A Physics-informed Deep Neural Network for Harmonization of CT Images

Mojtaba Zarei, Saman Sotoudeh Paima, Cindy McCabe, Ehsan Abadi,and Ehsan Samei,

***Abstract*— Objective: Computed Tomography (CT) quantification is affected by the variability in image acquisition and rendition. This paper aimed to reduce this variability by harmonizing the images utilizing physics-based deep neural networks (DNNs). Methods: An adversarial generative network was trained on virtual CT images acquired under various imaging conditions using a virtual imaging platform with 40 computational patient models. These models featured anthropomorphic lungs with different levels of pulmonary diseases, including nodules and emphysema. Imaging was conducted using a validated CT simulator at two dose levels and varying reconstruction kernels. The trained model was tested on an independent virtual test dataset and two clinical datasets. Results: On the virtual test set, the harmonizer improved the structural similarity index from 79.3±16.4% to 95.8±1.7%, normalized mean squared error from 16.7±9.7% to 9.2±1.7%, and peak signal-to-noise ratio from 27.7±3.7 dB to 32.2±1.6 dB. Moreover, the harmonized images yielded more precise quantification of emphysema-based imaging biomarkers for lung attenuation, LAA -950 from 5.6±8.7% to 0.23±0.16%, Perc 15 from 43.4±45.4 HU to 20.0±7.5 HU, and Lung Mass from 0.3±0.3 g to 0.1±0.2 g. In clinical data, the harmonizer reduced biomarker variability by an average of 70%. For lung nodules, harmonized images improved the detectability index by 6.5-fold and DNN-based precision by 6%. Conclusion: The proposed harmonizer significantly enhances image quality and quantification accuracy in CT imaging. Significance: The study demonstrated the potential utility of image harmonization for consistent CT image quality and reliable quantification, which is crucial for clinical applications and patient management.**

***Index Terms*— Harmonization, Computed Tomography, Physics-informed Deep Learning Model, Quantification.**

I. Introduction

QUANTITATIVE analysis of medical images offers valuable insights into assessing a patient's disease condition and progress, underscored by advancements in personalized treatment and novel imaging systems [1], [2]. This value is dependent on the reliability of quantitative results, a need that is influenced by the variability of imaging settings that cause inconsistencies in the image information and rendition. This is of particular relevance in computed tomography (CT) which has emerged as a major modality of interest in quantitative imaging. Inconsistencies in image formation can be mitigated by devising and utilizing standardized guidelines for protocol selections. It is challenging to identify what such standardized protocols should be, particularly considering the variabilities in clinical tasks [3], [4]. An additional or alternative approach is to devise algorithms that harmonize images post-acquisition. Notable examples of such algorithms include parametric empirical Bayes methods [5] and ComBat [6], [7]. While these approaches are promising, they are influenced by their sensitivity to assumptions about the data distribution and linearity, batch size, data scaling, parameter initialization, and potential removal of true biological variability [8]. Another approach is to mitigate spatial variations and biases by incorporating physics-based attributes of noise power spectra (NPSs) [9] and modulation transfer functions (MTFs) [10]. Despite their successful implementations for some clinical cases [11], their limitations for different kernels and systems have limited their applications over a diverse set of available systems and reconstruction kernels [12]. Deep neural networks (DNNs) offer a methodology to align distributions and learn invariant features, and thus an opportunity for harmonizing medical images [13]. Noteworthy potentials in this domain include maximum mean discrepancy [14], correlation distance [15], generative adversarial models [16], and disentangled representation [17]. A series of studies have demonstrated the potential of Generative Adversarial Networks (GANs) in improving the reproducibility and performance of radiomic features in CT scans. In [18], [19], GANs were employed to standardize and normalize CT images, effectively reducing variability in radiomic features. The study in [20] further demonstrated that GANs can enhance the reproducibility and discriminative power of radiomic features in radiography images, particularly addressing inter-manufacturer variability. The utilization of CycleGAN in [21] showcased the capability of GAN models to denoise low-dose CT scans, thereby

"This work was funded in part by the National Institutes of Health (P41-EB028744 and R01HL155293)."

Mojtaba Zarei is with the Center for Virtual Imaging Trials, Department of Electrical and Computer Engineering and Department of Radiology, Duke University, NC, USA (e-mail:mojtaba.zarei@duke.edu).

Saman Sotoudeh Paima is with the Center for Virtual Imaging Trials, Department of Electrical and Computer Engineering and Department of Radiology, Duke University, NC, USA (e-mail: saman.sotoudeh@duke.edu).

Cindy McCabe is with the Center for Virtual Imaging Trials, Department of Radiology, Duke University School of Medicine, Duke University, NC, USA (e-mail: cindy.mccabe@duke.edu).

Ehsan Abadi is with the Center for Virtual Imaging Trials, Department of Electrical and Computer Engineering and Department of Radiology, Duke University, NC, USA (e-mail: ehsan.abadi@duke.edu).

Ehsan Samei is with the Center for Virtual Imaging Trials, Department of Electrical and Computer Engineering and Department of Radiology, Duke University, NC, USA (e-mail: ehsan.samei@duke.edu).

improving the reproducibility and performance of radiomic features. The methodology proposed in [22] introduced a CT denoising method based on GANs, which was evaluated through radiomic feature reproducibility analysis and outperformed traditional methods. Additionally, incorporating frequency information content for training models has shown promising results. [23] introduced WaveGAN, a model for few-shot image generation that disentangles features into frequency components, thereby preserving structural information and enhancing fine detail synthesis. The approach in [24] presented a frequency-guided diffusion model for zero-shot medical image translation, utilizing frequency-domain filters to maintain structural information and outperforming existing methods across various metrics. These studies collectively highlight the significance of frequency-aware and structure-preserving techniques in image generation and medical imaging applications. They underscore the potential of GANs in addressing the challenges of variability and noise in radiomic features, thereby enhancing the reliability and accuracy of CT-based diagnostics.

These algorithms demand ample training data covering diverse imaging conditions and paired gold standards which are often not available, particularly at sufficiently-high image quality only possible high radiation dose acquisition – standard and low dose CT data suffers from scanner non-idealities, potentially introducing artificial and unrealistic textures in the harmonized images. This limitation can be mitigated by utilizing virtual, in *silico* image data which inherently provide known ground truth along with their images reflecting attributes of varying clinical imaging conditions [25], [26].

The objective of this study was to develop a harmonization algorithm employing a physics-informed generative adversarial neural network (GAN) model, trained by virtual imaging data. This harmonization model was crafted to enhance CT quantification of disease biomarkers and radiomics features, as well as CT image attributes such as noise magnitude and detectability. The paper reports the development, training, and validation of this methodology.

II. Methods

This study focused on chest CT scans. The following sections detail the methodology employed for creating training and reference images. Subsequently, the network structure is expounded upon, along with the loss function, augmentations, and regularization methods utilized. Lastly, the evaluation criteria are introduced. Throughout this paper, "H" and "NH" represent Harmonized and Non-harmonized images, respectively.

A. Training images and paired ground truth data

Forty distinct computational patient models, XCATs [27], were included. Among these, 10 XCAT phantoms had variable emphysema models, created based on real patient data [25], each with two different representations of emphysematous lung tissue, varying in severity, distribution, and size. The remaining 30 phantoms represented lung cancer nodules of varying size and degree of spiculation at different locations within the lungs. The patient models were imaged under varied imaging conditions emulating a commercial CT scanner (Force, Siemens) using a validated CT simulator (DukeSim [26]). The scans were done at ranges of clinically-used dose levels of 1.3 and 6.5 mGy CTDI_{vol}. The projection images were reconstructed using algorithms and kernels recommended by the American Association of Physicists in Medicine (AAPM) [28], [29]. These clinical kernels comprise different weighted-filter back-projections and iterative reconstructions including Br32f, Br32f(3), Br40f, Br40f(3), Br49f, Br49f(3), Br59f, Br59f(3), Br62f, Br64f, Br69f, Br69f(3), Qr32f, Qr49f, Qr59f, Qr61f. The number in the parenthesis indicates the strength of iterative kernels. All reconstructions were done at a field of view of 500 mm. The reference ground truth (GT), emulating CT images of the XCAT phantoms, is devoid of any scanner-related degradation such as noise, scatter, blur, or artifacts. Hence, the GT is a direct map from tissues in the associated XCAT phantom to the attenuation domain at 120 kVp in HU values without any noise or blurring effect. This means that all renditions of input images for one XCAT phantom would have the same GT reference. The methodology was tested using data from two phantoms (two COPD XCATs and one Nodule XCAT) that were not used in the training process. Therefore, the test data set included 32 renditions of one COPD and one Nodule XCAT phantom. The remaining 3 XCATs (32 renditions for each XCAT) were used for training and validation purposes. We utilized 75% of the data for training and the remaining 25% for validation.

B. Physics-informed Network Architecture

While our approach was primarily based on the image pixel value, following the strategy in [30], our network architecture employed the MTF of the images as a physics-informed input to further improve our harmonization algorithm. The MTF quantifies the spatial resolution characteristics of CT systems. MTF depends on both image acquisition and reconstruction kernels. Following steps described in [31], the patient's body is segmented utilizing a multithresholding technique, resulting in the creation of a binary volume. Subsequently, a tetrahedral mesh of the patient is generated using the iso2mesh toolbox within Matlab, with careful optimization of mesh size to ensure a balance between data representation and computational efficiency. The Edge Spread Function (ESF) is then measured along the air-skin interface, with contaminated measurements filtered out, and the right tail of the ESF is reconstructed to ensure precision. The ESF measurements are aligned and grouped based on radial distance, resulting in an oversampled ESF, which is subsequently binned. Following this, differentiation of the ESF yields the Line Spread Function (LSF), which undergoes a Fourier transform before normalization, ultimately yielding the desired

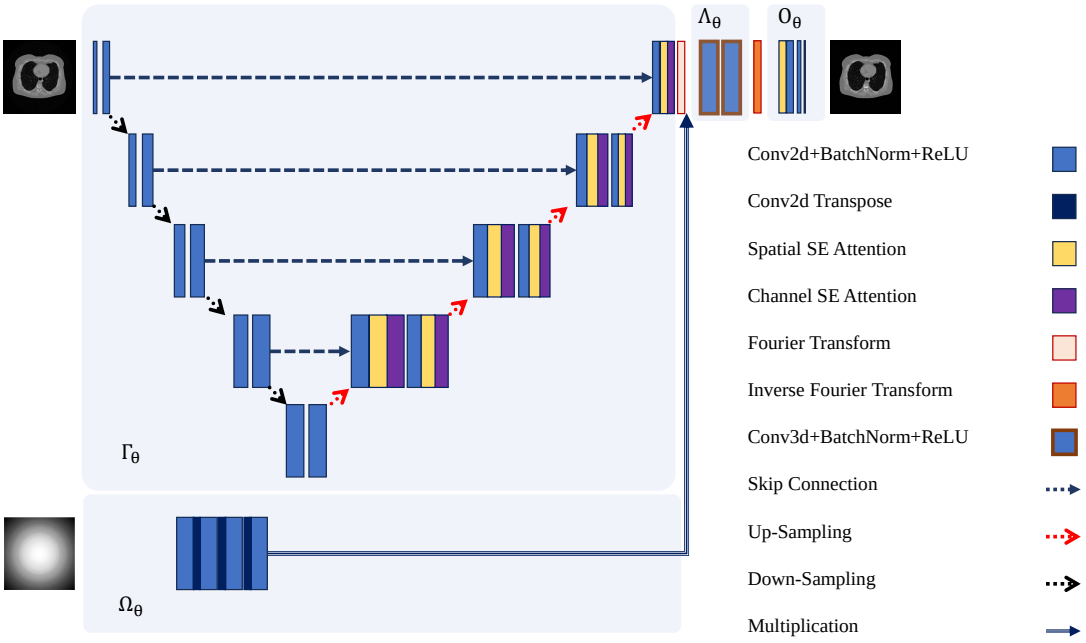


Fig. 1. The schematic of the developed DNN model. Two inputs are a single CT slice and the two-dimensional modulation transfer function (MTF) of the scanner. The output is the corresponding harmonized CT image to the given input CT slice.

MTF. In theory, under the assumption that the Fourier domain linearity holds for the projected 2D point spread function (PSF) in the XY-scan plane,

$$I_b(x,y) = \mathcal{F}^{-1}\Big[\mathcal{F}\big[I_a(x,y)\big] \times \frac{\mathcal{F}\big[PSF_b(x,y)\big]}{\mathcal{F}\big[PSF_a(x,y)\big]}\Big], \quad (1)$$

where I stands for a reconstructed image, and subscripts a and b stand for kernel a and b. $PSF(.)$, $\mathcal{F}(.)$, and $\mathcal{F}^{-1}(.)$ are point spread functions of the imaging systems, Fourier transfer function and inverse Fourier transfer function, respectively. Under isotropic system assumption, we have

$$I_b(x,y) = \mathcal{F}^{-1}\Big[\mathcal{F}\big[I_a(x,y)\big] \times \mathrm{MTF}_{ratio}(w)\Big]. \quad (2)$$

where MTF stands for the modulated transfer function. In practice, these assumptions may not be always valid, particularly in conditions where MTF_{ratio} cannot be accurately measured. As such we deployed parameterized deep neural networks (DNNs) to acquire nonlinear filters that match the input CT slice with its true reference counterpart. Labeling the image created using kernel a as the non-harmonized image (I_{nh}) and the harmonized images are indicated as (I_h),

$$\begin{aligned} I_h(x,y) =& O_\theta\Big(\mathcal{F}^{-1}\Big[\Lambda_\theta\Big(\mathcal{F}\big[\Gamma_\theta(I_{nh}(x,y))\big], \Omega_\theta(\mathrm{MTF}_{2D})\Big)\Big]\Big) \\ =& G_\theta(I_{nh}(x,y), \mathrm{MTF}_{2D}), \end{aligned} \quad (3)$$

where $O_\theta(.)$, $\Lambda_\theta(.)$, $\Gamma_\theta(.)$, Ω_θ and $G_\theta(.)$ are parameterized DNN with θ being their trainable parameter, and MTF_{2D} is the 2D modulated transfer function of the system measured at the skin-air boundary of input images at 64 frequency samples [31]. We utilized a U-Net architecture described in [32] for $\Gamma_\theta(.)$ network.$\Gamma_\theta(.)$ uses an attention mechanism for spatial encoding and network connectivity. To adapt a modified non-linear MTF_{2D}, $\Omega_\theta(.)$ was designed with two convolutional transpose layers to match the output size of $\mathcal{F}\big[\Gamma_\theta(.)\big]$. To incorporate the impact of the real and imaginary parts of the Fourier domain, we used a 3D convolution layer as the $\Lambda_\theta(.)$ network. $O_\theta(.)$ acts as a post-processing step over the attenuation domain which consists of two convolutional layers and one long attention mechanism from input data. We employed Pix2Pix adversarial networks training approach [33] to train the networks, where $G_\theta(.)$ served as the generator and a fully convolutional neural network with five layers acting as the discriminator. The generator network's schematic is illustrated Fig. 1.

C. Training

The training was conducted by sampling CT images from the training dataset. Each $G_\theta(.)$ input is the CT image ($I_{nh}(x,y)$) and the 2D MTF profile of the scanner, and its output is the corresponding harmonized image ($I_h(x,y)$). The discriminator's input is the CT image concatenated with corresponding harmonized images (or GT ($I_r(x,y)$))). The CT and GT images were normalized by subtracting 1024 Hounsfield Units (HU) and dividing them by 2048. In the training process, we used a combination of three different loss functions: Huber loss [34] for distance loss, Wasserstein loss with gradient penalty (WGANGP) [35] for adversarial loss, and perceptual style loss [36] for texture loss. The combination of these loss functions allows for the model to strike a balance between pixel-by-pixel distance loss, adversarial loss, and texture loss, allowing the model to preserve the pixel intensity, sharpness, and texture of the harmonized images. The

distance loss is defined as

$$\mathcal{L}_d(I_h, I_r) = \begin{cases} 0.5 \cdot (I_h - I_r)^2, & \text{if } |I_h - I_r| \leqslant \delta \\ \delta \cdot (|I_h - I_r| - 0.5 \cdot \delta), & \text{if } |I_h - I_r| > \delta \end{cases}, \quad (4)$$

with threshold δ set to 0.5. WGANGP is expressed as

$$\begin{aligned} \text{WGAN}_{\text{GP}} = D_\theta(I_{r,nh}) - \mathbb{E}[D_\theta(I_{h,nh})] \\ + \lambda_{GP}\mathbb{E}[(\|\nabla_{\tilde{I}} D_\theta(\tilde{I})\|_2 - 1)^2], \end{aligned} \quad (5)$$

where D stands for the discriminator network, $\hat{I} = \alpha I_{h,nh} + (1-\alpha) I_{r,nh}$, and α is a random number between 0 and 1. We set the gradient regularization coefficient, $\lambda_{\text{GP}} = 10$. Hence, the adversarial loss for training the generator is

$$\mathcal{L}_a(I_h) = \mathbb{E}[D_\theta(I_{h,nh})]. \quad (6)$$

For the texture loss, a pre-trained VGG-19 net [37] was used to extract features from harmonized and GT images. To focus on lung textures, each harmonized and GT image was multiplied by their lung mask, and then passed to the VGG net to extract perceptual features. Normalizing the Gram matrix of features, we calculated the texture loss by adding the mean squared difference of the errors. The texture loss was defined as

$$\mathcal{L}_t(I_h, I_r) = \sum_{i=1}^{N} \text{MSE}\left(\frac{Gm(\text{VGG}_i(I_h))}{\| Gm(\text{VGG}_i(I_h)) \|}, \frac{Gm(\text{VGG}_i(I_r))}{\| Gm(\text{VGG}_i(I_r)) \|} \right), \quad (7)$$

where $Gm(.)$ represents the Gram matrix operator, and i indicates the i^{th} feature vector extracted from the i^{th} VGG's down sampling layer. The generator loss function is defined as

$$\mathcal{L}_H(I_h, I_r) = \lambda_d \mathcal{L}_d(I_h, I_r) + \lambda_a \mathcal{L}_a(I_h) + \lambda_t \mathcal{L}_t(I_h, I_r), \quad (8)$$

where the loss weights were empirically determined ($\lambda_d = 1000, \lambda_t = 50$, and $\lambda_a = 1$). The Adam optimizer was used to train the networks. We used the cyclic learning scheduler with the base learning rate set to 10^{-4}. For a more stable training, gradient normalization was applied. The encoder backbone of the U-Net was initialized using weights trained on the ImageNet dataset [38], while the remaining network weights were initialized using the Xavier method. Throughout the training process, we delineated two sets of augmentations. In the initial set, we employed identical transformations on both inputs and their corresponding ground truth (GT) images. This encompassed random flips (horizontal or vertical) as well as stochastic affine transformations such as rotation, translation, scale, and shear. The second set pertained solely to the input images, which underwent transformations involving Gaussian noise and color jitter augmentations. Additionally, Gaussian noise was introduced to the MTF_{2D} to enhance the network's resilience against variations in MTF measurement errors.

D. Evaluation

The effectiveness of developed harmonizer was assessed in both virtual and actual clinical datasets.

1) Evaluation on Virtual Images

The initial segment involved assessing the original and harmonized images of the test set alongside the ground truth, using standard quality metrics such as the structural similarity index measure (SSIM), normalized root mean squared error (NRMSE), and peak signal-to-noise ratio (PSNR) for each slice. We also measured the harmonization's effectiveness for the accuracy of quantifying tasks. We assessed biomarker accuracies from harmonized and non-harmonized images. For cases with COPD, density-based biomarkers were LAA -950 (low attenuation area with HU < -950), Perc 15, (15th percentile of lung HU histogram), and Lung Mass ($\frac{HU+1024}{1024} \times$ voxel volume $\times N_{\text{lung voxel}}$). For cases with lung nodules, the nodule-based tasks were clinically-relevant morphological radiomics features of volume (V), surface-to-volume (A/V), and sphericity ($\frac{(36\pi V^2)^{\frac{1}{3}}}{A}$), representing size and irregularity of the lesion shape. We investigated the variability and bias across these biomarkers. To do so, we used K-means as a fixed segmentation algorithm to derive the segmentation mask for three nodules in the GT, non-harmonized, and harmonized images. We calculated the volume and surface of the binarized lesion's segmentation mask, fitted to a mesh surface as the surface (A) and volume (V) features.

Moreover, to demonstrate the positive impact of the physics-informed GAN model, we conducted an ablation study by training a UNet model (without the physics-informed component) and the proposed model (with the physics-informed component) using the same training strategy and dataset. We selected a UNet with a VGG encoder backbone, which contained 29 million trainable parameters, comparable to the 26.5 million parameters in our model. For evaluation, we utilized the same COPD test set and trained the models using only 18 XCAT phantoms with COPD diseases. This approach highlights the impact of the physics-informed component in scenarios with limited diversity in training data, a common situation in real clinical datasets. The remaining data was used for validation. Both models were trained with adversarial training, employing the described losses in Equation (8) with the same loss weights, learning rate scheduling, and the optimizer's parameters. Training continued until either signs of over-fitting were observed or the losses stabilized without further decrease.

2) Evaluation on Clinical Images

COPD quantification: We assessed the efficacy of the developed harmonizer using real clinical images obtained from the COPDGene dataset [39] . From the COPDGene dataset, we selected 40 patients with multiple renditions captured at two dose levels (regular and low doses) and different reconstruction kernels resulting in total of 148 cases. Subsequently, we harmonized these images using our algorithm. We examined the deviations in LAA -950 and

Perc 15. Drawing from previous research [40], we know that images acquired with smooth kernels and high dose provide the most accurate quantification of density-based COPD biomarkers. Therefore, we investigated whether the biomarkers derived from the harmonized images exhibited a quantitative bias toward those rendered with smooth kernels and regular dose. To this end, we reported Bland-Altman plot [41] biomarkers difference with respect to the non-harmonized images acquired with higher dose and smoother kernel among the candidates obtained from the harmonized feature with ComBat method [6], harmonized images with proposed algorithm, and non-harmonized images. All renditions from the combinations of the dose (high and low) and kernel (smooth, medium, and sharp) were grouped into separate batches resulting in a total of six different comparisons. The ComBat method was informed with the reference batch rendition. Besides, for demonstration purposes, the detailed variation analysis of the biomarkers for one of the cases was reported separately.

Nodule Detection evaluation: We further assessed the efficacy of the developed harmonizer using real clinical images obtained from the Luna datasets [42]. This dataset contains over 600 lung CT scans from four manufacturers, 17 scanner models and 18 kernels. For the nodule detection task, we undertook a twofold analysis of detection performance: 1) The conventional approach of detectability analysis, and 2) The utilization of a DNN-based algorithm for nodule detection.

In the first part, we utilized detectability index (d') using Fisher observer modeling [43]. We compared the performance of nodule detection in non-harmonized and harmonized images. The d' was computed as

$$d' = \int_{\Omega} \frac{W^2 MTF^2 E^2}{E^2 NPS + \alpha D^2} dudv, \tag{9}$$

where MTF, NPS, W, E, and D are modulated transfer function, noise power spectrum, task function, eye filter model, and reader distance to the monitor, respectively. α is a small positive coefficient ($\alpha = 10^{-8}$) determined through experimental analysis. The task function is dependent on the lesion size and field of view. For this study we assume a reference task of detecting the average nodule size reported in the dataset (4 mm) in a fixed contrast of 100 HU. Equation (9) encompass a task and observer model with NPS and MTF as the sole variable parameters. Therefore, we assessed the variability of frequency at 50% of MTF (MTFf_{50}) and the standard deviation of the noise as surrogates of the MTF and NPS, respectively. We measured the MTF in the air skin portion of each image and NPS over a uniform region inside the body.

The second part of detection performance analysis was centered around assessing the efficacy of a DNN-powered detection model with and without harmonized images. We employed a competition winner DNN model [44] for the detection task. The model's performance to detect nodules in harmonized images was measured and compared to that achieved with non-harmonized images. Additionally, we sought to further investigate harmonization advantages by training the model using harmonized images. During the training process, we persisted until the average of the last five training losses aligned with the average of the last five training losses obtained when training the model with non-harmonized images. Subsequently, we evaluated the Free Receiver Operating Curve (FROC), F_1 score, and precision of the predictions under three distinct scenarios. The first scenario involved training the model on non-harmonized images and testing it on non-harmonized images. In the second scenario, the model was trained on non-harmonized images and then tested on harmonized images. Finally, in the third scenario, the model was trained on harmonized images and tested on harmonized images. To carry out the training, we adhered to the same preprocessing and steps that were employed in training the original model on the non-harmonized images. To calculate the FROC curve, following the evaluation approach in [44], we conducted 1000 bootstrapping iterations to calculate the confidence interval of the FROC curves. Additionally, we applied the same exclusion criteria for lesions as was imposed on the original dataset.

TABLE I
QUANTITATIVE RESULTS ON THE COPD TEST SET. ALL METRICS WERE MEASURED ON A SINGLE 2D SLICE.

	NRMSE (%)	SSIM (%)	PSNR (dB)
H	**9.2±1.7**	**95.8±1.7**	**32.2±1.6**
NH	16.7±9.7	79.3±16.4	27.7±3.7
	|Δ|LAA -950 (%)	|Δ|Perc 15 (HU)	|Δ|Lung mass (g)
H	**0.23±0.16**	**20.0±7.5**	**0.1±0.2**
NH	5.6±8.7	43.4±45.4	0.3±0.3

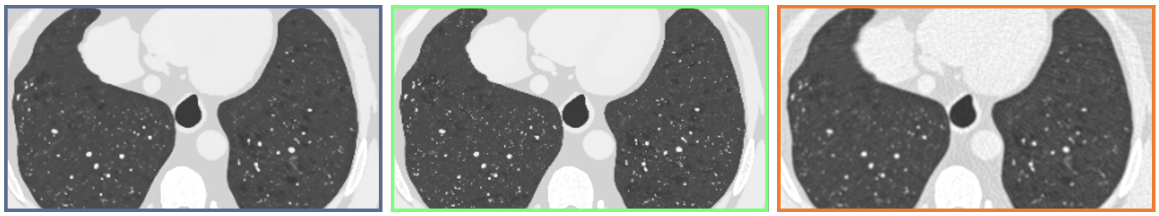
III. RESULTS

A. Evaluation on Virtual Image Domain

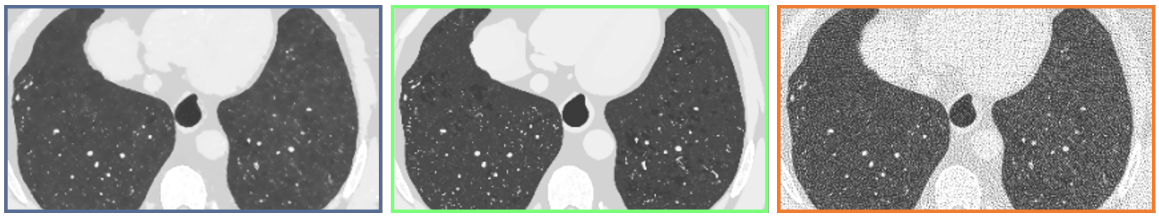
Figs. 2(a) and 2(b) depict ground truth and example images at the two ends of the image acquisition and reconstruction spectra including non-harmonized and harmonized, representing images at high dose and smooth kernel and images at low dose and a sharp kernel. The harmonized images offer higher sharpness and reduced noise compared to the non-harmonized images. These

TABLE II
QUANTITATIVE RESULTS ON THE TEST SET WITH LUNG NODULE INSERTED. THE FIRST THREE METRICS WERE MEASURED ON A SINGLE 2D SLICE, AND RADIOMICS FEATURES WERE MEASURED ON THE 3D SLICES.

	NRMSE (%)	SSIM (%)	PSNR (dB)
H	**11.5±1.9**	**95.1±1.3**	**31.3±1.1**
NH	14.2±4.0	87.4±9.1	29.6±2.3
	$V_{\text{NRMSE}(\%)}$	$A/V_{\text{NRMSE}(\%)}$	$S_{\text{NRMSE}(\%)}$
H	**2.1±1.3**	**14.2±3.0**	16.5±**3.8**
NH	3.4±2.5	14.8±6.9	**16.3**±8.45



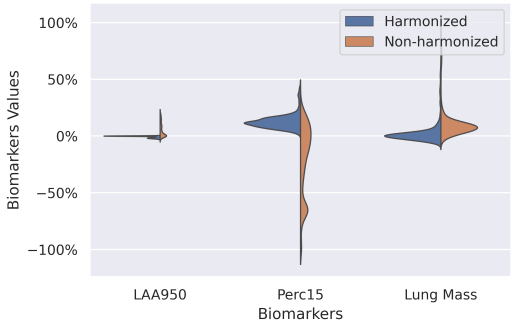
(a) High dose and smooth kernel rendition.



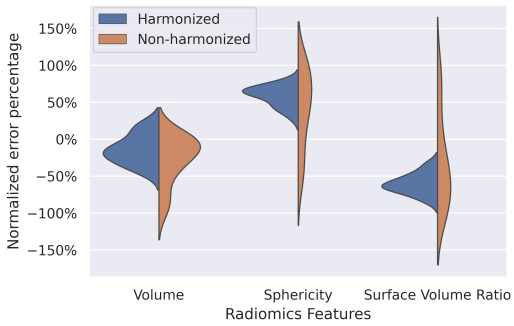
(b) Low dose and sharp kernel rendition.

Fig. 2. Harmonized images (in the blue square) versus non-harmonized images (in red square) visual differences. Non-harmonized image in the top row was acquired with high dose (100 mAs) and smooth kernel (Br32) and in the bottom row was acquired with a low dose (20 mAs) and sharp kernel (Br64). The corresponding GTs are in green squares. The GT is a direct conversion from XCAT to the HU domain (no reconstruction involved for the GTs). In other words, the image in green box represents an ideal CT image, aka ground truth, if there was no degradation caused by the imaging system. The images in the red squares are reconstructed from the same XCAT phantom, and it highlights how low-contrast intra-organ heterogeneity (texture) and image sharpness will be impacted due to the non-idealities of the acquisition and reconstruction.

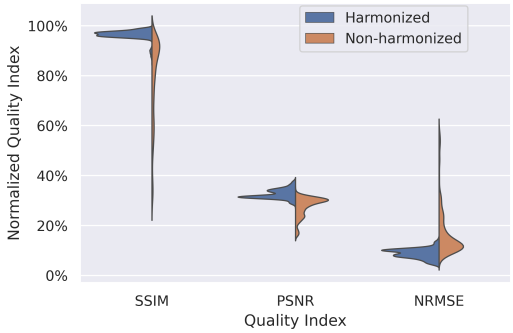
visual attributes are quantitatively affirmed in results shown in Fig. 3 in terms of the variability and bias of the quantification for the COPD and lung nodule cases. Fig. 4 represents the quantification variability and bias for each kernel separately. Utilizing a smooth kernel (Br32f) in non-harmonized images yields favorable outcomes, as evidenced by lower bias and variability in generic quality metrics and intensity-based biomarkers. Conversely, to improve bias in morphological features, sharper kernels (Br64f) exhibit higher accuracy. In the context of harmo-



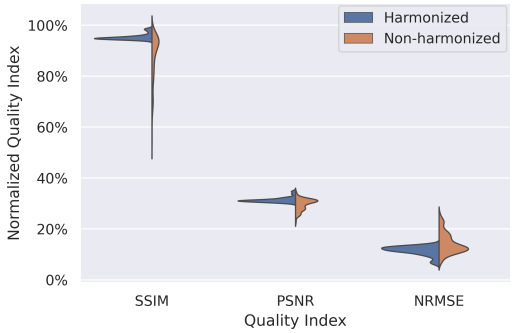
(a) Biomarkers difference for the COPD cases.



(b) Morphological radiomics feature difference for the nodule cases.

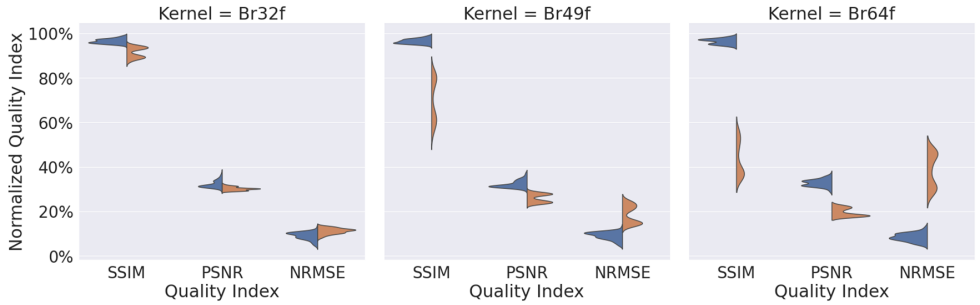


(c) Quality index error for the COPD cases.

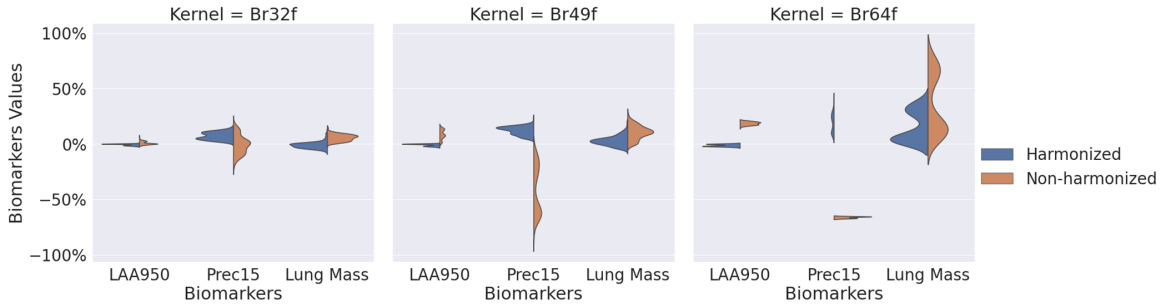


(d) Quality index error for the nodule cases.

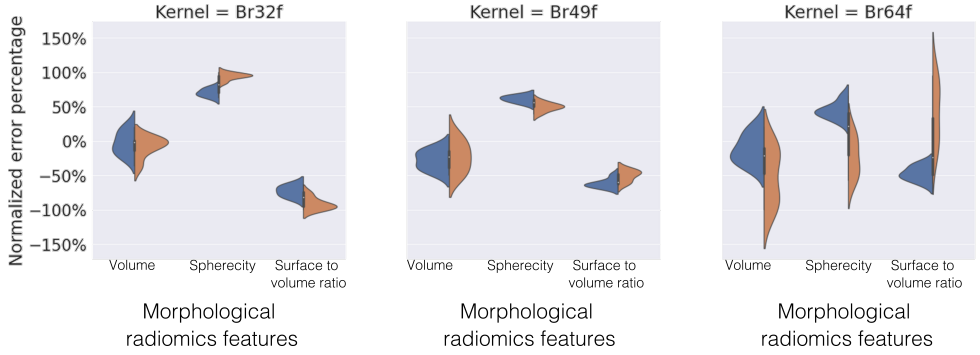
Fig. 3. The quantitative outcomes for the nodule cases (right column) and COPD cases (left column). To compare the PSNR values with other measurements, their values in dB were divided by 100.



(a) Quality index variability across kernels.

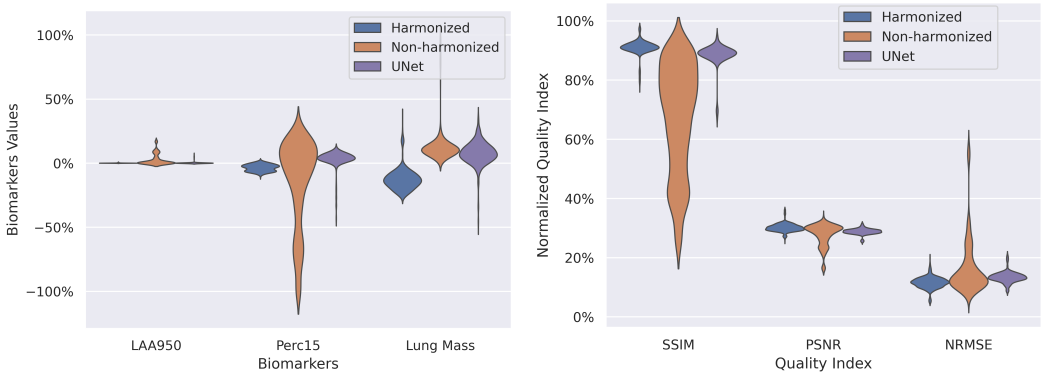


(b) COPD biomarkers variability across kernels.



(c) Morphological radiomics feature variability across kernels.

Fig. 4. Analyzing the differences in bias and variability between harmonized and non-harmonized images according to the deployed smooth, average, and sharp kernels.



(a) Biomarkers difference for the COPD cases. (b) Quality index error for the COPD cases.

Fig. 5. The quantitative outcomes for COPD cases in the ablation study. To compare the PSNR values with other measurements, their values in dB were divided by 100. The blue, red, and purple data are measured metrics and biomarkers from the harmonized, non-harmonized, and Unet output images, respectively.

TABLE III
QUANTITATIVE RESULTS ON THE COPD TEST SET FOR THE ABLATION STUDY. ALL METRICS WERE MEASURED ON A SINGLE 2D SLICE.

	NRMSE (%)	SSIM (%)	PSNR (dB)
H	**11.5±1.9**	**91.0±2.5**	**30.2**±1.4
UNet	13.4±1.9	88.4±4.2	28.8±**1.0**
	|Δ|LAA -950 (%)	|Δ|Perc 15 (HU)	|Δ|Lung mass (g)
H	**0.17±0.24**	**6.3±4.1**	0.3±0.12
UNet	0.54±1.93	10.7±10.9	**0.1±0.07**

TABLE IV
VARIABILITY ANALYSIS IN QUANTIFICATION OF THE COPD BIOMARKERS FOR HARMONIZED AND NON-HARMONIZED IMAGES.

	Kernel	Dose	LAA -950	Perc 15
H	-	-	**1.11**	**-915**
NH	Smooth	Regular	**1.91**	**-922**
H	-	-	**5.17**	**-932**
NH	Medium	Regular	11.81	-946
H	-	-	**3.87**	**-919**
NH	Smooth	Low	7.86	-931
H	-	-	**1.83**	**-920**
NH	Smooth	Low	6.19	-932
H	-	-	**1.86**	**-915**
NH	Smooth	Low	2.54	**-923**

nized images, the variation and bias remain unaffected by the specific kernel used, enabling a quantification that is largely independent of the kernel employed. Tables I and II present the errors associated with both harmonized and non-harmonized images for COPD cases and nodule cases, respectively. In the tables, the bold numbers indicate the superior scores. Table I focuses on the intensity-based biomarkers for COPD cases, while Table II examines morphological biomarkers. The tables provide generic quality metrics separately for each case.

The improvements associated with harmonization are evident in both generic image quality metrics and imaging biomarkers, enabling more robust and accurate quantification of CT images across varied targeted tasks. Across all cases, on average, the generic image quality of NRMSE, SSIM, PSNR improved by 24%, 17%, and 48%, respectively. The mean absolute error in intensity-based biomarkers of LAA -950, Perc 15 and Lung mass were reduced by 95%, 54%, and 66%, respectively.

Figures 4(a) and 4(b) depict the outcomes of the ablation study concerning both generic image quality metrics and pertinent clinical metrics. Table III presents the quantitative findings. In most of the evaluated metrics and biomarkers, the proposed model demonstrated statistically significant higher accuracy and lower variability compared to the UNet, with the sole exception being lung mass for both accuracy and variability and PSNR for variability.

B. Evaluation in Clinical Images

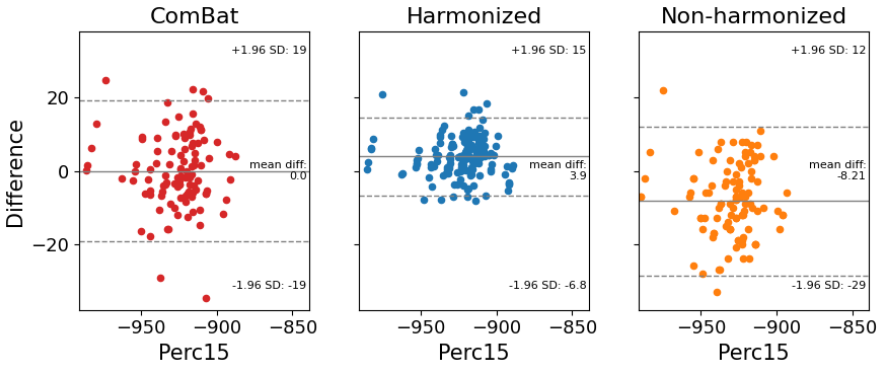
1) COPD quantification

Fig. 6 depicts the Bland-Altman plot showcasing the COPD biomarkers within the COPDGene dataset before and after applying harmonization through both ComBat and our developed harmonizer. Our proposed harmonizer decreased the variability of the Perc 15 and LAA -950 measurements compared to the non-harmonized measurements, reducing their variability from 20.5 and 8.65 to 10.9 and 4.4, respectively. Additionally, our developed harmonization method exhibited superior performance compared to the ComBat method [6], achieving a substantial reduction of 42.6% and 47.6% in the variability of Perc 15 and LAA -950, respectively. For a detailed comparison, the evaluation results for one individual case are reported in Table IV. The standard deviations of LAA -950 and Perc 15 from the non-harmonized images were 3.9 and 6.3. For the harmonized images, these dropped to 0.9 and 2.3, demonstrating the harmonizer's ability to reduce variability. Smooth kernels at high doses are expected to provide more precise measurement of biomarkers [40], as shown in the first row in Table IV. Even so, harmonization offers an added advantage of further improvement.

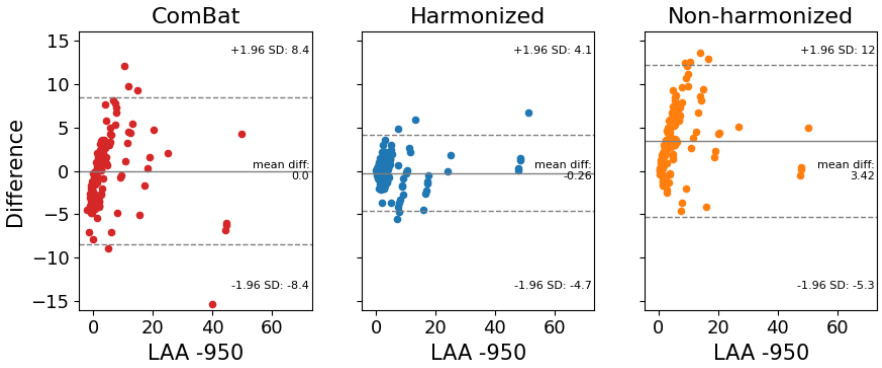
2) Nodule Detection evaluation

Figure 7 illustrates the variability in measurement of the standard deviation of the noise (σ), the MTFf_{50}, and resultant d'. As expected, due to the lower noise and higher sharpness of the harmonized images, the detectability index of the harmonized images is higher than the non-harmonized images. The d' values in the harmonized images are statistically significant when compared to the non-harmonized images. As noted earlier, the analysis of the effect of harmonized images on lung detection included how the detection model's performance would change when tested with non-harmonized and harmonized images after it had been initially trained with as well as harmonized images. The FROC results for three plausible scenarios, depicted in Fig. 8, are similar but show slightly improved performance with the training on harmonized images and testing on the harmonized scenario. The number of false positives in harmonized images at the decision threshold (0.816 for trained on harmonized and 0.823 for trained on non-harmonized images) are improved by 18.7% and 0.3% depending on training with harmonized or non-harmonized images, respectively. Table V summarized the precision, F_1 score, and Area Under the ROC Curve (AUC) of the three cases. Fig. 9 illustrates four distinct CT images from the Luna dataset, captured using scanners manufactured by GE, Siemens, Philips, and Toshiba. Comparing the harmonized versions of these CT scans to their non-harmonized counterparts, the harmonized images exhibit reduced noise and enhanced sharpness. These results demonstrate that the harmonizer, trained solely on virtual images, is capable of effectively adapting to clinical data from various vendors.

The analysis of the effect of harmonized images on the performance of the in-house detection model within the Luna dataset focused on two primary aspects. Initially,

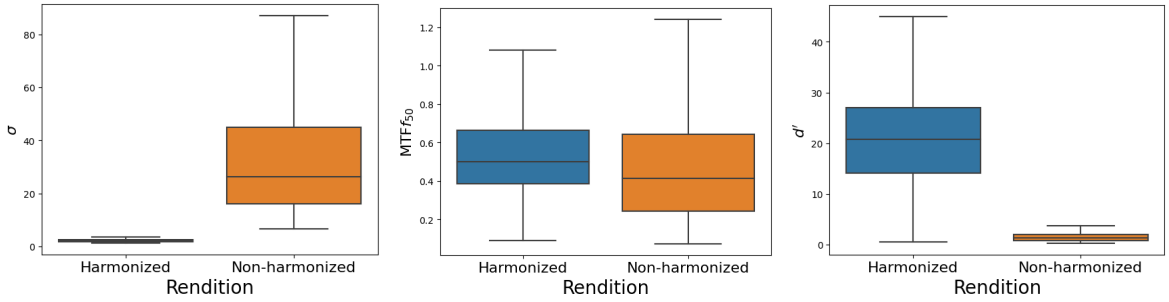


(a) Perc 15 variability.



(b) LAA -950 variability.

Fig. 6. COPD biomarkers variability and bias with respect to the smooth kernel and high dose renditions in COPDGene dataset. The results, from left to right, were obtained from the ComBat feature harmonization algorithm, proposed harmonizer, and non-harmonized images, respectively.



(a) Variability of the standard deviation of the noise (σ).

(b) Variability of the MTFf_{50}.

(c) Variability of the d'.

Fig. 7. Detectability analysis of the harmonized image versus non-harmonized image over Luna dataset [42]. The higher variability observed in the d' value is attributed to the smaller standard deviation of noise present in the harmonized images.

we examined how the model's performance would change when tested with harmonized images after it had been initially trained with non-harmonized images. Secondly, we examined the effects of training the model with harmonized images until it achieved the same loss level as the original model. The FROC curves for three plausible scenarios are depicted in Figure 8. The obtained results show that the model performance on the harmonized images has the same performance, and only a subtle degradation will happen when the trained model with non-harmonized images is used to detect the lesions in harmonized images. However, when the model is trained with the harmonized images, the sensitivity of the model slightly increases. Especially for lower false positives per scan within 0.5 to 1. If we look at the number of false positives in three cases, we realized that by training and testing on the harmonized images, the false positives at the cut-off operation point are 13% less than the false positives when we test the original model with non-harmonized images.

IV. Discussion

Our study presented a comprehensive development of CT harmonization and an analysis of its utility across

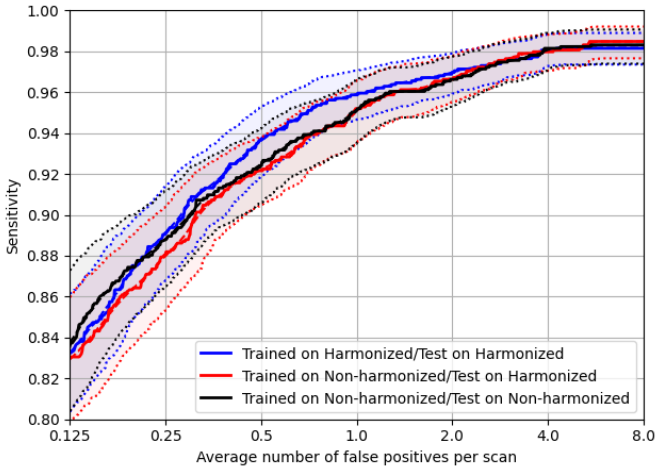


Fig. 8. FROC curves for three scenarios. The black solid curve pertaining to the test performed with the original model over the non-harmonized images. The solid red curve is the test result when we used the original model with harmonized images. The solid blue curve represents the result when we trained the model with harmonized images and tested the harmonized images.

TABLE V
DETECTION SCORES FOR ADOPTED THREE CASE STUDIES.

Trained on/Tested on	Precision	F_1	AUC
NH / NH	0.76	0.83	**0.97**
NH / H	0.75	0.83	**0.97**
H / H	**0.80**	**0.86**	**0.97**

various quantitative tasks. Through comparisons of non-harmonized and harmonized images, both virtual and clinical, the results illustrate the benefits of harmonization. We showed that by using harmonized images, conventional mathematical and data-driven methods could be improved in terms of quantification outcomes. Harmonized images consistently exhibit improved sharpness and reduced noise.

We validated the developed harmonizer in both the virtual and clinical domains. In the virtual domain, using the ground truth, we verified the bias correctness of the harmonized images in both conventional and photon counting CT.We further explored the influence of the physics-informed components on the proposed model's efficacy through an ablation study. The results indicated that, even with limited data, our proposed model outperforms in general image quality metrics. Additionally, the proposed model significantly better preserves clinical information in most relevant biomarkers compared to the UNet model. However, the UNet model shows better performance in lung mass variability reduction. In the clinical domain, we validated harmonization performance with clinical images acquired at different times and from different sites and vendors. Our findings were aligned with previous research in highlighting the importance of kernel choice and dose level in quantification bias and variability. The harmonized images remained resilient to the specific acquisition setting used, ensuring consistent quantification across diverse acquisition settings and manufacturers. By reducing variability and enhancing alignment with true values, harmonized images provided more accurate and robust measurements for COPD and lung nodule cases. In comparison to conventional harmonization techniques such as ComBat, the proposed harmonizer demonstrated superior performance.

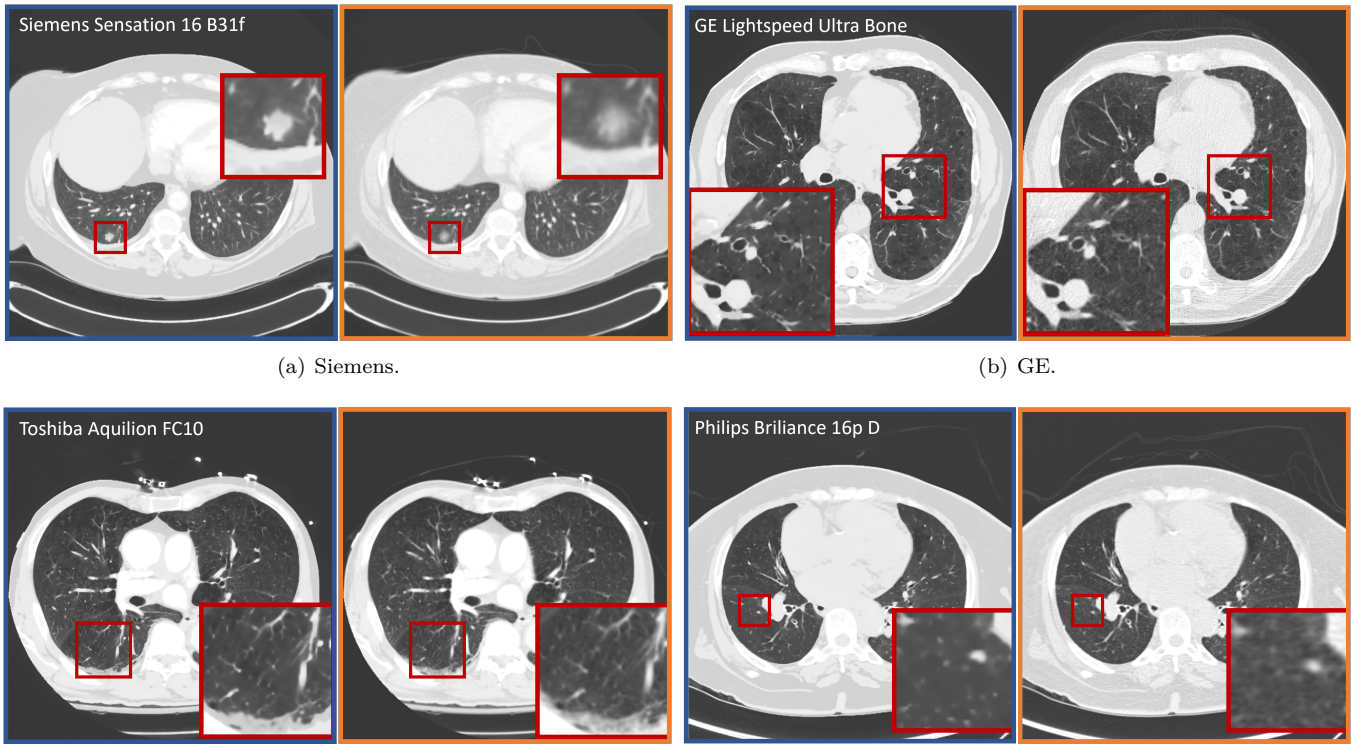
Our study provides compelling evidence of the positive impact of the developed harmonizer on quantitative measurements and clinical evaluations in medical imaging. By improving image sharpness, reducing noise, and enhancing alignment with true values, harmonized images offer a valuable tool for enhancing diagnostic accuracy and enabling more reliable clinical decisions. The findings open avenues for further research and highlight the potential of image harmonization as a transformative technique in the field of medical imaging and radiology.

This research was limited to 2D CT images obtained from 40 XCAT phantoms with COPD and lung nodules. We anticipate better performance by employing volumetric data from a broader range of phantoms that encompass various groups of diseases. Further, a systematic method for measuring MTF within lungs can provide additional advantages, especially when harmonization efforts revolve around lung-related biomarkers. The dataset's diversity, encompassing different noise patterns, reconstruction methods, and kernel properties, enables the network to generalize effectively to unseen data. However, during evaluation, we did not investigate the influence of various metric parameters (e.g., SSIM parameters or COPD feature thresholds) and relied on fixed components such as the segmentation method for morphological feature extraction analysis. Additionally, for generating training and testing data via VIT, we employed a uniform distribution sampling method for the acquisition parameters, potentially constraining the study's findings. Also, some of the variability shown in the non-harmonized images might be less if the acquisition variability is less.

Future studies will seek to assess how unexplored variables affect the effectiveness of harmonization, and explore additional quality measures beyond those employed in this research to highlight distinctions between harmonized and non-harmonized images. Specifically, we will explore alternative metrics to address situations where various rendition of a same patient may yield comparable SSIM values, given its limitations in detecting differences. Moreover, utilizing an active training approach, we aim to incorporate more unseen and rare conditions in the training dataset to minimize the possibility of fabricating unrealistic features or removing relevant clinical features from the harmonized images. Exploring the performance of harmonization combined with other data-driven third-party algorithms, such as trained radiomics feature extraction, are other possible future works.

V. CONCLUSION

The presented results highlight the effectiveness of the proposed algorithm in harmonizing CT images and its



(a) Siemens. (b) GE.

(c) Toshiba. (d) Philips.

Fig. 9. Harmonization results across different vendors. Non-harmonized (in orange square) and the corresponding harmonized (in blue square) images.

positive impact on quantification of image quality metrics, and clinical evaluations. The algorithm successfully reduces bias and variability, leading to more robust and accurate quantification of relevant biomarkers. Furthermore, harmonization enhances image quality by reducing noise and improving sharpness. The harmonizer's adaptability to images from different vendors and its compatibility with existing detection models further suggests its potential for broader clinical applications.

Acknowledgment

We would like to express our sincere gratitude to COPDGene for generously providing the invaluable data used in this study. This work was supported in part by the NIH (P41-EB028744 and R01HL155293).

References

[1] G. Pahn *et al.*, "Toward standardized quantitative image quality (IQ) assessment in computed tomography (CT): A comprehensive framework for automated and comparative IQ analysis based on ICRU Report 87," *Physica Medica*, vol. 32, no. 1, pp. 104–115, 2016.

[2] D. L. Raunig *et al.*, "Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment," *Statistical methods in medical research*, vol. 24, no. 1, pp. 27–67, 2015.

[3] J. Hsieh, *Computed tomography: principles, design, artifacts, and recent advances*. SPIE press, 2003, vol. 114.

[4] M. Zarei *et al.*, "Multi-factorial optimization of imaging parameters for quantifying coronary stenosis in cardiac CT," in *Medical Imaging 2021: Physics of Medical Imaging*, vol. 11595. International Society for Optics and Photonics, 2021, p. 1159504.

[5] C. N. Morris, "Parametric empirical bayes inference: theory and applications," *Journal of the American statistical Association*, vol. 78, no. 381, pp. 47–55, 1983.

[6] J.-P. Fortin *et al.*, "Harmonization of cortical thickness measurements across scanners and sites," *Neuroimage*, vol. 167, pp. 104–120, 2018.

[7] H. Horng *et al.*, "Generalized ComBat harmonization methods for radiomic features with multi-modal distributions and multiple batch effects," *Scientific reports*, vol. 12, no. 1, p. 4493, 2022.

[8] S. Cetin-Karayumak *et al.*, "Exploring the limits of combat method for multi-site diffusion MRI harmonization," *bioRxiv*, pp. 2020–11, 2020.

[9] G. Vegas-Sánchez-Ferrero *et al.*, "Harmonization of chest CT scans for different doses and reconstruction methods," *Medical physics*, vol. 46, no. 7, pp. 3117–3132, 2019.

[10] N. Tanabe *et al.*, "Kernel conversion for robust quantitative measurements of archived chest computed tomography using deep learning-based image-to-image translation," *Frontiers in Artificial Intelligence*, vol. 4, 2021.

[11] M. A. Juntunen *et al.*, "Harmonization of technical image quality in computed tomography: comparison between different reconstruction algorithms and kernels from six scanners," *Biomedical Physics & Engineering Express*, vol. 8, no. 3, p. 037002, 2022.

[12] S. Sotoudeh-Paima *et al.*, "CT-HARMONICA: physics-based CT harmonization for reliable lung density quantification," in *Medical Imaging 2023: Computer-Aided Diagnosis*, vol. 12465. SPIE, 2023, pp. 354–360.

[13] R. Wang *et al.*, "Embracing the disharmony in medical imaging: A simple and effective framework for domain adaptation," *Medical Image Analysis*, vol. 76, p. 102309, 2022.

[14] E. Tzeng *et al.*, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[15] B. Sun *et al.*, "Return of frustratingly easy domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[16] N. K. Dinsdale *et al.*, "Unlearning scanner bias for mri har-

monisation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 369–378.

[17] B. E. Dewey *et al.*, "A disentangled latent space for cross-site MRI harmonization," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 720–729.

[18] L. Wei *et al.*, "Using a generative adversarial network for CT normalization and its impact on radiomic features," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 844–848.

[19] M. Selim *et al.*, "Stan-ct: Standardizing CT image using generative adversarial networks," in *AMIA Annual Symposium Proceedings*, vol. 2020. American Medical Informatics Association, 2020, p. 1100.

[20] S. Marcadent *et al.*, "Generative adversarial networks improve the reproducibility and discriminative power of radiomic features," *Radiology: Artificial Intelligence*, vol. 2, no. 3, p. e190035, 2020.

[21] J. Chen *et al.*, "Improving reproducibility and performance of radiomics in low-dose CT using cycle GANs," *Journal of Applied Clinical Medical Physics*, vol. 23, no. 10, p. e13739, 2022.

[22] J. Lee *et al.*, "Generative adversarial network with radiomic feature reproducibility analysis for computed tomography denoising," *Computers in Biology and Medicine*, vol. 159, p. 106931, 2023.

[23] M. Yang *et al.*, "Wavegan: Frequency-aware gan for high-fidelity few-shot image generation," in *European Conference on Computer Vision*. Springer, 2022, pp. 1–17.

[24] Y. Li *et al.*, "Zero-shot medical image translation via frequency-guided diffusion models," *IEEE transactions on medical imaging*, 2023.

[25] E. Abadi *et al.*, "Modeling lung architecture in the XCAT series of phantoms: physiologically based airways, arteries and veins," *IEEE transactions on medical imaging*, vol. 37, no. 3, pp. 693–702, 2017.

[26] ——, "Dukesim: a realistic, rapid, and scanner-specific simulation framework in computed tomography," *IEEE transactions on medical imaging*, vol. 38, no. 6, pp. 1457–1465, 2018.

[27] ——, "Modeling "textured" bones in virtual human phantoms," *IEEE transactions on radiation and plasma medical sciences*, vol. 3, no. 1, pp. 47–53, 2018.

[28] A. A. of Physicists in Medicine *et al.*, "Adult routine chest CT protocols version 2.0. 2016."

[29] ——, "Lung cancer screening CT protocols, version 5.1. september 13, 2019."

[30] M. Ohkubo *et al.*, "Image filtering as an alternative to the application of a different reconstruction kernel in CT imaging: feasibility study in lung cancer screening," *Medical physics*, vol. 38, no. 7, pp. 3915–3923, 2011.

[31] J. Sanders *et al.*, "Patient-specific quantification of image quality: an automated method for measuring spatial resolution in clinical CT images," *Medical physics*, vol. 43, no. 10, pp. 5330–5338, 2016.

[32] A. G. Roy *et al.*, "Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks," *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 540–549, 2018.

[33] P. Isola *et al.*, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[34] P. J. Huber, "Robust estimation of a location parameter," *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518, 1992.

[35] I. Gulrajani *et al.*, "Improved training of wasserstein gans," *Advances in neural information processing systems*, vol. 30, 2017.

[36] J. Johnson *et al.*, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 694–711.

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

[38] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[39] E. A. Regan *et al.*, "Genetic epidemiology of COPD (COPDgene) study design," *COPD: Journal of Chronic Obstructive Pulmonary Disease*, vol. 7, no. 1, pp. 32–43, 2011.

[40] S. Sotoudeh-Paima *et al.*, "Photon-counting CT versus conventional CT for COPD quantifications: intra-scanner optimization and inter-scanner assessments using virtual imaging trials," in *Medical Imaging 2022: Physics of Medical Imaging*, vol. 12031. SPIE, 2022, pp. 625–633.

[41] D. G. Altman and J. M. Bland, "Measurement in medicine: the analysis of method comparison studies," *Journal of the Royal Statistical Society Series D: The Statistician*, vol. 32, no. 3, pp. 307–317, 1983.

[42] S. G. Armato III *et al.*, "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans," *Medical physics*, vol. 38, no. 2, pp. 915–931, 2011.

[43] S. Richard and J. H. Siewerdsen, "Comparison of model and human observer performance for detection and discrimination tasks using dual-energy x-ray images," *Medical physics*, vol. 35, no. 11, pp. 5043–5053, 2008.

[44] T.-Y. Lin *et al.*, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.