels were used to create a variety of noise settings. We reconstructed raw data using a medium smooth body reconstruction filter (30f) via filtered back projection (FBP) and iterative reconstruction (IR) of strength 3 (SAFIRE). The target protocol was 120 kVp × 60 mAs with an IR of 30f, which is the routine scan protocol of Seoul National University Hospital. Each protocol was scanned five times to obtain the natural uncertainty between the same-protocol images. The number of slice images in the scan was 146. We divided the dataset into training and internal test datasets. Four of the same-protocol CT images were used for training, and the remaining CT images were used for internal testing. For the target protocol, an image averaged from five scans was used as the target image to minimize the influence of random noise.

For external testing, CT scans were conducted on the 5-year-old age-equivalent anthropomorphic pediatric phantom (ATOM model 705; CIRS, Norfolk, VA, USA) with another CT scanner, SOMATOM Force (Siemens Medical Systems, Erlangen, Germany). The scans were obtained under the three different dose levels with two reconstruction techniques (FBP and IR with ADMIRE strength 3) and a body regular smooth filter level of 40. Each scan file has 87 slice images. Additionally, target images were acquired in the same manner as in the internal testing to obtain the target protocol image of the external testing phantom.

The following helical scanning parameters were kept constant in all scans: detector collimation, 128 × 0.6 mm; pitch, 1.5; field of view, 315 × 315 mm; slice thickness, 3 mm; source-to-detector distance, 1085.6 cm with an image size of 512 × 512. The rotation time was 0.28 seconds for training and internal testing and 0.25 seconds for external testing. The CT scan protocols are listed in Table 1, and Table 2 summarizes the number of images in each dataset.

B. PREPROCESSING

For model training, we experimentally employed the adaptive image normalization method that utilized each image's maximum and minimum values instead of a fixed normalization approach with predetermined maximum and minimum values. The images were first normalized from the CT Hounsfield value (HU) to [−1, 1], and each 2D axial image's minimum and maximum values were stored for denormalization. Gaussian noise with a mean gray value of 0 and standard deviation of 0.05 was multiplied by 0.01 and randomly added to the source images to increase model generalization.

TABLE 1. CT scan protocols with dose parameters and reconstruction kernels.

	Protocol No.	Voltage (kVp)	Current (mAs)	Reconstruction Kernel
Internal	Int-P1	100	30	IR 30f
	Int-P2	100	30	FBP 30f
	Int-P3	100	60	IR 30f
	Int-P4	100	60	FBP 30f
	Int-P5	100	120	IR 30f
	Int-P6	100	120	FBP 30f
	Int-P7	120	30	IR 30f
	Int-P8	120	30	FBP 30f
	Int-P9	120	60	FBP 30f
External	Ext-P1	80	60	IR Br40
	Ext-P2	80	60	FBP Br40
	Ext-P3	100	40	IR Br40
	Ext-P4	100	40	FBP Br40
	Ext-P5	120	60	IR Br40
	Ext-P6	120	60	FBP Br40
Target	Target	120	60	IR 30f

C. NETWORK TRAINING

In prior works [10], [28], G was trained with randomly sampled local patches; however, we trained G with the original image size of 512×512, and the output image was the same size. This was done to measure perceptual and JST losses over the image rather than a local patch. Additionally, the discriminator network D was trained to discriminate between local textures. We sampled five patches from the exact location in the generated and target images as the discriminator input. The patch size was 64×64, and the patches were augmented with a random left-right flip and up–down flip to maximize the generalization capability. We used the adaptive moment estimation optimizer [29] with $\beta_1 = 0.9$ and $\beta_2 = 0.99$, and the learning rate was set to 10^{-5}. The mini-batch size was set to 1 for the accuracy of the gradient descent direction and was trained for 1500 epochs. All experiments were performed using TensorFlow 1.15 on an NVIDIA RTX 2080 Ti GPU with 11 GB of memory.

D. EVALUATION OF THE IMAGE QUALITY

Image quality was evaluated using two objective metrics: structural similarity index (SSIM) [30] and peak signal-to-noise ratio (PSNR) [31]. SSIM is a perception-based model that measures the structural similarity between two images

and is defined as follows:

$$SSIM\left(\hat{y}, y\right) = \frac{\left(2\mu_{\hat{y}}\mu_y + C_1\right)\left(2\sigma_{\hat{y}y} + C_2\right)}{\left(\mu_{\hat{y}}^2 + \mu_y^2 + C_1\right)\left(\sigma_{\hat{y}}^2 + \sigma_y^2 + C_2\right)},$$
$$C_1 = (K_1 \cdot max(y))^2,$$
$$C_2 = (K_2 \cdot max(y))^2, \quad (9)$$

where $\hat{y}$ is the source or generated image. $\mu_{\hat{y}}$ and μ_y are the means of images $\hat{y}$ and y, respectively, and $\sigma_{\hat{y}}$ and σ_y are the standard deviations of images $\hat{y}$ and y, respectively. $\sigma_{\hat{y}y}$ is the cross-covariance between images $\hat{y}$ and y. C_1 and C_2 are constants used to stabilize the equation and prevent division by zero, where $max(y)$ is the maximum pixel value of the target image, 255, and $K_1 = 0.01$ and $K_2 = 0.03$. PSNR measures the ratio of the maximum pixel value of the target image to the pixel-wise MSE between the target and generated images and is defined as follows:

$$PSNR = 20 \cdot log_{10}(max(y)) - 10 \cdot log_{10}(MSE),$$
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad (10)$$

where MSE is the average of the squared differences between the two images $\hat{y}$ and y. n is the total number of pixels, and y_i and $\hat{y}_i$ are the value of the ith pixel of image y and $\hat{y}$, respectively. The SSIM and PSNR values were calculated for all slices and reported as the mean and standard deviation.

E. EVALUATION OF RADIOMIC FEATURE REPRODUCIBILITY

To ensure the reproducibility of the radiomic workflow in our study, the CheckList for EvaluAtion of Radiomics research (CLEAR) [32] was utilized as a comprehensive, step-by-step reporting guideline, as detailed in Supplementary Materials.

We defined 10 anatomically verified locations, including tissue, vessel, and muscle, to extract radiomic features from the test image set. We sampled 32×32 local patches based on the defined locations, and identical patch samples were extracted from all protocols. A total of 819 radiomic features, including first-order (intensity), second-order (texture), and wavelet features, were extracted using Pyradiomics (Pyradiomics library version 3.0.1) [33]. Eighteen first-order statistics-based intensity features, 22 gray-level co-occurrence matrix (GLCM), 16 gray-level run length matrix (GLRLM), 16 gray-level size zone matrix (GLSZM), 14 gray-level dependence matrix (GLDM), and five neighboring gray-tone difference matrix (NGTDM) based texture features were extracted. In addition, 728 wavelet features extracted in the same manner as the image transformed using eight wavelet filters were used. The feature extraction settings mostly adhered to the default values of Pyradiomics. The bin width for discretizing image gray levels was set to 25, consistent with the Pyradiomics default commonly used in CT Radiomics studies [34], [35], [36], [37]. In contrast, for wavelet-transformed images, it was adjusted to 10 [33]. Coiflets 1 was selected for the wavelet analysis, which is

TABLE 2. Number of protocols, scans, and images per dataset.

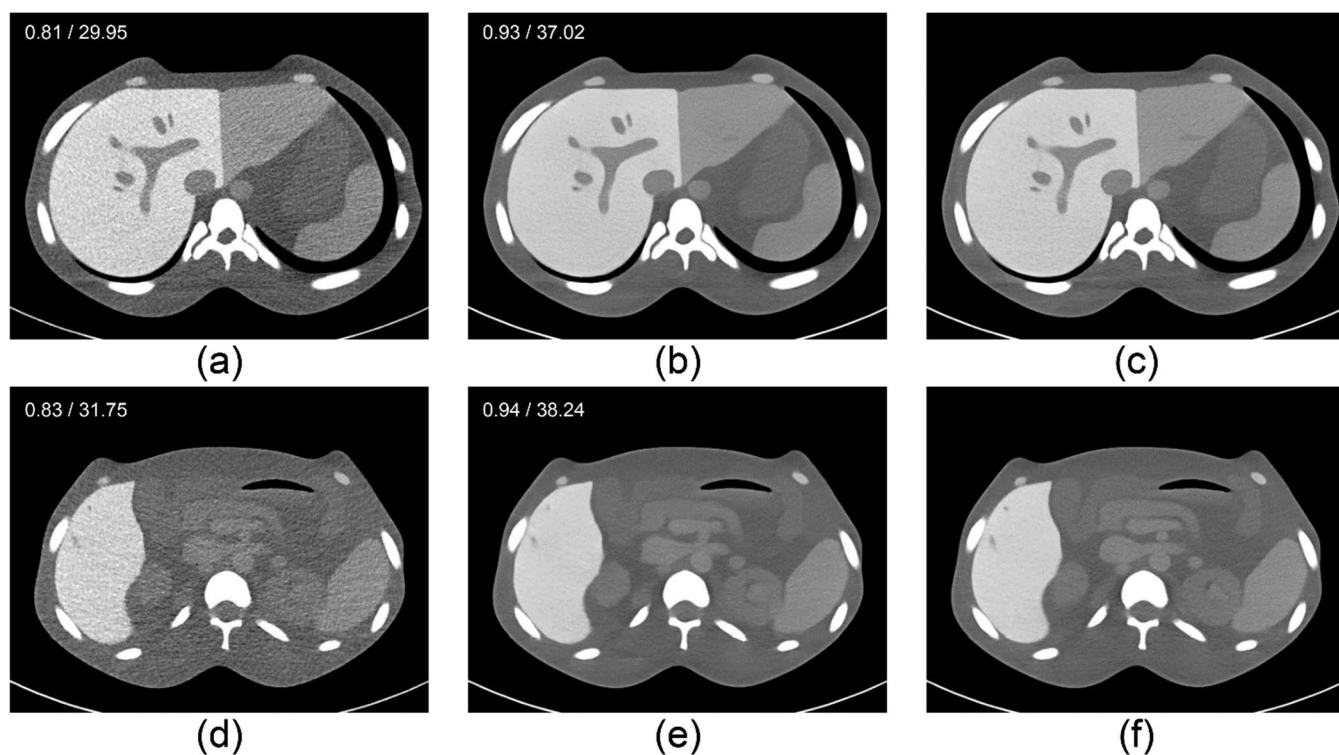| Datasets		Number of Protocols	Number of CT Scans per Protocol	Number of Slices per CT Scan	Total Images
Training	Source	9	4	146	5256
	Target	1	5 (averaged)	146	146
Internal Tests	Source	9	1	146	1314
	Target	1	5 (averaged)	146	146
External Tests	Source	6	1	87	522
	Target	1	5 (averaged)	87	87

FIGURE 3. Examples of original computed tomography (CT) images and generated CT images using the proposed model for the image. (a) and (b) are the source and generated CT images of protocol 100 kVp × 120 mAs with filtered back projection (FBP) 30f. (c) is the target protocol 120 kVp × 60 mAs with IR 30f. (d) and (e) are the source and generated CT images of protocol 120 kVp × 30 mAs with iterative reconstruction (IR) 30f. (f) is the target protocol 120 kVp × 60 mAs with IR 30f. The values in the corner are SSIM/PSNR values of each slice. The display window is [−160, 240] HU.

Pyradiomics default and has been utilized by several studies [38], [39], [40].

Lin's concordance correlation coefficient (CCC) [41], which was used to evaluate feature reproducibility in both internal and external tests, represents the agreement of radiomic feature values between the source and target patches and those between the generated and target patches across all protocols in the internal and external test sets. CCC ranges from −1 (entirely negative agreement) to 1 (entirely positive agreement). The analysis was performed using NumPy version 1.21.2 and Pingouin 0.5.3 [42]. We considered a radiomic feature with CCC ≥ 0.85 as a high-agreement feature [43], [44].

The proposed EHFS network was compared with four methods: 1) block-matching and 3D transform-domain collaborative filtering (BM3D) [45], 2) Choe et al.'s method [11], 3) RadiomicGAN [15], and 4) HFS network [16]. BM3D is the most popular filtering method and has demonstrated successful denoising performance in CT images [46], [47]. Choe et al.'s CNN-based method, GAN-based RadiomicGAN, and the HFS network showed improvements in CT radiomics reproducibility in previous studies.

TABLE 3. Image quality analysis with structural similarity index and peak signal-to-noise ratio.

Protocol	SSIM		PSNR	
	Source	Generated	Source	Generated
Int-P1	0.81±0.07	0.93±0.03	28.85±1.89	36.10±1.71
Int-P2	0.76±0.07	0.93±0.03	26.78±2.10	35.73±1.77
Int-P3	0.85±0.06	0.94±0.03	30.38±1.55	36.69±1.72
Int-P4	0.80±0.07	0.93±0.03	28.60±1.79	36.48±1.75
Int-P5	0.89±0.05	0.94±0.02	30.82±1.15	37.11±1.68
Int-P6	0.84±0.06	0.94±0.03	29.43±1.36	37.00±1.69
Int-P7	0.84±0.06	0.94±0.03	31.90±2.09	36.26±1.30
Int-P8	0.79±0.07	0.93±0.03	29.54±2.24	36.11±1.49
Int-P9	0.85±0.06	0.94±0.03	30.38±1.55	36.69±1.72
Average	0.83±0.07	0.94±0.03	29.90±2.51	36.46±1.66

F. COMPREHENSIVE ABLATION STUDY

We performed a comprehensive ablation study to demonstrate the effectiveness of sequential channel and spatial attention modules in the reproducibility of radiomic features. The HFS network [16] without the attention module is defined as a baseline model to demonstrate the contribution of the employed sequential attention module. We tested four models: baseline (HFS without attention), baseline + channel attention, baseline + spatial attention, and baseline + channel + spatial attention (proposed model). Because the sequential application of channel and spatial attention modules has better feature attention performance than the parallel application, the parallel application was not compared. The loss functions were not studied because this study is an extension of the prior HFS network, and the same loss functions of the HFS network were employed.

IV. RESULTS

A. IMAGE QUALITY EVALUATION RESULTS

The image processing results are presented in Fig. 3. The generated images exhibited a significantly higher SSIM than the source images (source vs. generated: 0.83 ± 0.07 vs. 0.94 ± 0.03). The SSIM of all protocols increased after image conversion. The protocols can be grouped into low-dose (Int-P1 to Int-P6) and high-dose (Int-P7 to Int-P9) groups. Before the image conversion, the SSIM of the low-dose group was 0.82 ± 0.07, and that of the high-dose group was 0.84 ± 0.07. After the image conversion, both groups showed similar SSIM improvements (0.93 ± 0.03 vs. 0.94 ± 0.03). The generated images showed a significantly high PSNR (source vs. generated: 29.90 ± 2.51 vs. 36.46 ± 1.66). The PSNRs of the low-dose and high-dose groups were 29.14 ± 2.12 vs.

36.52 ± 1.78 and 31.41 ± 2.54 vs. 36.34 ± 1.37, respectively. The PSNR increased in both groups, and the difference between the groups decreased from 2.27 to 0.18. The image quality was objectively improved through image conversion, and the difference between dose groups was reduced, making these dose groups more homogeneous. The evaluated SSIM and PSNR values are presented in Table 3.

B. IMPROVED FEATURE REPRODUCIBILITY IN THE GENERATED CT IMAGES

We analyzed the reproducibility of the radiomic features. The local sample patches used for radiomic feature extraction are shown in Fig. 4.

We counted the significantly reproduced features (CCC ≥ 0.85) to present the image conversion results effectively in the internal test set. The proposed method increased the number of significantly reproduced radiomic features in all the protocols. The averages of the number of significantly reproduced features in the source and generated images were 241 ± 38 and 565 ± 11, respectively. For first-order features, all features of the generated images were significantly reproduced in all protocols. On average, reproducible features increased from 36% to 84% for second-order features and from 27% to 67% for wavelet features. The number of significantly reproduced radiomic features is summarized in Table 4. Fig. 5 shows the protocols' CCC heat map for all radiomic features. We compared the proposed method with the other methods, as summarized in Table 5. The averages of the number of significantly reproduced features were as follows: BM3D: 305 ± 9; the method reported by Choe et al.: 381 ± 19; RadiomicGAN: 442 ± 32; HFS network: 518 ± 30; and EHFS network: 324 ± 16.

In the external test set, the number of significantly reproduced features increased from 239 ± 74 to 324 ± 16 with the EHFS network. As shown in Table 6, the results of external testing for all comparative methods were as follows: BM3D: 226 ± 8; the method reported by Choe et al.: 281 ± 20; RadiomicGAN: 296 ± 34; and HFS network: 306 ± 6. Through CCC analysis, we confirmed that the EHFS network converts images well, and the radiomic features were reproducible, similar to the target protocol images relative to other comparative models. This confirmation was true for internal and external testing, consisting of a new phantom and previously unseen protocols.

C. ABLATION STUDY RESULTS

The source images had an average of 241 ± 38 significantly reproduced features (CCC ≥ 0.85). The baseline model (without attention) had an average of 515 ± 28 significantly reproduced features. When the channel and spatial attention modules were independently applied to the baseline model, the average numbers of the significantly reproduced features were 549 ± 19 and 526 ± 39, respectively. This result indicates that channel attention contributes more than spatial attention to improving feature reproducibility. The EHFS network, the baseline model with a
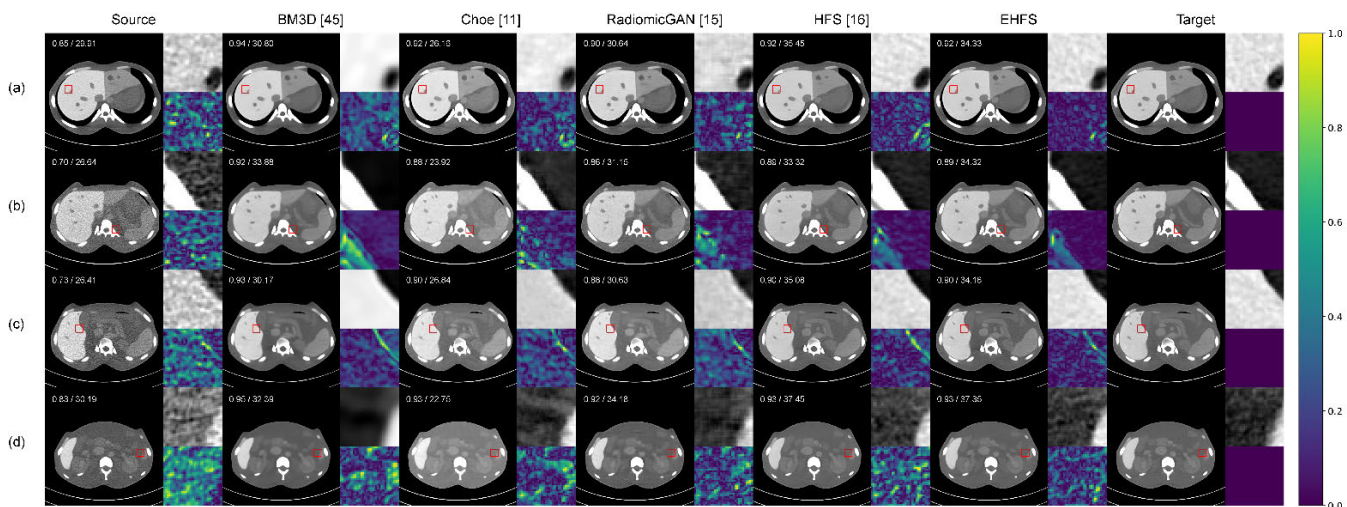
FIGURE 4. Examples of local patch images with extracted radiomic features. (a) Int-P5 protocol 100 kVp × 120 mAs with IR 30f. (b) Int-P8 protocol 120 kVp × 30 mAs with FBP 30f. (c) Int-P4 100 kVp × 60 mAs with FBP 30f. (d) Int-P3 protocol 100 kVp × 60 mAs with IR 30f. The values in the corner are the SSIM/PSNR values of each slice. The color bar for the images is located to the right of the figure. The display window is [−160, 240] HU.
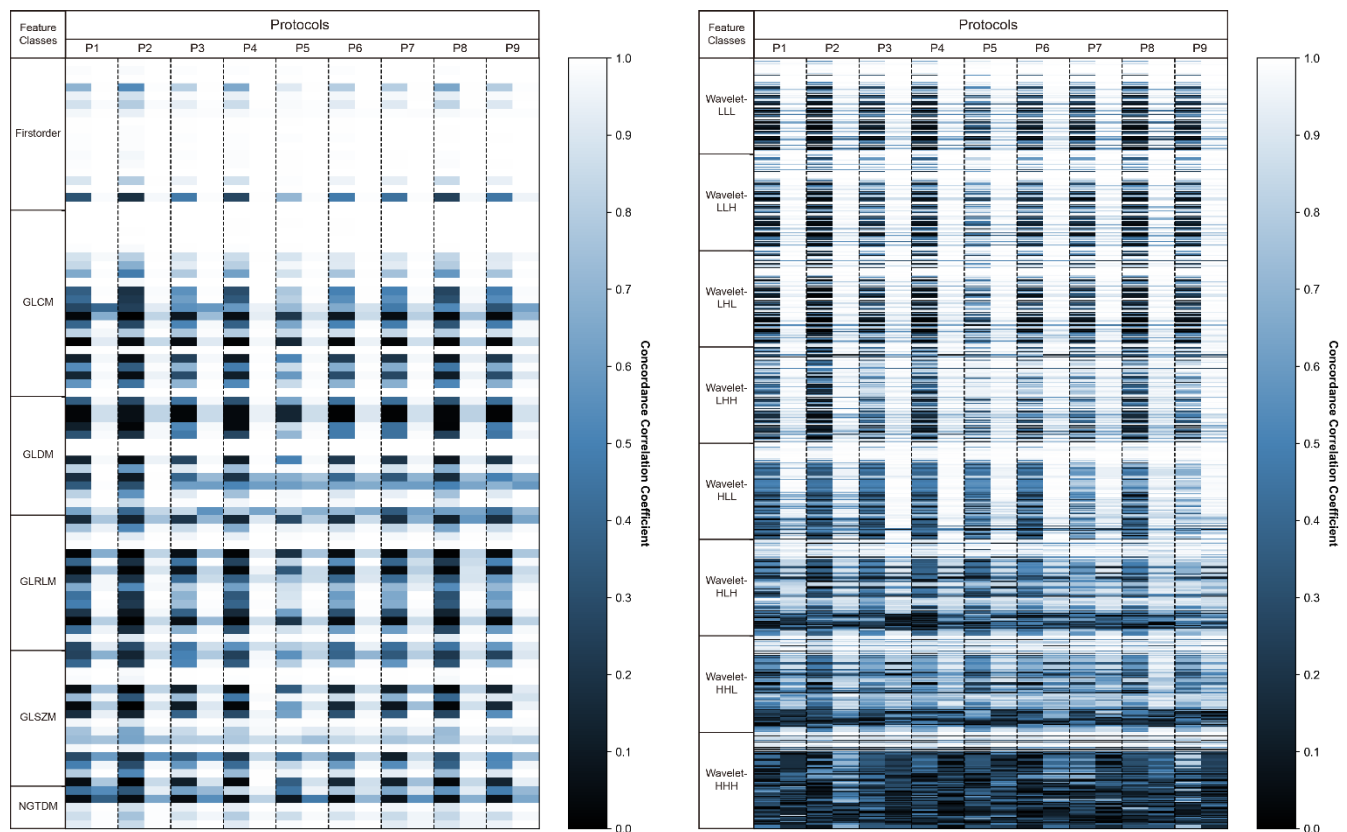

FIGURE 5. Concordance correlation coefficient (CCC) heat map of all radiomic features. The heat maps display each protocol's CCC of the 819 radiomic features.

sequential application of the channel and spatial attention modules, had an average of 565 ± 11 significantly reproduced features. The ablation study results are summarized in Table 7.

V. DISCUSSION

In this study, we developed an EHFS network that successfully improves the reproducibility of radiomic features across various CT protocols and reconstruction kernels. The network

TABLE 4. Number of significantly reproduced radiomic features by the protocol (CCC ≥ 0.85).

Protocol	First-order (Intensity)		Second-order (Texture)		Wavelet		Total	
	Source	Generated	Source	Generated	Source	Generated	Source	Generated
Int-P1	16 (89%)	18 (100%)	22 (30%)	59 (81%)	184 (25%)	466 (64%)	222 (27%)	**543 (66%)**
Int-P2	14 (78%)	18 (100%)	19 (26%)	61 (84%)	153 (21%)	481 (66%)	186 (23%)	**560 (68%)**
Int-P3	16 (89%)	18 (100%)	28 (38%)	57 (78%)	196 (27%)	488 (67%)	240 (29%)	**563 (69%)**
Int-P4	16 (89%)	18 (100%)	21 (29%)	68 (93%)	163 (22%)	500 (69%)	200 (24%)	**586 (72%)**
Int-P5	17 (94%)	18 (100%)	36 (49%)	61 (84%)	256 (35%)	490 (67%)	309 (38%)	**569 (69%)**
Int-P6	16 (89%)	18 (100%)	30 (41%)	66 (90%)	202 (28%)	474 (65%)	248 (30%)	**558 (68%)**
Int-P7	16 (89%)	18 (100%)	26 (36%)	60 (82%)	214 (29%)	483 (66%)	256 (31%)	**561 (68%)**
Int-P8	14 (78%)	18 (100%)	20 (27%)	60 (82%)	181 (25%)	491 (67%)	215 (26%)	**569 (69%)**
Int-P9	16 (89%)	18 (100%)	30 (41%)	60 (82%)	245 (34%)	498 (68%)	291 (36%)	**576 (70%)**
Average	16 ± 1	18 ± 0	26 ± 5	61 ± 3	199 ± 33	486 ± 10	241 ± 38	**565 ± 11**

TABLE 5. Algorithm comparison with number of significantly reproduced radiomic features (CCC ≥ 0.85).

Protocol	Source	BM3D [45]	Choe et al [11]	RadiomicGAN [15]	HFS network [16]	EHFS network
Int-P1	222 (27%)	318 (39%)	384 (47%)	427 (52%)	**550 (67%)**	543 (66%)
Int-P2	186 (23%)	315 (38%)	340 (42%)	390 (48%)	532 (65%)	**560 (68%)**
Int-P3	240 (29%)	300 (37%)	371 (45%)	501 (61%)	527 (64%)	**563 (69%)**
Int-P4	200 (24%)	290 (35%)	378 (46%)	428 (52%)	546 (67%)	**586 (72%)**
Int-P5	309 (38%)	299 (37%)	369 (45%)	485 (59%)	495 (60%)	**569 (69%)**
Int-P6	248 (30%)	293 (36%)	386 (47%)	423 (52%)	553 (68%)	**558 (68%)**
Int-P7	256 (31%)	308 (38%)	401 (49%)	446 (54%)	466 (57%)	**561 (68%)**
Int-P8	215 (26%)	310 (38%)	402 (49%)	444 (54%)	517 (63%)	**569 (69%)**
Int-P9	291 (36%)	313 (38%)	398 (49%)	430 (53%)	476 (58%)	**576 (70%)**
Average	241 ± 38	305 ± 9	381 ± 19	442 ± 32	518 ± 30	**565 ± 11**

outperforms existing methods in reproducibility and generalization, as demonstrated by comprehensive experiments on both internal and external datasets.

Radiomics has shown potential as an imaging biomarker for improving diagnostic and prognostic performance in many tumor diseases, such as lung cancer, head-and-neck cancer, breast cancer, and brain tumors [48], [49], [50]. However, the reproducibility of radiomic features in radiomics studies must be established to obtain generalizable results. We experimentally evaluated the improvement of radiomic features reproducibility in internal and external testing. Thus, the proposed EHFS network could be used as a preprocessing

step in CT radiomics studies to minimize the effects of protocol differences before quantitative analysis and could be integrated with other methods for improved reproducibility.

Traditional image quality assessment methods, such as PSNR and SSIM, are defined based on pixel-wise image differences; therefore, their ability to capture perceptually relevant differences, such as high texture detail, is limited. SSIM is relevant to human perception but does not explicitly incorporate information about image texture [20], [51]. Filter-based BM3D and CNN-based Choe's networks showed no significant difference in SSIM, and only PSNR decreased in Choe (SSIM/PSNR, BM3D: $0.95 \pm 0.02/32.80 \pm 2.26$,

TABLE 6. Algorithm comparison with number of significantly reproduced radiomic features in external testing (CCC ≥ 0.85).

Protocol	Source	BM3D [45]	Choe et al [11]	RadiomicGAN [15]	HFS network [16]	EHFS network
Ext-P1	170 (21%)	217 (26%)	242 (30%)	246 (30%)	307 (37%)	**310 (38%)**
Ext-P2	154 (19%)	227 (28%)	274 (33%)	254 (31%)	**305 (37%)**	301 (37%)
Ext-P3	262 (32%)	239 (29%)	288 (35%)	305 (37%)	311 (38%)	**319 (39%)**
Ext-P4	201 (25%)	220 (27%)	302 (37%)	312 (38%)	315 (38%)	**333 (41%)**
Ext-P5	**373 (46%)**	233 (28%)	284 (35%)	322 (39%)	297 (36%)	346 (42%)
Ext-P6	273 (33%)	217 (26%)	298 (36%)	335 (41%)	298 (36%)	**338 (41%)**
Average	239 ± 74	226 ± 8	281 ± 20	296 ± 34	306 ± 6	**324 ± 16**

TABLE 7. Number of significantly reproduced radiomic features for ablation study (CCC ≥ 0.85).

Protocol	Source	Baseline	Baseline + Channel Attention	Baseline + Spatial Attention	EHFS network (Baseline + Channel + Spatial)
Int-P1	222 (27%)	526 (64%)	551 (67%)	**573 (70%)**	543 (66%)
Int-P2	186 (23%)	509 (62%)	513 (63%)	553 (68%)	**560 (68%)**
Int-P3	240 (29%)	549 (67%)	544 (66%)	558 (68%)	**563 (69%)**
Int-P4	200 (24%)	534 (65%)	558 (68%)	539 (66%)	**586 (72%)**
Int-P5	309 (38%)	520 (63%)	524 (64%)	502 (61%)	**569 (69%)**
Int-P6	248 (30%)	550 (67%)	567 (69%)	**569 (69%)**	558 (68%)
Int-P7	256 (31%)	471 (58%)	550 (67%)	477 (58%)	**561 (68%)**
Int-P8	215 (26%)	469 (57%)	560 (68%)	499 (61%)	**569 (69%)**
Int-P9	291 (36%)	504 (62%)	575 (70%)	460 (56%)	**576 (70%)**
Average	241 ± 38	515 ± 28	549 ± 19	526 ± 39	**565 ± 11**

Choe: 0.93 ± 0.03/25.15 ± 2.72). However, radiomic reproducibility was significantly higher for Choe than for BM3D, as confirmed through internal and external testing (Internal/External, BM3D: 305 ± 9/226 ± 8, Choe: 381 ± 19/281 ± 20). Similarly, the SSIM and PSNR values in GAN-based RadiomicGAN, HFS, and EHFS were similar (SSIM/PSNR, RadiomicGAN: 0.92 ± 0.03/32.81 ± 2.00, HFS: 0.94 ± 0.03/35.87 ± 1.64, EHFS: 0.94 ± 0.03/36.46 ± 1.66), but a significant difference was obtained in the reproducibility of radiomic features (Internal/External, RadiomicGAN: 442 ± 32/296 ± 34, HFS: 518 ± 30/306 ± 6, EHFS: 565 ± 11/324 ± 16). This indicates that a high SSIM and PSNR do not guarantee improved reproducibility of radiomic features. The superior radiomics reproducibility observed in RadiomicGAN, HFS, and EHFS compared to BM3D and Choe suggests that more than simple filter techniques or shallow architectures are required to convert CT images of various protocols with different noise patterns. In addition, the proposed EHFS network significantly improved radiomic feature reproducibility compared to RadiomicGAN and the HFS network. The EHFS network performed consistently well in both internal and external testings, demonstrating its robustness and effectiveness on both seen and unseen data. It is especially noteworthy for unseen protocols, which are often challenging for other methods.

Our results confirm that both the newly designed generator and the adoption of adversarial training techniques significantly enhance radiomics reproducibility. As shown in Table 7, the baseline increased the number of reproducible features from 241 to 515 compared to the source image; and adversarial training improved radiomic feature reproducibility. Moreover, adding the channel and spatial attention modules to the generator network further enhanced the performance from 515 to 565 reproducible features. This finding underscores the positive contribution of generator

enhancements in improving the overall reproducibility of radiomic features.

Recently, many studies have attempted to overcome the lack of repeatability and reproducibility of radiomic features. Peng et al. [52] explored the primary source of variability in radiomic features by testing test-retest, inter-CT (manufacturers and models), and intra-CT protocols (acquisition and reconstruction parameters). They obtained stable, informative, and nonredundant radiomic features using hierarchical clustering. This approach can obtain substantial radiomic features after CT image conversion using the proposed EHFS network. Shafiq-ul-Hassan et al. [53] applied voxel size and gray-level normalization to CT images to increase the reproducibility of target radiomic features. This method may improve the robustness of the proposed EHFS network as a post-processing technique. However, integrating image conversion and radiomic feature reproducibility is difficult. Because the radiomic features are based on a non-differentiable texture matrix, gradient backpropagation is not possible, so they cannot be directly used for the training.

Our study has several limitations: First, it was performed using scanners from a single vendor. CT vendors affect the reproducibility of the radiomic features. Therefore, further testing using scanners from multiple vendors is required. Second, an experiment was performed using phantom images. However, we set as many scan protocols and reconstruction kernels as possible to assume various noise patterns to overcome this limitation. Finally, we assigned equal importance to all radiomic features. In the future, we will study the reproducibility of the essential features that can help diagnose diseases.

VI. CONCLUSION

The proposed EHFS network can improve the reproducibility of radiomic features, and it has the potential to act as a baseline for the harmonization of radiomic research. Furthermore, deep-learning-based CT image conversion could be a solution for standardization in ongoing radiomic research.

REFERENCES

[1] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, Feb. 2016, doi: 10.1148/radiol.2015151169.

[2] S. Rizzo, F. Botta, S. Raimondi, D. Origgi, C. Fanciullo, A. G. Morganti, and M. Bellomi, "Radiomics: The facts and the challenges of image analysis," *Eur. Radiol. Experim.*, vol. 2, no. 1, pp. 1–8, Dec. 2018, doi: 10.1186/s41747-018-0068-z.

[3] R. Berenguer, M. D. R. Pastor-Juan, J. Canales-Vázquez, M. Castro-García, M. V. Villas, F. M. Legorburo, and S. Sabater, "Radiomics of CT features may be nonreproducible and redundant: Influence of CT acquisition parameters," *Radiology*, vol. 288, no. 2, pp. 407–415, Aug. 2018, doi: 10.1148/radiol.2018172361.

[4] J.-P. Fortin, D. Parker, B. Tunc, T. Watanabe, M. A. Elliott, K. Ruparel, D. R. Roalf, T. D. Satterthwaite, R. C. Gur, R. E. Gur, and R. T. Schultz, "Harmonization of multi-site diffusion tensor imaging data," *NeuroImage*, vol. 161, pp. 149–170, Nov. 2017, doi: 10.1016/j.neuroimage.2017.08.047.

[5] J.-P. Fortin, N. Cullen, Y. I. Sheline, W. D. Taylor, I. Aselcioglu, P. A. Cook, P. Adams, C. Cooper, M. Fava, P. J. McGrath, and M. McInnis, "Harmonization of cortical thickness measurements across scanners and sites," *NeuroImage*, vol. 167, pp. 104–120, Feb. 2018, doi: 10.1016/j.neuroimage.2017.11.024.

[6] Y. Nan, J. Del Ser, S. Walsh, C. Schonlieb, M. Roberts, I. Selby, K. Howard, J. Owen, J. Neville, J. Guiot, and B. Ernst, "Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic review, meta-analysis and future research directions," *Inf. Fusion*, vol. 82, pp. 99–122, Jun. 2022, doi: 10.1016/j.inffus.2022.01.001.

[7] A. Midya, J. Chakraborty, M. Gönen, R. K. G. Do, and A. L. Simpson, "Influence of CT acquisition and reconstruction parameters on radiomic feature reproducibility," *J. Med. Imag.*, vol. 5, no. 1, Feb. 2018, Art. no. 011020, doi: 10.1117/1.jmi.5.1.011020.

[8] M. Meyer, J. Ronald, F. Vernuccio, R. C. Nelson, J. C. Ramirez-Giraldo, J. Solomon, B. N. Patel, E. Samei, and D. Marin, "Reproducibility of CT radiomic features within the same patient: Influence of radiation dose and CT reconstruction settings," *Radiology*, vol. 293, no. 3, pp. 583–591, Dec. 2019.

[9] S. Lefkimmiatis, "Non-local color image denoising with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5882–5891.

[10] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang, "Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1348–1357, Jun. 2018, doi: 10.1109/TMI.2018.2827462.

[11] J. Choe, S. M. Lee, K.-H. Do, G. Lee, J.-G. Lee, S. M. Lee, and J. B. Seo, "Deep Learning–based image conversion of CT reconstruction kernels improves radiomics reproducibility for pulmonary nodules or masses," *Radiology*, vol. 292, no. 2, pp. 365–373, Aug. 2019, doi: 10.1148/radiol.2019181960.

[12] E. C. de Farias, C. di Noia, C. Han, E. Sala, M. Castelli, and L. Rundo, "Impact of GAN-based lesion-focused medical image super-resolution on the robustness of radiomic features," *Sci. Rep.*, vol. 11, no. 1, p. 21361, Nov. 2021, doi: 10.1038/s41598-021-00898-z.

[13] S. Marcadent, J. Hofmeister, M. G. Preti, S. P. Martin, D. Van De Ville, and X. Montet, "Generative adversarial networks improve the reproducibility and discriminative power of radiomic features," *Radiol., Artif. Intell.*, vol. 2, no. 3, May 2020, Art. no. e190035, doi: 10.1148/ryai.2020190035.

[14] J. Chen, L. Wee, A. Dekker, and I. Bermejo, "Improving reproducibility and performance of radiomics in low-dose CT using cycle GANs," *J. Appl. Clin. Med. Phys.*, vol. 23, no. 10, Oct. 2022, Art. no. e13739, doi: 10.1002/acm2.13739.

[15] M. Selim, J. Zhang, B. Fei, G.-Q. Zhang, and J. Chen, "CT image harmonization for enhancing radiomics studies," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Houston, TX, USA, Dec. 2021, pp. 1057–1062, doi: 10.1109/BIBM52615.2021.9669448.

[16] S. B. Lee, Y. J. Cho, Y. Hong, D. Jeong, J. Lee, S.-H. Kim, S. Lee, and Y. H. Choi, "Deep learning-based image conversion improves the reproducibility of computed tomography radiomics features: A phantom study," *Investigative Radiol.*, vol. 57, no. 5, pp. 308–317, 2022, doi: 10.1097/rli.0000000000000839.

[17] J. Liu, W. Zhang, Y. Tang, J. Tang, and G. Wu, "Residual feature aggregation network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 2359–2368, doi: 10.1109/cvpr42600.2020.00243.

[18] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)* (Lecture Notes in Computer Science), 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[20] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690, doi: 10.1109/CVPR.2017.19.

[21] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, Jun. 2013, vol. 30, no. 1, p. 3.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034, doi: 10.1109/ICCV.2015.123.

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[24] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[25] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2016, pp. 694–711, doi: 10.1007/978-3-319-46475-6_43.

[26] J. Bruna, P. Sprechmann, and Y. LeCun, "Super-resolution with deep convolutional sufficient statistics," 2015, *arXiv:1511.05666*.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[28] C. You, G. Li, Y. Zhang, X. Zhang, H. Shan, S. Ju, Z. Zhao, Z. Zhang, W. Cong, M. W. Vannier, P. K. Saha, and G. Wang, "CT super-resolution GAN constrained by the identical, residual, and cycle learning ensemble (GAN-CIRCLE)," *IEEE Trans. Med. Imag.*, vol. 39, no. 1, pp. 188–203, Jan. 2020, doi: 10.1109/TMI.2019.2922960.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[30] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

[31] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2366–2369, doi: 10.1109/ICPR.2010.579.

[32] B. Kocak, B. Baessler, S. Bakas, R. Cuocolo, A. Fedorov, L. Maier-Hein, N. Mercaldo, H. Müller, F. Orlhac, D. Pinto dos Santos, A. Stanzione, L. Ugga, and A. Zwanenburg, "CheckList for EvaluAtion of radiomics research (CLEAR): A step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII," *Insights Imag.*, vol. 14, no. 1, pp. 1–13, May 2023, doi: 10.1186/s13244-023-01415-8.

[33] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer Res.*, vol. 77, no. 21, pp. 104–107, Nov. 2017, doi: 10.1158/0008-5472.can-17-0339.

[34] S. Feng, M. Gong, D. Zhou, R. Yuan, J. Kong, F. Jiang, L. Zhang, W. Chen, and Y. Li, "A CT-based radiomics nomogram for differentiation of benign and malignant small renal masses (≤4 cm)," *Transl. Oncol.*, vol. 29, Mar. 2023, Art. no. 101627, doi: 10.1016/j.tranon.2023.101627.

[35] V. Bevilacqua, N. Altini, B. Prencipe, A. Brunetti, L. Villani, A. Sacco, C. Morelli, M. Ciaccia, and A. Scardapane, "Lung segmentation and characterization in COVID-19 patients for assessing pulmonary thromboembolism: An approach based on deep learning and radiomics," *Electronics*, vol. 10, no. 20, p. 2475, Oct. 2021, doi: 10.3390/electronics10202475.

[36] K. Flouris, O. Jimenez-del-Toro, C. Aberle, M. Bach, R. Schaer, M. M. Obmann, B. Stieltjes, H. Müller, A. Depeursinge, and E. Konukoglu, "Assessing radiomics feature stability with simulated CT acquisitions," *Sci. Rep.*, vol. 12, no. 1, p. 4732, Mar. 2022, doi: 10.1038/s41598-022-08301-1.

[37] S. Mahmoudi, S. S. Martin, J. Ackermann, Y. Zhdanovich, I. Koch, T. J. Vogl, M. H. Albrecht, L. Lenga, and S. Bernatz, "Potential of high dimensional radiomic features to assess blood components in intraaortic vessels in non-contrast CT scans," *BMC Med. Imag.*, vol. 21, no. 1, pp. 1–10, Dec. 2021, doi: 10.1186/s12880-021-00654-9.

[38] Z. Shi, I. Zhovannik, A. Traverso, F. J. W. M. Dankers, T. M. Deist, P. Kalendralis, R. Monshouwer, J. Bussink, R. Fijten, H. J. W. L. Aerts, A. Dekker, and L. Wee, "Distributed radiomics as a signature validation study using the personal health train infrastructure," *Sci. Data*, vol. 6, no. 1, p. 218, Oct. 2019, doi: 10.1038/s41597-019-0241-0.

[39] M. Bologna, V. Corino, S. Cavalieri, G. Calareso, S. E. Gazzani, T. Poli, M. Ravanelli, D. Mattavelli, P. de Graaf, I. Nauta, K. Scheckenbach, L. Licitra, and L. Mainardi, "Prognostic radiomic signature for head and neck cancer: Development and validation on a multi-centric MRI dataset," *Radiotherapy Oncol.*, vol. 183, Jun. 2023, Art. no. 109638, doi: 10.1016/j.radonc.2023.109638.

[40] F. Prinzi, C. Militello, V. Conti, and S. Vitabile, "Impact of wavelet kernels on predictive capability of radiomic features: A case study on COVID-19 chest X-ray images," *J. Imag.*, vol. 9, no. 2, p. 32, Jan. 2023, doi: 10.3390/jimaging9020032.

[41] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 5, pp. 255–268, Mar. 1989, doi: 10.2307/2532051.

[42] R. Vallat, "Pingouin: Statistics in Python," *J. Open Source Softw.*, vol. 3, no. 31, p. 1026, 2018.

[43] J. Peerlings, H. C. Woodruff, J. M. Winfield, A. Ibrahim, B. E. Van Beers, A. Heerschap, A. Jackson, J. E. Wildberger, F. M. Mottaghy, N. M. DeSouza, and P. Lambin, "Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial," *Sci. Rep.*, vol. 9, no. 1, p. 4800, Mar. 2019, doi: 10.1038/s41598-019-41344-5.

[44] B. Zhao, Y. Tan, W.-Y. Tsai, J. Qi, C. Xie, L. Lu, and L. H. Schwartz, "Reproducibility of radiomics for deciphering tumor phenotype with imaging," *Sci. Rep.*, vol. 6, no. 1, p. 23428, Mar. 2016, doi: 10.1038/srep23428.

[45] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007, doi: 10.1109/TIP.2007.901238.

[46] P. F. Feruglio, C. Vinegoni, J. Gros, A. Sbarbati, and R. Weissleder, "Block matching 3D random noise filtering for absorption optical projection tomography," *Phys. Med. Biol.*, vol. 55, no. 18, pp. 5401–5415, Sep. 2010, doi: 10.1088/0031-9155/55/18/009.

[47] D. Kang, P. Slomka, R. Nakazato, J. Woo, D. S. Berman, C. C. J. Kuo, and D. Dey, "Image denoising of low-radiation dose coronary CT angiography by an adaptive block-matching 3D algorithm," in *Proc. SPIE*, vol. 8669, pp. 671–676, Mar. 2013, doi: 10.1117/12.2006907.

[48] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Commun.*, vol. 5, no. 1, p. 4006, Jun. 2014, doi: 10.1038/ncomms5006.

[49] A. S. Tagliafico, M. Piana, D. Schenone, R. Lai, A. M. Massone, and N. Houssami, "Overview of radiomics in breast cancer diagnosis and prognostication," *Breast*, vol. 49, pp. 74–80, Feb. 2020, doi: 10.1016/j.breast.2019.10.018.

[50] G. Singh, S. Manjila, N. Sakla, A. True, A. H. Wardeh, N. Beig, A. Vaysberg, J. Matthews, P. Prasanna, and V. Spektor, "Radiomics and radiogenomics in gliomas: A contemporary update," *Brit. J. Cancer*, vol. 125, no. 5, pp. 641–657, Aug. 2021, doi: 10.1038/s41416-021-01387-w.

[51] E. Wharton, K. Panetta, and S. Agaian, "Human visual system based similarity metrics," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, Oct. 2008, pp. 685–690, doi: 10.1109/ICSMC.2008.4811357.

[52] X. Peng, S. Yang, L. Zhou, Y. Mei, L. Shi, R. Zhang, F. Shan, and L. Liu, "Repeatability and reproducibility of computed tomography radiomics for pulmonary nodules: A multicenter phantom study," *Investigative Radiol.*, vol. 57, no. 4, pp. 242–253, 2022, doi: 10.1097/rli.0000000000000834.

[53] M. Shafiq-ul-Hassan, K. Latifi, G. Zhang, G. Ullah, R. Gillies, and E. Moros, "Voxel size and gray level normalization of CT radiomic features in lung cancer," *Sci. Rep.*, vol. 8, no. 1, p. 10545, Jul. 2018, doi: 10.1038/s41598-018-28895-9.

DAWUN JEONG received the B.S. degree in biomedical engineering from the Hankuk University of Foreign Study, Gyeonggi-do, Republic of Korea, in 2021. She is currently pursuing the Ph.D. degree in medical science with Yonsei University College of Medicine, Seoul, Republic of Korea. Her research interests include computer vision, medical image analysis, and radiomics.

YOUNGTAEK HONG (Member, IEEE) received the B.S. degree in digital information engineering from the Hankuk University of Foreign Study, Gyeonggi-do, Republic of Korea, in 2012, and the Ph.D. degree in medical science from Yonsei University College of Medicine, Seoul, Republic of Korea, in 2018. He was a Visiting Scholar with the Biomedical Imaging Research Institute, Cedars-Sinai Medical Center, CA, USA, from 2017 to 2018. He was a Postdoctoral Research Fellow with Yonsei University College of Medicine, in 2018, where he was a Research Assistant Professor, from 2020 to 2022. Since 2022, he has been the Research Director of Ontact Health Inc., Seoul. His research interests include computer vision, medical image analysis, and radiomics.

JINA LEE received the B.S. degree in digital media from Ajou University, Suwon, Republic of Korea, in 2019. She is currently pursuing the Ph.D. degree in medical science with the Yonsei University College of Medicine, Seoul, Republic of Korea. Her research interests include computer vision, medical image analysis, and radiomics.

SEUL BI LEE received the B.S. and M.S. degrees from the College of Medicine, Inje University, Busan, Republic of Korea, in 2014 and 2018, respectively. She is currently pursuing the Ph.D. degree with the College of Medicine, Seoul National University, Seoul, Republic of Korea. She received radiology residency training with Haeundae Paik Hospital and a Pediatric Radiology Fellowship with Seoul National University Hospital. Since 2021, she has been an Assistant Professor with the Department of Radiology, Seoul National University Hospital.
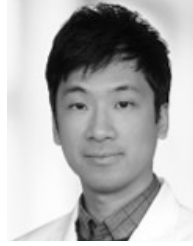
YEON JIN CHO received the B.S., M.S., and Ph.D. degrees from the College of Medicine, Seoul National University, Seoul, Republic of Korea, in 2012, 2017, and 2022, respectively, and the medical degree from Seoul National University Hospital. He received radiology residency training and a fellowship with Seoul National University Hospital. From 2019 to 2022, he was an Assistant Professor with the Department of Radiology, College of Medicine, Seoul National University.

HACKJOON SHIM received the B.S. and Ph.D. degrees from Seoul National University, Seoul, South Korea, in 1993 and 2007, respectively. From 2008 to 2010, he was a Research Professor with the Department of Electrical and Computer Engineering, Seoul National University College of Engineering, Seoul. Since 2011, he has been a Research Professor with the CONNECT-AI Research Center, Yonsei University College of Medicine, Seoul. His research interests include computer vision and medical image analysis.
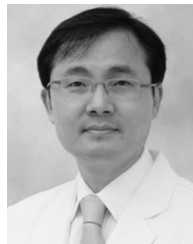
HYUK-JAE CHANG received the M.D. and M.M. degrees from the College of Medicine, Yonsei University, Seoul, South Korea, in 1994 and 1999, respectively, and the Ph.D. degree in medicine from the Graduate School of Medicine, Ajou University, Suwon, South Korea, in 2003. From 2002 to 2004, he was an Assistant Professor with the Department of Internal Medicine, College of Medicine, Ajou University. From 2004 to 2009, he was an Assistant Professor with the Department of Internal Medicine, Seoul National University College of Medicine, Seoul. From 2007 to 2008, he was a Postdoctoral Fellow with the Department of Internal Medicine, Johns Hopkins University, Baltimore, MD, USA. Since 2009, he has been a Professor with the Department of Internal Medicine, Yonsei University College of Medicine. Since 2016, he has been the CIO with the Division of Medical Information and Technology, Severance Hospital. From 2017 to 2020, he was the CEO of PHI Digital Healthcare Inc., Seoul. Currently, he is with Ontact Healthcare as the CEO and the Founder. His research interests include cardiovascular disease, pulmonary hypertension, and medical image processing.

● ● ●