No-reference perceptual CT image quality assessment based on a self-supervised learning framework

IOP Publishing Mach. Learn.: Sci. Technol. 3 (2022) 045033 https://doi.org/10.1088/2632-2153/aca87d

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
6 September 2022

REVISED
25 November 2022

ACCEPTED FOR PUBLICATION
2 December 2022

PUBLISHED
29 December 2022

PAPER

No-reference perceptual CT image quality assessment based on a self-supervised learning framework

Wonkyeong Lee[1,5], Eunbyeol Cho[1,5], Wonjin Kim[1,5], Hyebin Choi[1], Kyongmin Sarah Beck[2], Hyun Jung Yoon[3], Jongduk Baek[4] and Jang-Hwan Choi[1,*]

[1] Division of Mechanical and Biomedical Engineering, Graduate Program in System Health Science and Engineering, Ewha Womans University, Seoul, Republic of Korea
[2] Department of Radiology, Seoul St. Mary's Hospital, College of Medicine, the Catholic University of Republic of Korea, Seoul, Republic of Korea
[3] Department of Radiology, Veterans Health Service Medical Center, Seoul, Republic of Korea
[4] Department of Artificial Intelligence, Yonsei University, Seoul, Republic of Korea
[5] These authors contributed equally to this work.
* Author to whom any correspondence should be addressed.

E-mail: choij@ewha.ac.kr

Keywords: computed tomography, perceptual image quality, radiation dose, self-supervised learning, no-reference image quality assessment

Abstract

Accurate image quality assessment (IQA) is crucial to optimize computed tomography (CT) image protocols while keeping the radiation dose as low as reasonably achievable. In the medical domain, IQA is based on how well an image provides a useful and efficient presentation necessary for physicians to make a diagnosis. Moreover, IQA results should be consistent with radiologists' opinions on image quality, which is accepted as the gold standard for medical IQA. As such, the goals of medical IQA are greatly different from those of natural IQA. In addition, the lack of pristine reference images or radiologists' opinions in a real-time clinical environment makes IQA challenging. Thus, no-reference IQA (NR-IQA) is more desirable in clinical settings than full-reference IQA (FR-IQA). Leveraging an innovative self-supervised training strategy for object detection models by detecting virtually inserted objects with geometrically simple forms, we propose a novel NR-IQA method, named deep detector IQA (D2IQA), that can automatically calculate the quantitative quality of CT images. Extensive experimental evaluations on clinical and anthropomorphic phantom CT images demonstrate that our D2IQA is capable of robustly computing perceptual image quality as it varies according to relative dose levels. Moreover, when considering the correlation between the evaluation results of IQA metrics and radiologists' quality scores, our D2IQA is marginally superior to other NR-IQA metrics and even shows performance competitive with FR-IQA metrics.

1. Introduction

Computed tomography (CT) is one of the fundamental tools for diagnosing patients. One of the most important principles of using CT is to keep the radiation dose to 'as low as reasonably achievable (ALARA)' [1]. To achieve the aim of this principle, we need to optimize the tradeoff between CT image quality and radiation dose. By navigating this tradeoff, it is possible to find the optimal radiation dose with acceptable CT image quality. However, the estimation of CT image quality still suffers from the absence of standard assessment; thus, image quality assessment (IQA) is an active area of research in the field of image processing and technology.

Although research on the development of natural image perceptual quality evaluation with RGB color channels has been actively conducted [2], studies on the assessment of grayscale CT image quality are relatively weak. This is because there are many difficulties in developing CT IQA metrics. First, each modality

of medical imaging has its own characteristics and artifacts that are not found in natural images. Second, there is a lack of datasets for CT IQA, while there exist many publicly available datasets for natural IQA, such as LIVE [3], CSIQ [4], TID2013 [5], and PIPAL [6]. Third, considering radiation-induced risk to the patient, high-dose reference images cannot be easily acquired. This limits the performance testing of developed CT IQA and the development of full-reference IQA (FR-IQA). Last but not least, in medical images, it is difficult to define overall image quality with one quantitative evaluation measure. The reason is that when evaluating medical images, not only the visual quality of an image but also its diagnostic quality must be considered [7]. In clinical practice, diagnostic quality should be evaluated very specifically for each pathology, as local anatomical information to be analyzed in an image differs across each pathology. For example, when radiologists evaluate abdominal CT images, they focus on whether the visibility of major organ structures (e.g. small and large bowel, liver parenchyma, kidney, etc) is suitable for diagnosis [8, 9] rather than placing the same 'attention' on the entire image.

For these reasons, no-reference IQA (NR-IQA) is desirable for CT in clinical settings and should be developed while utilizing the existing datasets, assessing without reference images, and gauging the diagnostic accuracy. Especially regarding the last reason, model observers have been widely researched in CT IQA, which can estimate specific task performance on CT images [10]. However, the existing tasks of model observers are mainly binary classification (e.g. classifying patients as either normal or pathological or signals as absent or present) and do not fully reflect the integrity of local anatomical information, which is different for each pathology across evaluations of diagnosis quality. Therefore, these approaches do not mimic the real task of daily diagnosis assumed by radiologists. Moreover, these task-based models require the synthesis of realistic lesions with complex textures and various sizes incorporating known internal noise and anatomical backgrounds for each organ (e.g. liver [11], lung [12], and breast [13]) and evaluate the quality of a CT image based on the detectability of the inserted lesion. However, the process of generating a lesion by reflecting system- or organ-specific information is difficult to reproduce, and, thus, it is not generally applicable to various organs or less well-known acquisition systems.

In this study, considering the above limitations of existing studies, we propose an automated self-supervision-based NR-IQA metric that is more clinically relevant and easily reproducible. Here, by simulating the experimental setup of image quality evaluation by radiologists, a convolutional neural network (CNN)-based object detection model reads stacks of images containing virtual low-contrast objects. In order to evaluate the quality of a single image, the detector attempts to find the virtual object inserted in the image multiple times, and the resulting mean average precision (mAP) value of these multiple trials is defined as a quantitative value of IQA. At this time, unlike prior model observer studies, not only the parameters (i.e. shape, contrast level, and size) of the virtual objects but also their position and number per image are changed randomly for every trial and unknown to the detection models. Note that the positioning of the object is not completely random, and, considering that radiologists mainly focus on major organ structures with rich textural details rather than uniform regions or background, the object is randomly placed within the found saliency area. In this respect, our proposed method can be said to be more clinically relevant.

Moreover, we model the virtual objects with geometric shapes to be inserted in CT images. Such simple virtual objects can be defined by simple mathematical formulae, and there is no need to consider system-specific noise, organ-specific background, or the complex shape of realistic lesions to generate the virtual objects. Therefore, our approach can generally be applicable to various organs or less well-known acquisition systems and is thus more reproducible.

Lastly, this approach enabled us to implement self-supervised learning to train the detector network, so no forms of annotations or labels are required. Once we trained the network with CT images with the virtual objects inserted, we did not use any high-quality reference images to estimate the IQA score of each image. Considering that a CT image corresponding to the maximum radiation dose cannot be easily obtained in a clinical environment in consideration of patient health, our approach can be more easily applied to clinical practice.

The main contributions of this paper are as follows:

- We present a novel NR-IQA metric for CT images with virtual objects inserted utilizing a CNN-based object detection model. The resulting detection performance (mAP) is converted to an accurate measure of IQA score.
- We propose a self-supervised training strategy for CT IQA by detecting the inserted virtual objects. The objects are of geometrically simple forms and thus can be generally applicable.
- Moreover, the configurations of the virtual objects (i.e. shape, contrast level, size, position, and number per image) change randomly for every detection trial and are unknown to the detection models, lending to greater clinical relevance compared to model observers.

- Rigorous evaluations of clinical and phantom data reveal that our IQA metric showed superior performance over existing NR-IQA metrics and even comparable performance to FR-IQA metrics in terms of correlation with radiologists' perceptions.
- Lastly, to promote research in this field, we have constructed and opened a library of CT images with their associated IQA scores provided by radiologists.

2. Related works

IQA metrics have been developed to measure the perceptual image quality, which can be used to provide scores related to how humans perceive images from the perspective of perceptual quality. They are used by giving the metrics of algorithms images after degradation or post-processing. IQA can be divided into FR-IQA and NR-IQA. FR-IQA methods measure the similarity between an image of interest and a high-quality reference image. They have been widely used in the evaluation of image/video coding, restoration, and communication. Mean-squared error (MSE), peak signal-to-noise ratio (PSNR), and structural similarity (SSIM) [14] are some metrics that are still widely used for evaluation. However, these have limitations because they cannot fully represent human perception. For this reason, data-driven methods were also investigated for IQA [15, 16]. NR-IQA metrics are proposed to assess image quality without reference images because reference images are not always available for IQA. Natural image quality evaluator (NIQE) [17], Ma *et al* [18], blind/referenceless image spatial quality evaluator (BRISQUE) [19], and perception-based IQE (PIQE) [20] are representative NR-IQA methods. Blau *et al* [21] combined FR-IQA and NR-IQA methods to evaluate image restoration algorithms. Even though many algorithms are developed, only a few IQA methods (e.g. PSNR and SSIM) are commonly used for the evaluation of image restoration methods.

CT IQA metrics have been researched considering their own characteristics. MSE, PSNR, and SSIM have been used for CT IQA as basic guidelines for the evaluation of algorithms; however, they do not correlate well with human perception and have little relationship with diagnostic utility [22]. Some classical methods for the estimation of CT IQA are the modulation transfer function and the noise power spectrum (NPS) [23, 24]. These methods assume that the imaging systems are shift-invariant and linear. Thus, if these assumptions are not valid, new methods for CT IQA need to be developed [25, 26].

3. Methods

3.1. Task definition

We propose a CT IQA framework that can be seen as clinically relevant to the simulation of radiologists' IQA tasks in a clinical setup. To mimic the radiologists' IQA task, our model reads stacks of images containing lesion-like virtual objects without knowing their configurations (i.e. shape, contrast, size, location, and number per image). Since our proposed IQA model is based on a deep learning-based detector, we named it the deep detector IQA (D2IQA). The goal of the D2IQA's task is to detect lesion-like virtual objects in CT images that are distorted with an unknown level of noise using a deep learning-based detector. This strategy was motivated by the hypothesis that the accuracy of detecting virtual objects would be highest in high-quality full-dose CT images, and objects distorted due to noise would degrade detection accuracy. In other words, the mAP would be lower for more distorted images. Under this assumption, we set the average mAP as the score of image quality. Additionally, the figure of merit is defined as the correlation coefficient between the mean scores of radiologist evaluations and the scores predicted by D2IQA.

3.2. Overall framework of our proposed model

In this section, we describe the architecture of D2IQA and define our proposed model based on the data, detector, and metrics of CT image quality. The data consists of different contrasts, sizes, shapes, and numbers of virtual objects in arbitrary locations in CT images with various dose levels. Many object detectors are trained with full supervision, but we propose a novel self-supervised task-based assessment method, D2IQA, that does not require any form of ground truth score labels or reference images. Finally, using mAP from the detector, we measure the image quality of CT images with different dose levels. The overall architecture of our proposed method is shown in figure 1. D2IQA consists of two key parts, a data generator and an image quality estimator. We propose a simple but efficient method to automatically generate synthetic lesion-like virtual objects and their associated labels (i.e. object location and shape/class) via a data generator. These self-made labels allow the detector to be trained in a self-supervised manner.

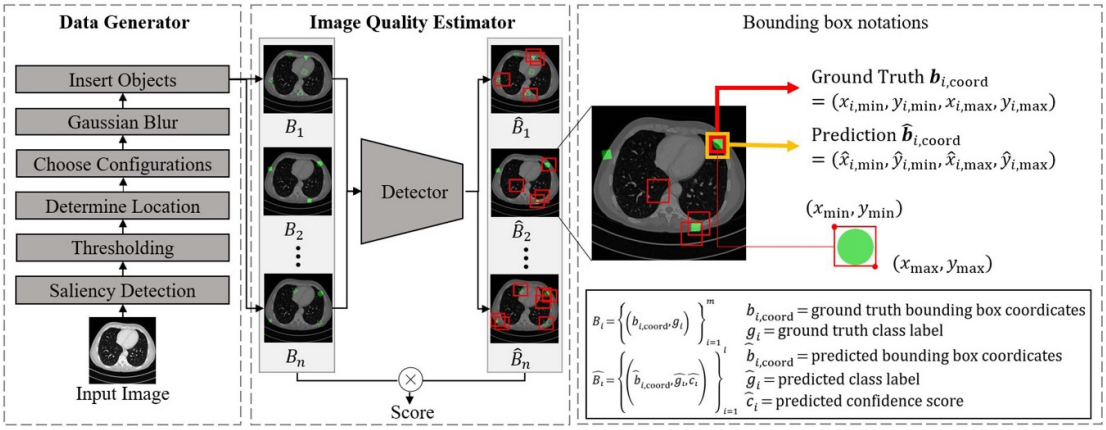


Figure 1. The overall framework of the proposed deep detector image quality assessment (D2IQA). ⊗ represents the mAP calculation between ground truth and the prediction.

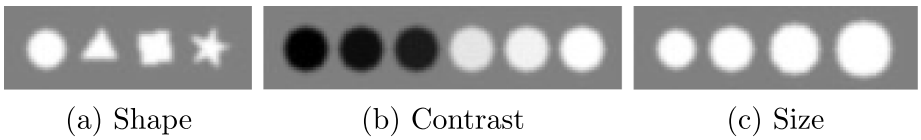


Figure 2. Visualization of shape (a), contrast (b), and size (c) configurations of the virtual objects. The window level and width are set to 0 and 0.2, respectively, for the convenience of visualization.

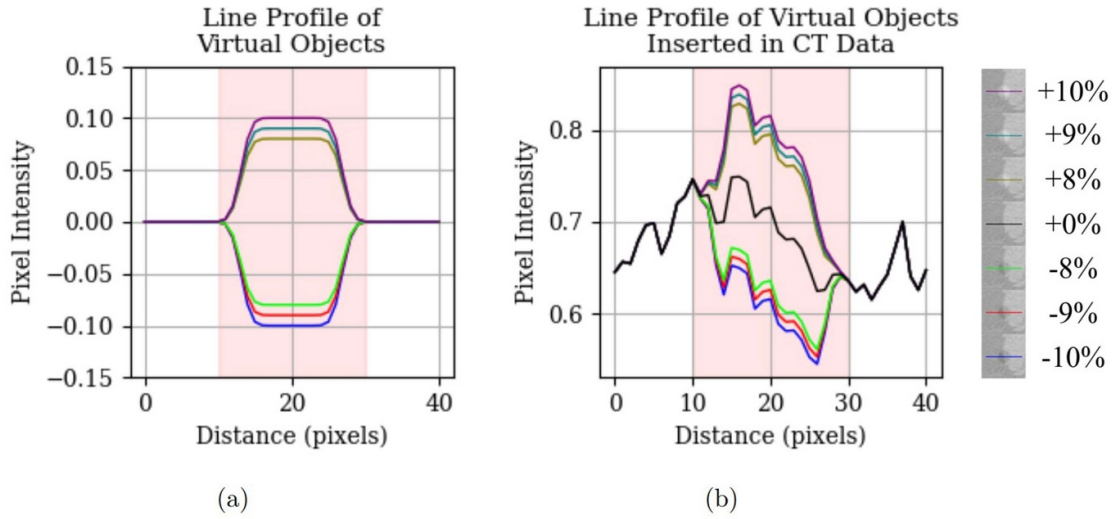


Figure 3. Line profiles comparing the contrast [−10% to +10%] of the virtual objects themselves (a) and that of the virtual objects inserted in CT images (b). (b) The detailed textures in the original image, even with the object overlaid, are preserved.

3.2.1. Data generator

To train our proposed detector to mimic the task of radiologists, we modeled a data generator to insert lesion-like virtual objects in CT images. We considered two things when we modeled the data generator: (a) virtual objects should be simple enough to guarantee adequate model generalization, and (b) it should have complex configurations so that the model can also discover useful feature presentations. A random combination of four different shapes (circle, star, triangle, and square), six different contrast levels (±8%, ±9%, and ±10%), and four different sizes (7, 8, 9, and 10 pixels) was set for each virtual object (see figure 2). We inserted the virtual objects with such configurations, varying at each detection trial, at random positions at a frequency ranging from two to five per CT image. The resulting line profiles of a virtual object inserted in a CT image are illustrated in figure 3. These contrast and size values were chosen empirically to ensure that the score predicted by D2IQA degrades monotonically as the image score goes down.

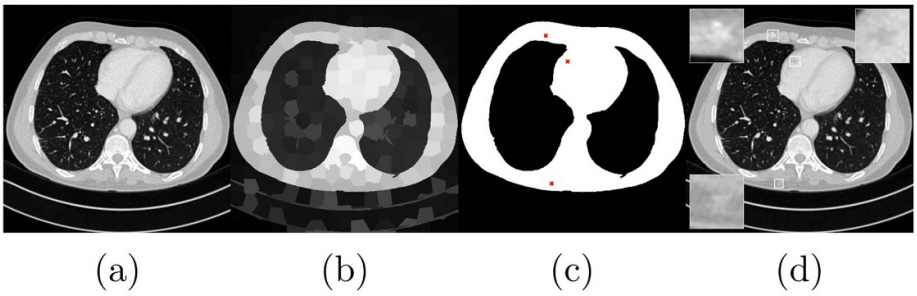


Figure 4. Examples of intermediate result images shown during the virtual object insertion procedure. (a) shows the original image, (b) is the saliency map, (c) shows arbitrary locations (marked with a red ×) of the virtual objects in the binary map, and (d) displays the resulting image with the objects inserted.

Table 1. Measurements of the 2D correlation coefficient (i.e. r) in 2D NPS and SSIM between the original images and the same images with the maximum number (i.e. 5) of virtual objects to evaluate the similarity of the noise texture. The values in the parentheses indicate standard deviations of the measurements.

	Full-dose image	Quarter-dose image
r in 2D NPS	0.999 96 (1.5761×10^{-5})	0.999 97 (1.0568×10^{-5})
SSIM	0.998 97 (2.5171×10^{-5})	0.999 45 (1.7551×10^{-5})

Using these carefully chosen values of contrast and size, the data generator inserts virtual objects. First, a saliency map from each CT image is obtained using the methods developed by Perazzi *et al* [27]. With this obtained saliency map, we discriminate patient anatomical bodies with rich textural details in CT images from the image background. Then, we convert this saliency map to a binary map with a value of 0.4 so that we can identify the anatomical background from the air. The virtual objects will be located in the binary map at high-value indices. The next step is that we randomly choose several configurations for virtual objects including the number of objects (2–5), size (7–10 pixels), and contrast (8%–10 %). We ensured that the virtual objects could not be overlapped by locating each object at least 50 pixels apart from one another. Then, these virtual objects are blurred by Gaussian blur with a 5×5 kernel to make their edges blend well with the background textures. Moreover, this edge blurring process prevents the detector model from being trained to rely on sharp edge textures of high-frequency components that do not exist in human organs. Finally, each virtual object is inserted at the sampled coordinates and rotated at an arbitrary angle.

Figure 4 represents the intermediate results of each step. By inserting artificial lesion-like objects in this way, their corresponding labels (i.e. object location and shape/class) can be automatically obtained, which allows us to train D2IQA in a self-supervised manner. Moreover, considering that the r in 2D NPS and SSIM reported in table 1 are almost 1, the insertion of the low-contrast virtual objects in the image hardly changes the original noise and image texture.

3.2.2. Image quality estimator.

We adopted the existing object detector, Cascade R-CNN [28] with a ResNet-50 [29] backbone, as a detection model for the proposed image quality estimator. The image quality estimator and the existing deep learning-based detector have the same method for detecting an object, but the image quality estimator is trained in a self-supervised way with our data generator.

The inputs of the model are data pairs that consist of $D = [(X_1, B_1), \ldots, (X_n, B_n)]$. X_i represents the CT image containing the virtual object generated by the data generator, and B_i contains 2 to 5 annotations of b, which is the self-made label of image patch x containing the bounding box coordinates b_{coord} and type of each object g. When given an input CT image X_i, the model predicts $\hat{B}_i = (\hat{b_1}, \ldots, \hat{b_n})$. $\hat{b}$ represents the annotation labels for each object; these labels consist of the set of bounding box coordinates $\hat{b}_{\text{coord}} = (\hat{x}_{\min}, \hat{y}_{\min}, \hat{x}_{\max}, \hat{y}_{\max})$, the confidence score $\hat{c}$, and the predicted class label $\hat{g}$. All parameters of the model are updated in a way that minimizes loss by comparing these predictions $(\hat{b})$ to the labels (b).

The detector Cascade R-CNN resamples by the cascade regression of the three R-CNN detection stages using different intersection over union (IoU) values. This cascade regression consecutively resamples objects starting with sample (x_i, b_i) to (x_i', b_i'). Therefore, the loss function for each stage t is the sum of the cross entropy L_{cls} for classification and the smoothed L_1 loss L_{reg} for the bounding box regression of samples produced by different IoU values, as was done in [28]:

$$L(\hat{b}^t, b^t) = L_{\text{cls}}(\hat{g}^t, f_t(\hat{b}^t, b^t)) + h(f_t(\hat{b}^t, b)) L_{\text{reg}}(\hat{b}^t_{\text{coord}}, b^t_{\text{coord}}) \quad (1)$$

in which

$$f_t(\hat{b}, b) = \begin{cases} g & \text{if IoU}(\hat{b}_{\text{coord}}, b_{\text{coord}}) \geqslant u_t, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

$$h(f) = \begin{cases} 1 & \text{if } f \geqslant 1, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $\hat{b}^t$ and b^t are the predicted and the ground truth annotations at stage t, respectively, and u_t is the IoU threshold of stage t, which corresponds to 0.5, 0.6, and 0.7, in the order of detection stages. By using this loss function, we used CT images to fine-tune the Cascade R-CNN model, which was pretrained on the ImageNet dataset.

Note that randomly determined virtual object configurations can sometimes be relatively favorable for the detector to predict. In order to alleviate the effect of some randomly chosen configurations that could bias image quality scores, the detection task was performed multiple times per CT image with different random configurations for each detection trial. Here, the average of the AP values calculated after multiple detection tasks using the Microsoft Common Objects in Context [30] style AP calculation, which utilizes the 101 point interpolation method, are defined as the image quality score.

$$AP = \frac{1}{101} \sum_{r \in R} \max_{\tilde{r} \geqslant r} Precision(\tilde{r}) \tag{4}$$

$$\text{mAP} = \frac{1}{C} \sum_{i}^{C} AP_i, \tag{5}$$

where R is a set of 101 numbers from 0 to 1 at interval of 0.01, and $\tilde{r}$ is the subset of R where each element is greater than the recall value r. In addition, C is 4, and it represents the number of classes (i.e. the number of shapes).

To find the optimal number of the model predictions, we found t satisfying the following condition:

$$\left| \frac{1}{t} \sum_{i}^{t} s_i - \frac{1}{t-1} \sum_{i}^{t-1} s_i \right| < \epsilon \tag{6}$$

where s_i represents the mean of a set of mAPs reflecting 100 CT images with randomly synthesized virtual objects, and t represents the number of sets of repetitions. We empirically obtained that the minimum value of t is 5 when ϵ is 0.01. Thus, we ran 500 predictions on CT images that were differently synthesized from one CT image, which can be computed in parallel. Then, we averaged the resulting mAP values to obtain the quality score for the image.

In this study, Cascade R-CNN was chosen as a detector in the quality estimator. However, to test the generalization ability of D2IQA, we conducted an experiment using three detector models. One of the earliest, mid-term, and recent detector models was selected, which was Faster R-CNN [31], Cascade R-CNN [28], and generalized focal loss [32], respectively. The models scored 4340 images with seven different noise levels obtained from two patients to handle various image qualities. The Kruskal–Wallis [33] test confirmed that the differences in the distribution of the image scores from the three different detectors were not statistically significant ($p \geqslant 0.05$) and the performance of D2IQA was not sensitive to the detector model selection. Therefore, D2IQA is a framework that can be applied to various detector models. Nevertheless, the reason for choosing the Cascade R-CNN model in this study is that the model is relatively sensitive to the difference in image quality, and, as a result, the score difference between noise levels is widely spread.

4. Data and preprocessing

4.1. CT datasets

4.1.1. In-vivo *clinical data*

We used an axial view of abdominal CT images of ten patients from the 2016 Low Dose CT Grand Challenge dataset [34], acquired using multi-detector CT scanners. The full-dose images were acquired using a

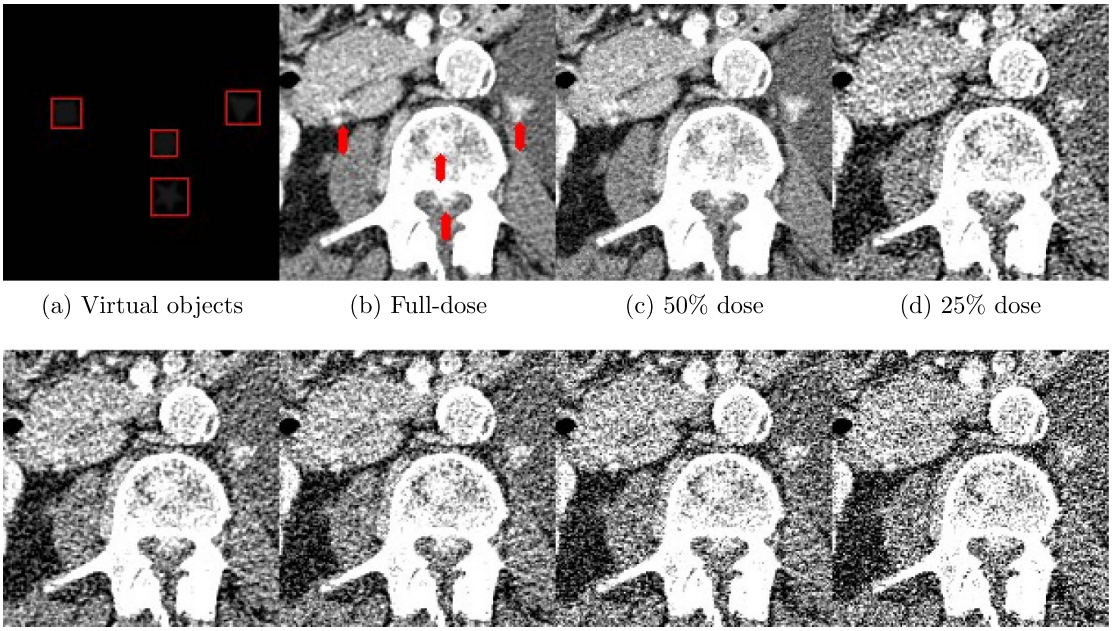


Figure 5. Examples of clinical testing images containing virtually inserted objects. (a) shows examples of virtual objects with only different configurations and locations. The objects in (a) are inserted in the clinical images at varying noise levels (b)–(h). The window level and width are set to 50 and 350.

reference tube potential of 120 kV and a quality reference effective mAs of 200. Because there are only two dose-level CT images (i.e. 100% and 25%) in this *in-vivo* clinical dataset, we followed an accurate physics-based noise generation pipeline [35, 36] to insert Poisson noise into the projection data to achieve seven different noise levels that corresponded to 100%, 50%, 25%, 20%, 12%, 7%, and 5% of the full-dose noise (see section 4.1.2).

Clinical data were split into training, validation, and test sets. Training and validation data were randomly divided from 7988 images from seven patients and finally consisted of 6390 training images and 1598 validation images. For the test data, we used the other two patients, containing a total of 1240 images. The representative clinical testing images with the virtual objects inserted are displayed in figure 5.

4.1.2. Noise generation procedure for in vivo data

Because there are only two dose levels represented in the CT images (i.e. 100% and 25%) from the *in-vivo* clinical dataset, images of various dose levels were created by synthesizing realistic noise with full-dose images. Note that N_0 is the average number of photons emitted during a given time interval. The mean number of transmitted photons is a nonlinear function of the line integral of the attenuation coefficient ($N_\mathrm{d} = N_0 \exp(-\int \mu ds)$), where μ and s are the linear attenuation coefficient and a length element along the photon path, respectively [37]. Then, the line integral in the detector bin i is computed as follows:

$$p_i = \ln \frac{N_0}{\kappa} = \ln N_0 - \ln \kappa \tag{7}$$

where $\kappa = \mathrm{Poisson}(N_\mathrm{d})$ represents the number of transmitted photons at the bin i, and it is a random variable following a Poisson distribution with a mean value N_d.

Following an accurate physics-based noise generation pipeline [35, 36], we inserted Poisson noise into the projection data to obtain five additional noise levels that corresponded to 50%, 20%, 12%, 7%, and 5% of the full-dose. Here, we acquired the projection data corresponding to the line integral by forward-projecting full-dose CT images, resulting in 360° of projection data with the gantry rotated over 2π. The generated noise in the 2D projection data is computed by

$$\tilde{n}_\mathrm{2D} = \mathrm{Poisson}(N_\mathrm{d}) - N_\mathrm{d}. \tag{8}$$

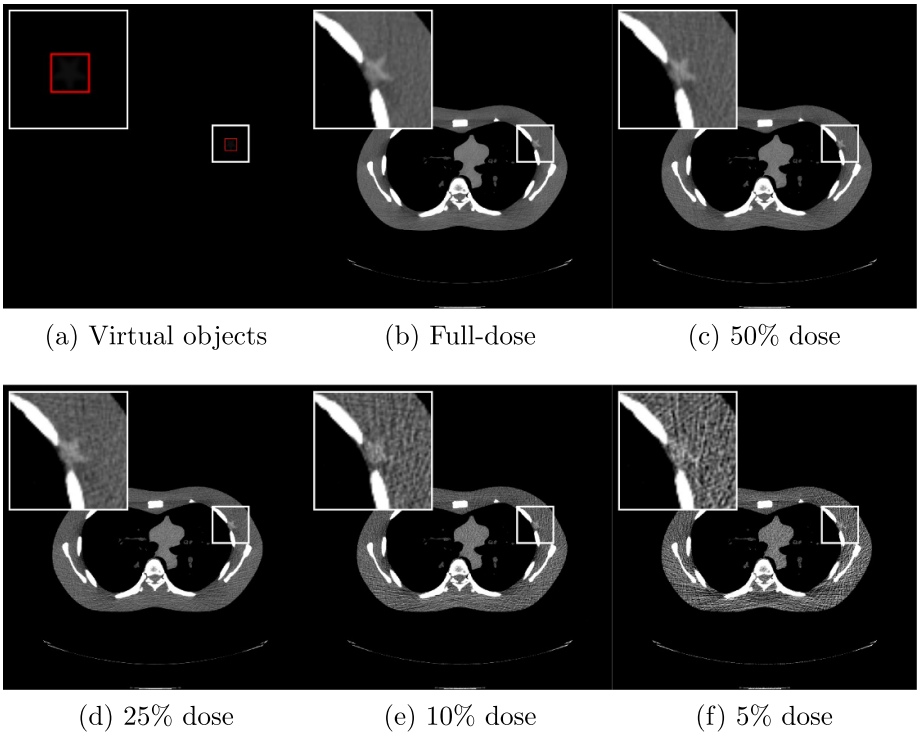


Figure 6. Examples of phantom test images containing a virtually inserted object (a). The images have the same content but contain different levels of noise (b)–(f). The window level and width are set to 50 and 350.

Then, the filtered back-projection algorithm was used to reconstruct the noisy 3D image $\tilde{n}_{3D}$ from $\tilde{n}_{2D}$ as follows:

$$\tilde{n}_{3D} = A^{T} h(\tilde{n}_{2D}), \tag{9}$$

where A^{T} and h denote the back-projector and the reconstruction kernel, respectively. Finally, we added the reconstructed $\tilde{n}_{3D}$ to the original full-dose CT images to obtain CT images with various dose levels. These processes were implemented in the CONRAD open-source package[6]. After we added these additional data to the original dataset, we obtained a dataset with seven different dose levels corresponding to 100%, 50%, 25%, 20%, 12%, 7%, and 5% of the full-dose.

4.1.3. Anthropomorphic phantom data with real noise.
We also used an axial view of CT images from a different domain (i.e. anthropomorphic phantom data with a relatively uniform background and real noise) to demonstrate that D2IQA is well-generalized across noise and anatomy domains and is accurate in quantifying CT image quality. We scanned an anthropomorphic phantom of the chest on a multislice CT scanner (GE). A fixed tube voltage of 120 kV was used for all images. After acquiring a high-dose image using the routine CT acquisition protocols, low-dose image pairs were acquired with a low tube current exposure time product (mAs) values corresponding to five different relative dose levels, including 100%, 50%, 25%, 10%, and 5%. The high-dose data were augmented by fivefold by creating five virtual object-inserted images for every one image to produce a total of 1204 training images and 301 validation images. The test data were unseen during training and model evaluation and consist of 50 images for each dose and 250 images in total. The representative phantom test images containing virtually inserted objects are displayed in figure 6.

4.2. Subjective image quality scores of expert radiologists

To evaluate the correlation between the scores generated by D2IQA and that of humans, we obtained image quality scores from three radiologists with more than ten years of experience reading CT scans. Prior to reading the images, all three radiologists underwent an instruction session on evaluation procedures, and examples of the best (100% dose level) and worst (5% dose level) images were shown to the evaluators so

[6] https://github.com/akmaier/CONRAD

Table 2. A table for image quality scoring criteria.

Numeric score	Verbal descriptive scale	Diagnostic quality criteria
0–1	Bad quality	Desired features are not shown
2–3	Poor quality	Diagnostic interpretation is impossible
4–6	Fair quality	Images are suitable for limited clinical interpretation
7–8	Good quality	Images are good for diagnostic interpretation
9–10	Excellent quality	Anatomical structure is highly visible

that they could establish the scoring criteria for image evaluation. Individual raters evaluated 30 sets of 150 images in total, where each set contained images with different random noise levels of each image. Each CT image was evaluated with abdominal soft-tissue windows (width/level: 350/40). Finally, the image quality scores for each image evaluated by the three radiologists were normalized to their mean and standard deviation and averaged to determine the final human perceptual score for each image.

As to the scoring criteria, we considered that the quality of medical images should be evaluated by reflecting the diagnostic quality of the radiographic images, and, thus, we carefully defined the clinically relevant criteria for the diagnostic IQA. Clinical image quality is often evaluated on a variety of Likert scales, mostly ranging from 3 to 5 [8, 38]. Therefore, a similar Likert scale made by these two studies was chosen for our study with two minor changes. First, diagnosis tasks were added to our scoring criteria since a previous study by Fang *et al* [8] pointed to the absence of specific diagnostic tasks in their subjective assessment as a limitation of their study. According to the opinions of the three radiologists who participated in this study, image noise, anatomical structure, and diagnostic interpretation including lesion detectability [39] were considered in the subjective IQA scoring criteria. Second, we expanded the 5-point Likert scale to a 11-point scale where a higher score represents higher image quality and vice versa. This was done to increase evaluation accuracy. Through this approach, the sensitivity of each clinical image to different image qualities will be better reflected in the scores. The image quality evaluation criteria are described in table 2.

5. Results and discussion

5.1. D2IQA score according to virtual object configuration and noise level

We first verified whether our proposed detector used in D2IQA works reasonably. The meaning of 'reasonably' here is whether our detector model shows a similar trend, just as the detectability of radiologists varies depending on the size and contrast of the virtual object. To be specific, as the size and contrast of the virtual object decreases, the detectability is expected to decrease. Table 3 shows the performance of our detector model according to changes in the size and contrast of objects inserted at fixed positions in the image. As expected, the detector's performance sensitively deteriorates with smaller lesion size and lower contrast, supporting that the detector model in D2IQA works reasonably.

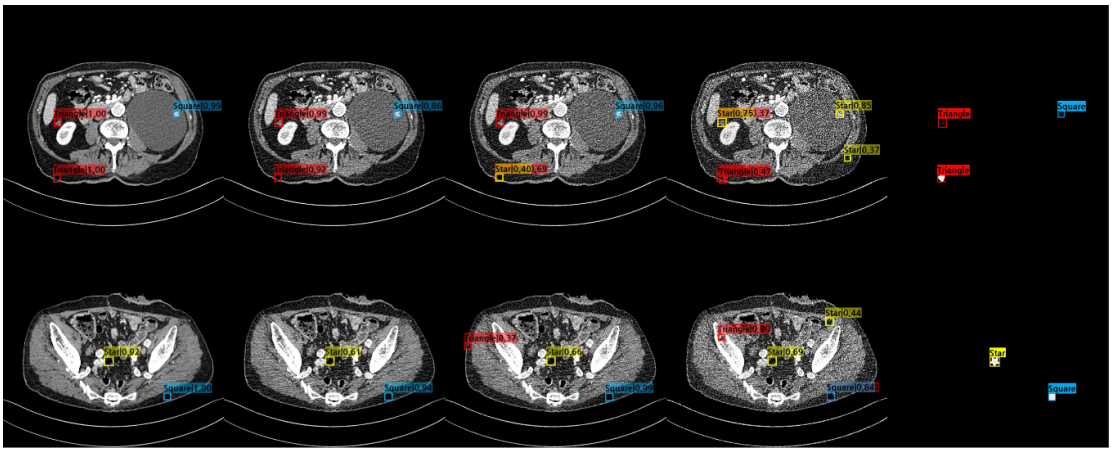
We also found that the performance of our detector gradually decreased as the noise level of the images increased, as shown in figure 7. This result supports our hypothesis that the accuracy of detecting virtual objects would be the highest in high-quality full-dose CT images, and distorted objects due to noise would cause the degradation of detection accuracy. The deterioration of the detector's performance can be explained by false positives and false negatives caused by erroneous prediction of the bounding box position or type (shape) of an object, and these occur relatively frequently in low-quality images. The two representative cases for the four different noise levels (100%, 25%, 12%, and 5% of the full-dose) of the *in-vivo* clinical data in figure 7 show that the lower the quality, the more likely the model is to recognize the virtual object type incorrectly or not to detect the virtual object at all. In addition, as the image quality deteriorates, the confidence in the true positives also gradually decreases. For instance, the prediction of bounding boxes is correct, with a high confidence score in both image cases with 100% of the full-dose (figure 7(a)). However, as the image quality degrades to that of a quarter-dose (figure 7(b)), the confidence scores begin to decrease. In addition, when the image quality degrades to 12% (figure 7(c)), false positive predictions begin to happen, and this phenomenon becomes intensified in CT images corresponding to 5% of the full-dose (figure 7(d)). When the image quality is degraded to 5%, the detector predicts false negatives, struggling to detect the square shape.

5.2. D2IQA performance dependency on the training dataset

This study used an image library containing both full-dose and quarter-dose images to train the detector of D2IQA. However, considering that the radiation dose in each clinical environment is different, we tested the model with different training data sets using cases at both ends of the clinical dose spectrum (i.e. routine full-dose only and quarter-dose only). Here, Pearson linear correlation coefficient (PLCC) and Spearman

Table 3. Results of the detector performance of D2IQA according to contrast and size. Contrast is the mean difference between the virtual object and the background, and size is the size of the object. The darker the color, the better the detection. In D2IQA, the larger the size of the object and the greater the contrast, the better the detection of the object.

Contrast (%)	10				9				8			
Size (pixel)	7	8	9	10	7	8	9	10	7	8	9	10
100%	0.371	0.490	0.549	0.563	0.321	0.449	0.514	0.524	0.255	0.391	0.461	0.470
50%	0.343	0.458	0.524	0.536	0.294	0.419	0.485	0.496	0.231	0.361	0.430	0.440
25%	0.198	0.330	0.408	0.419	0.150	0.280	0.360	0.370	0.101	0.217	0.298	0.315
20%	0.192	0.315	0.389	0.396	0.145	0.267	0.344	0.351	0.098	0.209	0.286	0.297
12%	0.170	0.283	0.360	0.365	0.128	0.240	0.314	0.319	0.084	0.187	0.257	0.264
7%	0.134	0.243	0.311	0.324	0.098	0.198	0.268	0.283	0.062	0.147	0.213	0.232
5%	0.112	0.206	0.268	0.276	0.082	0.164	0.226	0.236	0.052	0.118	0.176	0.190



(a) Full-dose (b) Quarter-dose (c) 12% dose (d) 5% dose (e) virtual object map

Figure 7. Detection performance degradation in two cases. Detection accuracy degrades as the image quality degrades. Only boxes with a confidence score of 0.3 or more are displayed to make the performance degradation noticeable. The window level and width are set to 50 and 350.

rank-order correlation coefficient (SROCC) values were used to evaluate the prediction performance of the model in comparison with the averaged radiologist scores of image quality. The PLCC is calculated as follows:

$$\mathrm{PLCC} = \frac{\sum(\hat{s}-\hat{\mu})(s-\mu)}{\sqrt{\Sigma(\hat{s}-\hat{\mu})^2\Sigma(s-\mu)^2}} \tag{10}$$

where $\hat{s}$ and s are the predicted image quality score and averaged radiologist score, respectively, and $\hat{\mu}$ and μ are the means of the predicted scores and the radiologist scores, respectively. In addition, SROCC is calculated as follows:

$$\mathrm{SROCC} = 1 - \frac{6\Sigma(\hat{S}_i - S_i)^2}{n(n^2-1)} \tag{11}$$

where $\hat{S}_i$ and S_i are the values of the ith element in the sorted list of the predicted image score $\hat{S}$ and the value of the ith element in the sorted list of the radiologist scores S. Moreover, n denotes the number of images. As shown in figure 8, PLCC and SROCC were ranked in the order of 'full-dose only,' 'full- and quarter-doses,' and 'quarter-dose only.' However, all three cases still showed superior performance compared to the other comparative NR-IQA metrics (table 4). These results indicate that our D2IQA can produce robust prediction performance regardless of the dose level of the training dataset, and, therefore, our D2IQA can generally be applicable to various clinical environments with different dose levels.

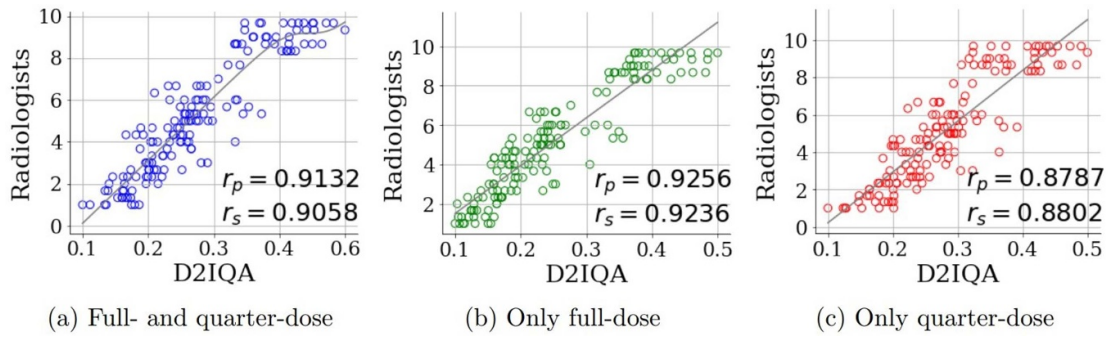


Figure 8. Scatter graph showing the correlation between the scores predicted by D2IQA and those predicted by the radiologists according to the radiation dose of the data used. r_p and r_s are the Pearson linear correlation coefficient and Spearman rank-order correlation coefficient, respectively.

5.3. D2IQA performance comparison with other IQA metrics

A statistical study was also conducted to verify D2IQA's performance in comparison with other representative FR- and NR-IQA methods, including root mean square error (RMSE), PSNR, SSIM, multi-scale (MS)-SSIM [40], gradient magnitude similarity deviation (GMSD) [41], feature similarity index (FSIM) [42], noise quality measure (NQM) [43], visual information fidelity (VIF) [44], NIQE [17], BRISQUE [19], PIQE [20], and non-pre-whitening with eye filter (NPWE) [45]. The model observer NPWE with anthropomorphic channels was considered for the comparison with IQA in the medical domain.

IQA scores predicted by each metric are compared with the mean score estimated by three radiologists, as shown in figure 9, and the PLCC and SROCC values calculated for each metric are summarized in table 4. A nonlinear least squares regression model, as described in equation (12) [3], was used to visualize correlation trends.

$$\text{Quality}(x) = \beta_1 \left\{ \frac{1}{2} - \frac{1}{1+\exp[\beta_2(x-\beta_3)]} \right\} + \beta_4 x + \beta_5, \tag{12}$$

x is the original IQA score, and β are the regression model parameters. Some IQA metrics, such as RMSE, GMSD, NIQE, BRISQUE, and PIQE, are negatively correlated with radiologists' scores. Thus, the x-axis of their corresponding graphs is inverted to facilitate easy comparison with other positively correlated metrics.

From the regression trends shown in figure 9, we observed that FR-IQA methods showed a relatively superior correlation with the radiologists' scores compared to NR-IQA methods. Moreover, some NR-IQA metrics (PIQE and NPWE) failed to show positive correlation patterns with the radiologists' scores. Also, in general, the individual data points of NR-IQA metrics, except for D2IQA, are scattered far apart from the regression graph compared to those of FR-IQA metrics. Individual data points of D2IQA are located close to the regression graph, and the regression graph shows a monotonically increasing pattern similar to that of FR-IQA metrics.

We also quantitatively analyzed the correlation coefficient values as summarized in table 4. FR-IQA metrics all exceed 0.9 in both PLCC and SROCC, but NR-IQA metrics show poor performance, with much lower PLCC and SROCC values compared to FR-IQA metrics. Even PIQE, which shows the highest correlation among the existing NR-IQA metrics, has considerably lower PLCC and SROCC values than all FR-IQA metrics. However, even though our D2IQA is NR-IQA, it shows high correlation values (0.9132 for PLCC and 0.9058 for SROCC) that are marginally better than other NR-IQA metrics and are even comparable with those of existing FR-IQA metrics.

Figure 10 shows the scores of representative IQA metrics versus the radiation dose level of the clinical image. SSIM and NIQE have the highest PLCC and SROCC and are chosen as representative metrics for FR- and NR-IQA metrics, respectively. Because NIQE is negatively correlated with the radiologists' scores, its normalized average IQA score per dose level is subtracted from 1 to show an identical trend with other metrics. One key finding is that, as the dose decreased, the radiologists' scores showed a nonlinear decreasing trend, and our D2IQA most closely mimics this trend among tested metrics.

Another interesting observation in figure 10 is that although SSIM displays the overall trend of decreasing image quality as the dose decreases, it does not correlate well with the radiologists' scores at some dose levels. To be specific, radiologists and D2IQA reported that there was no significant difference in image

Table 4. Correlation coefficient values of each metric. Full-reference and no-reference metrics are used. PLCC and SROCC values are considered for both linear and non-linear relationships.

	Full-reference								No-reference				
Metric	RMSE	PSNR	SSIM [14]	MS-SSIM [40]	GMSD [41]	FSIM [42]	NQM [43]	VIF [44]	NIQE [17]	BRISQUE [19]	PIQE [20]	NPWE [45]	D2IQA
PLCC	−0.9548	0.9570	0.9617	0.9348	−0.9543	0.9105	0.9371	0.9416	−0.7953	−0.7528	−0.7430	0.1864	0.9132
SROCC	−0.9723	0.9601	0.9730	0.9653	−0.9280	0.9315	0.9210	0.9527	−0.8174	−0.7747	−0.6874	0.1036	0.9058

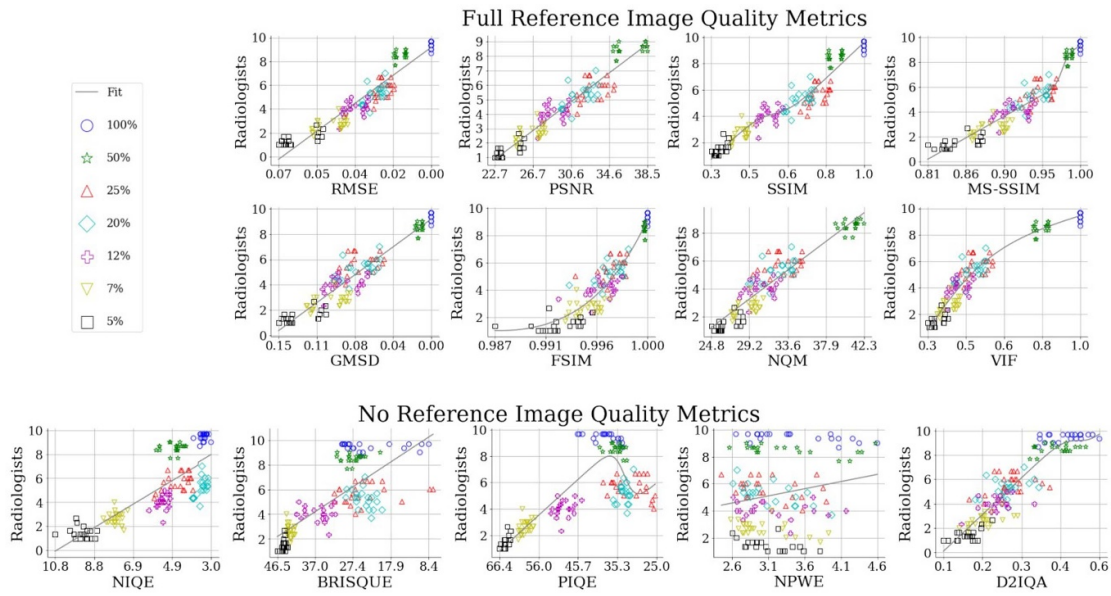


Figure 9. A visualization of the correlation between radiologists' scores and 13 metrics of FR- or NR-IQA. About 150 test images subdivided into seven different noise levels are used. Fitting is done by performing non-linear least squares using equation (12).

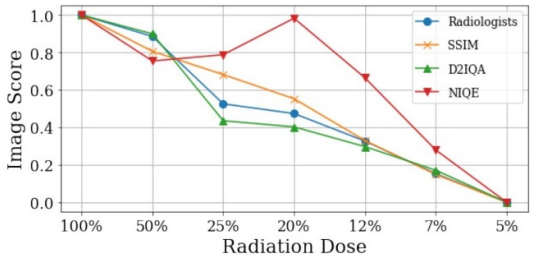


Figure 10. Normalized average IQA scores from radiologists, SSIM, NIQE, and D2IQA are plotted as a function of the noise level.

quality between the 100% and 50% doses, whereas the image quality decreased sharply between the 50% and 25% doses. However, SSIM shows a monotonically decreasing trend as the dose decreases. Similarly, previous studies reported that SSIM and PSNR did not show a strong correlation with radiologists' opinions of diagnostic image quality in the evaluation of magnetic resonance [46] and CT [47] images. This result indicates that although the SSIM metric has been widely accepted in the community, it is necessary to be careful in interpreting the results of SSIM-based diagnostic imaging evaluation. It also suggests that our no-reference D2IQA may be a comparable or better alternative to the full-reference SSIM in the evaluation of CT image quality.

5.4. D2IQA performance on phantom data with real noise

Table 5 shows the predicted image quality score for each metric according to the change in the noise level. Overall, the performance patterns for each metric seen with the clinical images were similarly observed in the phantom images. First of all, FR-IQA metrics, except for NQM and VIF, showed a pattern in which the image quality score changed monotonically as the image quality decreased. However, among the NR-IQA metrics, only NIQE and D2IQA successfully recapitulated the pattern observed in radiologists' scores. Second, among the metrics that successfully showed monotonically changing patterns, we qualitatively confirmed that D2IQA responds most sensitively to image quality changes. To be specific, PSNR and SSIM drop evenly in all intervals of noise level, while D2IQA drops sharply between 25% (70.72 mAs) and 10% (26.52 mAs) dose levels. Considering that less than 40 mAs leads to a drastic degradation in image quality [48], D2IQA's interpretation is reasonable. The drastic deterioration in image quality is also qualitatively recognizable between 25% and 10% dose levels, as shown in figure 6. This result demonstrates that D2IQA is well

Table 5. Quality prediction score by metric for phantom images with various noise levels. Metrics negatively correlated with image quality are indicated by (−). The first, second, and third highest prediction scores for each metric are highlighted in **bold**, underlined, and double underlined, respectively. If the prediction score of a metric decreases monotonically well as the image quality deteriorates, the metric's the highest three scores appear in order from left to right.

		100% (282.88)		50% (141.44)		25% (70.72)		10% (26.52)		5% (13.26)	
Radiation dose (mAs)		mean	cv	mean	cv	mean	cv	mean	cv	mean	cv
FR	RMSE (−)	**0**	—	0.0087	0.2114	0.0135	0.1782	0.0183	0.3219	0.0290	0.3120
	PSNR	—	—	**41.4340**	0.0413	37.5380	0.0386	35.1964	0.0791	31.2007	0.0896
	SSIM [14]	**1**	0	0.9541	0.0226	0.9167	0.0357	0.8152	0.0924	0.6827	0.1417
	MS-SSIM [40]	**1**	0	0.9932	0.0037	0.9871	0.0060	0.9717	0.0160	0.9436	0.0284
	GMSD (−) [41]	**0**	—	0.0106	0.5151	0.0248	0.4164	0.0467	0.5210	0.0856	0.3779
	FSIM [42]	**1**	0	0.9932	0.0027	0.9816	0.0065	0.9545	0.0224	0.9110	0.0377
	NQM [43]	—	—	**41.0434**	0.0455	34.8592	0.0675	36.3521	0.1176	31.3428	0.1474
	VIF [44]	**1**	0	0.7032	0.0976	0.6021	0.0943	0.6156	0.1449	0.5337	0.1725
NR	NIQE (−) [17]	**3.9501**	0.1046	4.0218	0.0900	4.2805	0.0950	5.2763	0.1199	6.1214	0.1579
	BRISQUE (−) [19]	38.1813	0.1507	32.5669	0.1686	**30.5351**	0.1443	36.5365	0.1030	41.0559	0.0543
	PIQE (−) [20]	43.2872	0.1058	**41.6128**	0.0901	42.8507	0.0891	46.3149	0.0820	53.0219	0.0831
	NPWE [45]	2.2915	0.2285	2.2927	0.2339	**2.2928**	0.2356	2.2667	0.2218	2.2418	0.2220
	D2IQA	**0.6932**	0.0952	0.6717	0.0966	0.6259	0.1066	0.4791	0.1975	0.3076	0.3585

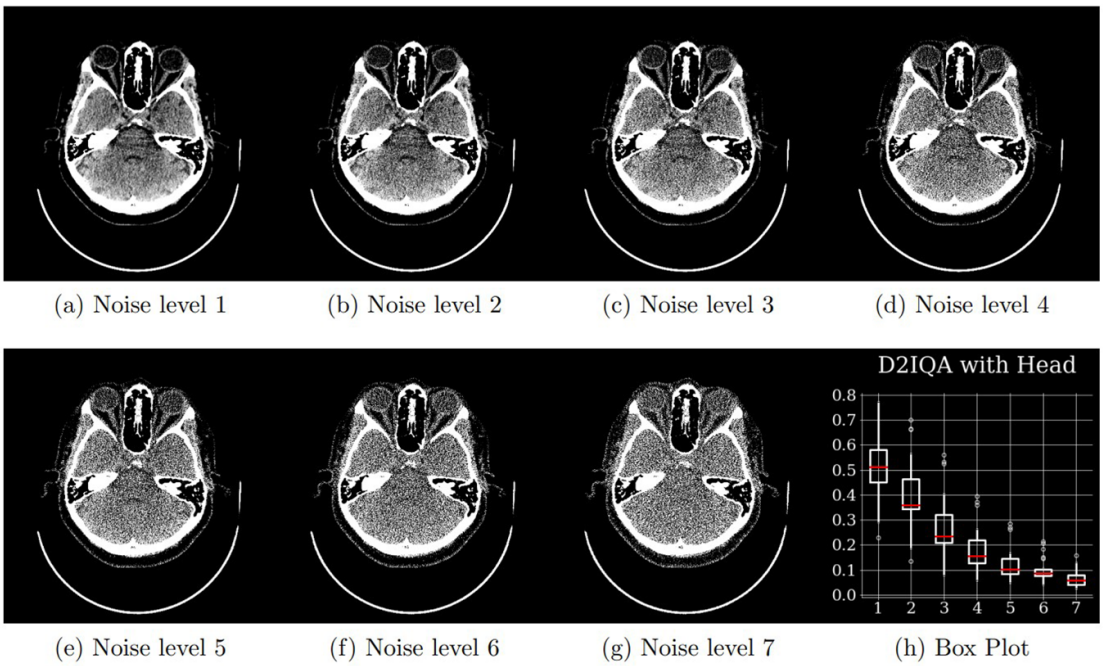


Figure 11. (a)–(g) Visualization of seven different noise levels of *in-vivo* clinical head data. The window level and width are set to 40 and 80. (h) A box plot of scores predicted by D2IQA at seven different noise levels. The *x*-axis and *y*-axis represent the noise level and scores predicted by D2IQA, respectively.

generalized across noise and anatomy domains (i.e. anthropomorphic phantom data with relatively uniform background and real noise) and is accurate in quantifying CT image quality.

5.5. Generalization ability of D2IQA

In this section, we validate the generalization ability of D2IQA. The generalization ability can be divided into two categories, which are generalization to different anatomical parts and generalization to different artifacts. First, in order to validate the ability to generalize to other anatomical parts, an additional clinical non-contrast head CT dataset CQ500 [49] was used to train and test a new D2IQA model. In this dataset, 811 slice images of six patients were used for training and validation, and 22 slice images of one patient were used for testing D2IQA. Because 811 images were insufficient to train D2IQA, the train and validation dataset was augmented fivefold by generating five virtual object-inserted images per slice image, which resulted in 3244 training images and 811 validation images. In addition, In order to acquire various noise levels in the test dataset, the identical noise generation pipeline in section 4.1.2 was used to generate seven different noise levels, which resulted in 154 images in total. Because these images were CT images without contrast and had a relatively simple background, the contrast of virtual objects was reduced to $\pm 6\%$, $\pm 7\%$ and $\pm 8\%$. As seen in figure 11, the scores predicted by D2IQA tend to degrade monotonically as the noise level increases. This indicates that D2IQA can be applied to head images, which suggests the possibility of generalization to other anatomical images.

Secondly, in practice, clinical images are usually distorted by various artifacts. Therefore, to fully address D2IQA's generalization ability to other distortion types, we tested its performance on additional dataset that contained *in-vivo* abdominal clinical images with streak artifacts. Figure 12 shows the *in-vivo* clinical images with seven different levels of streak artifacts and a box plot of scores predicted by D2IQA at each level. After we decreased the projection stack size from 360 to 270, 225, 180, 150, 120, and 90, we obtained test images with seven different levels of streak artifacts. The same D2IQA model that was used in section 5.3 and trained on full- and quarter-dose abdominal images of *in-vivo* clinical data was used to predict the image scores. As shown in figure 12, the image score predicted by D2IQA shows a tendency to drop monotonically as the image quality degrades due to streak artifacts. This result suggests the generalization of D2IQA to other distortion types and its capacity for generalization to actual clinical images.

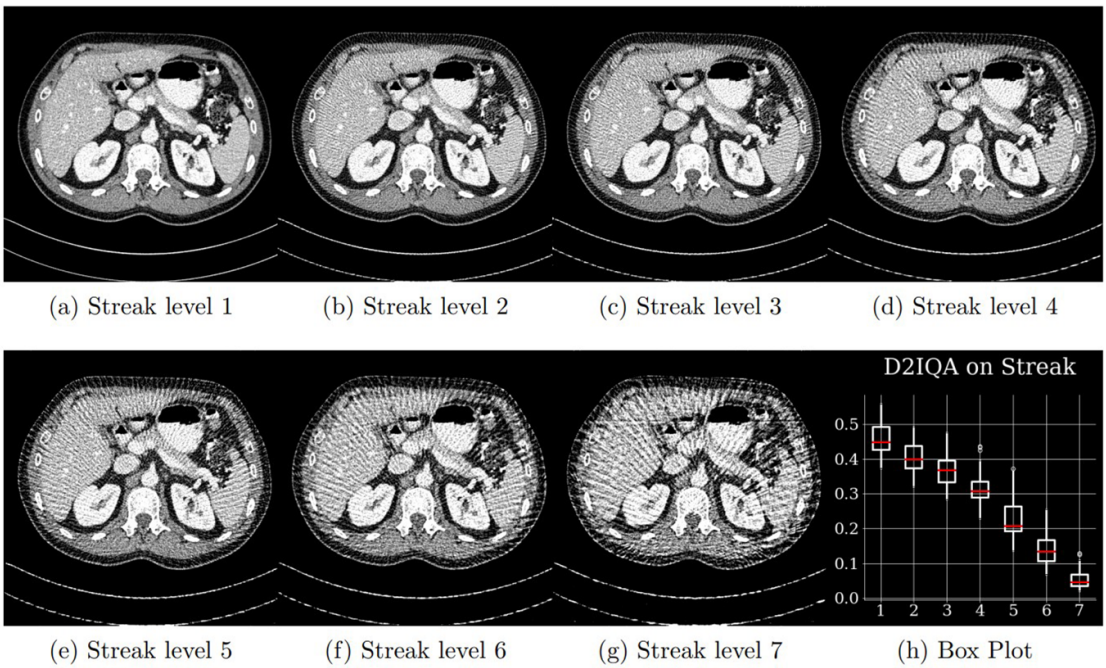


Figure 12. (a)–(g) Visualization of *in-vivo* clinical data with streak artifacts. Seven different streak levels were applied to *in-vivo* clinical data. The window level and width were set to 50 and 350. (h) A box plot of scores predicted by D2IQA on seven different streak levels. The numbers on the *x*-axis and *y*-axis represent the streak level and scores predicted by D2IQA, respectively.

6. Conclusion

By leveraging an innovative self-supervised training strategy for object detection models by detecting virtually inserted objects of geometrically simple forms, our proposed NR-IQA metric, D2IQA, can automatically compute the quantitative quality of CT images at varying dose levels. Rigorous evaluations of clinical and phantom CT image datasets with different domains reveal that our metric showed superior performance over existing NR-IQA metrics and even comparable performance to FR-IQA metrics in terms of correlating with the perception of radiologists'. Until now, the majority of the current IQA metrics for medical images were FR-IQA, but images without degradation are not readily available in clinical practice. Therefore, our D2IQA would make a great contribution to optimizing image post-processing algorithms for diagnosis or new image acquisition techniques. In future research, we plan to extend this study to design a universal medical NR-IQA metric incorporating various image characteristics across different imaging modalities. Lastly, to facilitate research in this field, we have built and opened a library of CT images with their associated IQA scores provided by radiologists.

Data availability statement

The data that support the findings of this study will be openly available following an embargo at the following URL/DOI: https://github.com/Ewha-AI/D2IQA.git. Data will be available from 31 December 2022.

Acknowledgments

This work was partly supported by the Technology development Program of MSS [S3146559], the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2022M3A9I2017587, NRF-2022R1A2C1092072), and by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Nos. 1711174276, RS-2020-KD000016). This work was also partly supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government (21YR2400, Development of image and medical intelligence core technology for rehabilitation diagnosis and treatment of brain and spinal cord diseases).

ORCID iD

Jang-Hwan Choi https://orcid.org/0000-0001-9273-034X

References

[1] Sodickson A, Baeyens P F, Andriole K P, Prevedello L M, Nawfel R D, Hanson R and Khorasani R 2009 *Radiology* **251** 175–84
[2] Gu J, Cai H, Dong C, Ren J S, Qiao Y, Gu S and Timofte R 2021 Ntire 2021 Challenge on perceptual image quality assessment *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 677–90
[3] Sheikh H R, Sabir M F and Bovik A C 2006 *IEEE Trans. Image Process.* **15** 3440–51
[4] Larson E C and Chandler D M 2010 *J. Electron. Imaging* **19** 011006
[5] Ponomarenko N *et al* 2015 *Signal Process., Image Commun.* **30** 57–77
[6] Jinjin G, Haoming C, Haoyu C, Xiaoxing Y, Ren J S and Chao D 2020 Pipal: a large-scale image quality assessment dataset for perceptual image restoration *European Conf. on Computer Vision* (Springer) pp 633–51
[7] Cavaro-Ménard C, Zhang L and Le Callet P 2010 Diagnostic quality assessment of medical images: challenges and trends *2010 2nd European Workshop on Visual Information Processing (EUVIP)* (IEEE) pp 277–84
[8] Fang T, Deng W, Law M W M, Luo L, Zheng L, Guo Y, Chen H and Huang B 2018 *Brit. J. Radiol.* **91** 20170448
[9] Speelman E, Brocx B, Wilbers J, de Bie M, Ivashchenko O, Tank Y and van der Molen A 2020 *Emerg. Radiol.* **27** 141–50
[10] Barrett H H, Yao J, Rolland J P and Myers K J 1993 *Proc. Natl Acad. Sci.* **90** 9758–65
[11] Gong H, Fletcher J G, Heiken J P, Wells M L, Leng S, McCollough C H and Yu L 2022 *Med. Phys.* **49** 70–83
[12] Sauer T J, Abadi E, Solomon J, Hoye J M and Samei E 2018 Realistic lesion simulation: application of hyperelastic deformation to lesion-local environment in lung CT *Proc. SPIE* **10573** 105731U
[13] Han M, Kim B and Baek J 2018 *PLoS One* **13** e0194408
[14] Wang Z, Bovik A, Sheikh H and Simoncelli E 2004 *IEEE Trans. Image Process.* **13** 600–12
[15] Prashnani E, Cai H, Mostofi Y and Sen P 2018 Pieapp: perceptual image-error assessment through pairwise preference *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 1808–17
[16] Zhang R, Isola P, Efros A A, Shechtman E and Wang O 2018 The unreasonable effectiveness of deep features as a perceptual metric *2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 586–95
[17] Mittal A, Soundararajan R and Bovik A C 2013 *IEEE Signal Process. Lett.* **20** 209–12
[18] Ma C, Yang C Y, Yang X and Yang M H 2017 *Comput. Vis. Image Underst.* **158** 1–16
[19] Mittal A, Moorthy A K and Bovik A C 2012 *IEEE Trans. Image Process.* **21** 4695–708
[20] Venkatanath N, Praneeth D, Maruthi Chandrasekhar B, Channappayya S S and Medasani S S 2015 Blind image quality evaluation using perception based features *2015 Twenty First National Conf. on Communications (NCC)* pp 1–6
[21] Blau Y and Michaeli T 2018 The perception-distortion tradeoff *2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 6228–37
[22] Erickson B J 2002 *J. Digit. Imaging* **15** 5–14
[23] Barrett H H and Myers K J 2013 *Foundations of Image Science* (New York: Wiley)
[24] Boedeker K L, Cooper V N and McNitt-Gray M F 2007 *Phys. Med. Biol.* **52** 4027
[25] Hudson H and Larkin R 1994 *IEEE Trans. Med. Imaging* **13** 601–9
[26] Hara A K, Paden R G, Silva A C, Kujak J L, Lawder H J and Pavlicek W 2009 *Am. J. Roentgenol.* **193** 764–71
[27] Perazzi F, Krähenbühl P, Pritch Y and Hornung A 2012 Saliency filters: contrast based filtering for salient region detection *2012 IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE) pp 733–40
[28] Cai Z and Vasconcelos N 2018 Cascade R-CNN: delving into high quality object detection *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 6154–62
[29] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 770–8
[30] Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L 2014 Microsoft COCO: common objects in context *European Conf. on Computer Vision* (Springer) pp 740–55
[31] Ren S, He K, Girshick R and Sun J 2015 *Adv. Neural Inf. Process. Syst.* **28** 1–9
[32] Li X, Wang W, Wu L, Chen S, Hu X, Li J, Tang J and Yang J 2020 *Adv. Neural Inf. Process. Syst.* **33** 21002–12
[33] Kruskal W H and Wallis W A 1952 *J. Am. Stat. Assoc.* **47** 583–621
[34] AAPM 2017 Low dose CT grand challenge (available at: www.aapm.org/GrandChallenge/LowDoseCT/)
[35] Gholizadeh-Ansari M, Alirezaie J and Babyn P 2020 *J. Digit. imaging* **33** 504–15
[36] Kim B, Shim H and Baek J 2021 *Med. Image Anal.* **71** 102065
[37] Macovski A 1983 *Medical Imaging Systems* (Hoboken, NJ: Prentice–Hall)
[38] Beenen L F, Sierink J C, Kolkman S, Nio C Y, Saltzherr T P, Dijkgraaf M G and Goslings J C 2015 *Acta Radiol.* **56** 873–80
[39] Singh R *et al* 2020 *Am. J. Roentgenol.* **214** 566–73
[40] Wang Z, Simoncelli E and Bovik A 2003 Multiscale structural similarity for image quality assessment *The 37th Asilomar Conf. on Signals, Systems Computers, 2003* vol 2 pp 1398–402
[41] Xue W, Zhang L, Mou X and Bovik A C 2014 *IEEE Trans. Image Process.* **23** 684–95
[42] Zhang L, Zhang L, Mou X and Zhang D 2011 *IEEE Trans. Image Process.* **20** 2378–86
[43] Damera-Venkata N, Kite T, Geisler W, Evans B and Bovik A 2000 *IEEE Trans. Image Process.* **9** 636–50
[44] Sheikh H and Bovik A 2006 *IEEE Trans. Image Process.* **15** 430–44
[45] Burgess A 1994 *J. Opt. Soc. Am.* A **11** 1237–42
[46] Mason A, Rioux J, Clarke S E, Costa A, Schmidt M, Keough V, Huynh T and Beyea S 2019 *IEEE Trans. Med. Imaging* **39** 1064–72
[47] Choi D, Kim W, Lee J, Han M, Baek J and Choi J H 2021 *Mach. Vis. Appl.* **32** 1–14
[48] Yan H, Cervino L, Jia X and Jiang S B 2012 *Phys. Med. Biol.* **57** 2063
[49] Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau N G, Venugopal V K, Mahajan V, Rao P and Warier P 2018 arXiv:1803.05854