Deep-learning-based model observer for a lung nodule detection task in computed tomography

Hao Gong, Qiyuan Hu, Andrew Walther, Chi Wan Koo, Edwin A. Takahashi, David L. Levin, Tucker F. Johnson, Megan J. Hora, Shuai Leng, Joel G. Fletcher, Cynthia H. McCollough, and Lifeng Yu*

Mayo Clinic, Department of Radiology, Rochester, Minnesota, United States

Abstract

Purpose: Task-based image quality assessment using model observers (MOs) is an effective approach to radiation dose and scanning protocol optimization in computed tomography (CT) imaging, once the correlation between MOs and radiologists can be established in well-defined clinically relevant tasks. Conventional MO studies were typically simplified to detection, classification, or localization tasks using tissue-mimicking phantoms, as traditional MOs cannot be readily used in complex anatomical background. However, anatomical variability can affect human diagnostic performance.

Approach: To address this challenge, we developed a deep-learning-based MO (DL-MO) for localization tasks and validated in a lung nodule detection task, using previously validated projection-based lesion-/noise-insertion techniques. The DL-MO performance was compared with 4 radiologist readers over 12 experimental conditions, involving varying radiation dose levels, nodule sizes, nodule types, and reconstruction types. Each condition consisted of 100 trials (i.e., 30 images per trial) generated from a patient cohort of 50 cases. DL-MO was trained using small image volume-of-interests extracted across the entire volume of training cases. For each testing trial, the nodule searching of DL-MO was confined to a 3-mm thick volume to improve computational efficiency, and radiologist readers were tasked to review the entire volume.

Results: A strong correlation between DL-MO and human readers was observed (Pearson's correlation coefficient: 0.980 with a 95% confidence interval of [0.924, 0.994]). The averaged performance bias between DL-MO and human readers was 0.57%.

Conclusion: The experimental results indicated the potential of using the proposed DL-MO for diagnostic image quality assessment in realistic chest CT tasks.

 [DOI: 10.1117/1.JMI.7.4.042807]

Keywords: model observer; deep learning; lung nodule detection; x-ray computed tomography; task based image quality assessment.

Paper 19236SSR received Sep. 9, 2019; accepted for publication Jun. 15, 2020; published online Jun. 30, 2020.

1 Introduction

Objective and quantitative image quality assessment is critical for optimizing radiation dose and scanning protocols in computed tomography (CT). The reference standard using multicase multireader studies is frequently challenged by intra- and interobserver performance variability and intensive resource requirements. For example, it is usually very expensive and time-consuming to collect sufficient positive cases with target pathology and establish ground truth via clinical follow-up.[1] Further, traditional image quality metrics (e.g., modulation transfer

*Address all correspondence to Lifeng Yu, E-mail: Yu.Lifeng@mayo.edu

2329-4302/2020/$28.00 © 2020 SPIE

function and noise power spectrum) are inappropriate in the newer CT systems using iterative reconstruction (IR) or other algorithms that involve nonlinear postprocessing,[2–5] and cannot serve as the complete descriptors of CT diagnostic image quality. In contrast, mathematical model observers (MOs) have demonstrated potential as a task-based image quality descriptor, i.e., predictor of human reader diagnostic performance, once a statistically significant correlation can be determined between MOs and human readers for clinically relevant CT tasks.

The performance of traditional MOs (e.g., channelized Hotelling observers) had been typically validated in simplified object detection tasks that used uniform phantom background and artificial lesion-mimicking inserts as the surrogate of patient CT exams.[6–8] To date, it is not yet clear if the traditional MOs validated in the phantom studies can be readily applied for radiation dose and scanning protocol optimization in various routine CT tasks. To achieve greater clinical realism, multiple prior studies have presented statistical MOs incorporating models of human visual processing. For instance, Gifford validated visual-search MOs that involved a two-stage holistic search process, using simulated emission tomography;[9,10] Lago et al.[11,12] proposed a foveated channelized Hotelling observer that involved the modeling of signal detectability across the visual field and validated it using simulated breast tomosynthesis. However, it is still unclear if these improved statistical MOs can be readily extended for x-ray CT tasks involving real human anatomy. Of note, human reader performance is affected by patient anatomical variability (e.g., soft-tissue heterogeneity) across different types of diagnostic tasks.[13–16] Meanwhile, it is challenging to acquire sufficient statistics of varying anatomical background from a general population, to use traditional MOs in patient CT images.

The emerging deep-learning-based MOs (DL-MOs) may provide an alternative solution to this problem, using the state-of-the-art convolutional neural networks (CNN). Alnowami et al.,[17] Murphy et al.,[18] and Massanes and Brankov[19] recently evaluated several DL-MOs in the simulated and/or clinical mammograms. Further, Zhou et al.[20] employed CNNs to approximate ideal observer and Hotelling observer for a binary signal detection task in the simulated mammograms. Meanwhile, Kopp et al.[21] validated a DL-MO in a CT phantom study that involved the detection of lesion-mimicking inserts in uniform phantom background. In these prior studies, DL-MOs were constructed using relatively shallow CNN models and validated using simplified detection tasks that lacked sophisticated anatomical background (e.g., patient chest/abdomen CT images) or complex visual searching process. Therefore, it is not clear if these DL-MOs can be readily applied in more challenging scenarios that involve complex patient anatomy and realistic CT tasks. To overcome this issue, we have recently proposed a DL-MO framework that was based on the ensemble modeling of a deep CNN model and a traditional statistical learning technique. We validated this method in a two alternative forced choice experiment that involved low-contrast liver metastases inserted in patient liver background.[22,23]

In this study, we proposed a modified DL-MO framework for more realistic lesion localization tasks, by incorporating sliding window strategy and nodule-searching process into our prior DL-MO framework. Then we proceeded to validate this new DL-MO framework in a virtual clinical trial that involved a realistic lung nodule localization task in patient chest background. To generate a large number of realistic and positive cases with known truth, our previously validated projection-based lesion-/noise-insertion techniques[24–27] were used. The preliminary results have been recently reported.[28,29] In this paper, we present the complete methodology with comprehensive experimental validation.

2 Method

2.1 *DL-MO Framework for a Localization Task*

The generic framework of the proposed DL-MO is illustrated in Fig. 1. This DL-MO framework includes four major components: a pretrained deep CNN, a partial least square regression discriminant analysis (PLS-DA) model, an internal noise component, and a nodule searching process. The first three components were similar to the ones in our prior work.[23] The main difference of the current framework is that it combined with a sliding window strategy[8] to

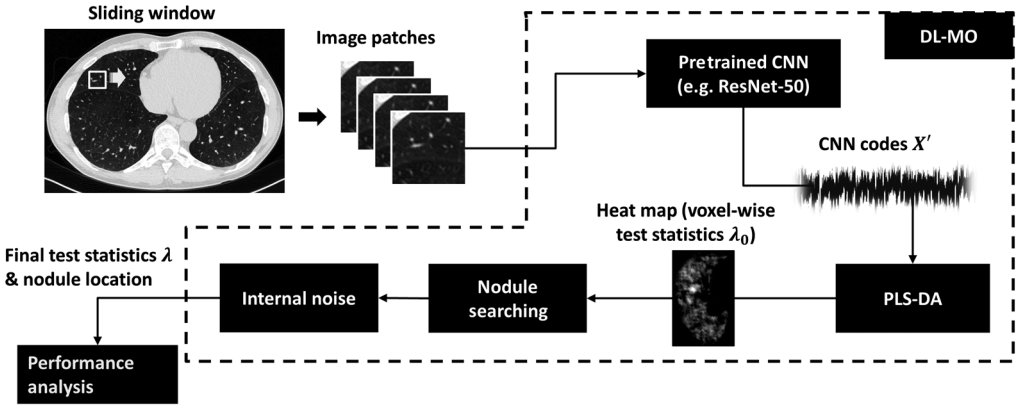

Fig. 1 The framework of the proposed DL-MO, including a pretrained deep CNN, a PLS-DA model, nodule searching process, and an internal noise component. DL-MO, deep learning-based model observer; CNN, convolutional neural network; and PLS-DA, partial least square regression discriminant analysis.

generate the test statistics across all potential nodule locations and then a nodule searching process was added to the DL-MO to determine the most likely location of lung nodules. Below we provide more details on this new framework, focusing on the component of nodule searching process. For further details of the other three major components, one can refer to Ref. 23.

The sliding window strategy was used to extract local image patches that were used as the inputs to DL-MO. The image patches were fed into a 50-layer residual net (i.e., ResNet50)[30] that was pretrained on a natural image database, i.e., ImageNet,[31] to extract image feature maps X' (termed as "CNN codes") from a preselected intermediate layer of the CNN. PLS-DA model was used to further engineer the CNN codes and generate the test statistics λ_0 without the internal noise. The test statistics λ_0 was calculated as the inner product between CNN codes X' and PLS regression coefficient vector B that was acquired by training PLS-DA model. Of note, λ_0 was assigned to the central voxel of each input image patch. The complete spatial distribution of λ_0 (termed as "heat map") was acquired after the sliding window scrolled through all potential nodule locations. A nodule searching process was used to identify the location of the voxel that coincided with the maximal value of λ_0, i.e., the most-likely location of lung nodules. Finally, an internal noise component was added to the maximal λ_0 to model the variation of human reader performance, i.e., $\lambda = \lambda_0 + \alpha \cdot x$, where λ denotes the final test statistics, α is the weighting factor, and x is a Gaussian random variable with a zero expectation and the same standard deviation as the test statistics $\lambda_{0,\mathrm{bkg}}$ of nodule-absent images.

2.2 *Patient Data Preparation*

2.2.1 *Case selection*

Fifty low-dose lung cancer screening patient CT exams were retrospectively selected from our clinical data registry. The inclusion criteria included that: all patients agreed to the use of medical records for research purpose; raw projection data were archived in our data registry; all CT exams were acquired from a 192-slice dual-source CT system (SOMATOM Force, Siemens Healthineers, Germany); and the midlevel lungs (middle lobe/lingula, within an ~3-cm range) were nodule-free. All cases that did not meet the inclusion criteria were excluded. In the original patient cohort, each patient was scanned twice, using a routine low-dose CT protocol [120 kV without tin filter, 50 quality-reference-mAs (QRM), and nominal CTDIvol 3.6 mGy] and an ultralow-dose CT protocol (100 kV with an additional tin filter, 50 QRM, nominal CTDIvol 0.17 mGy), respectively. However, only the projection data acquired with the routine low-dose protocol were used in this study since the radiation dose was relatively higher and several lower radiation dose levels would be simulated based on this relatively higher dose level.

Table 1 The configuration for all experimental trials.

Index	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12
Type	GGN	GGN	GGN	GGN	PSN	PSN	PSN	PSN	GGN	GGN	GGN	GGN
Size	5.4	5.4	5.4	5.4	3.4	3.4	3.4	3.4	3.4	7.4	5.4	5.4
Dose (%)	10	25	50	100	10	25	50	100	50	50	50	50
Recon	IR-2	IR-2	IR-2	IR-2	IR-2	IR-2	IR-2	IR-2	IR-2	IR-2	FBP	IR-4

Note. Type: GGN, ground glass nodule and PSN, partially solid nodule.
Size: the effective nodule diameter in mm.
Dose: the percentage relative to routine radiation dose level.
Recon: IR, iterative reconstruction (ADMIRE) using a medium sharpness kernel with the strength setting of 2 (i.e., IR-2) or 4 (i.e., IR-4) and FBP, filtered back projection with a medium sharpness kernel.

2.2.2 *Simulation of patient cases with multiple dose levels and nodule conditions*

We synthesized virtual chest CT exams to compare the performance of DL-MO and human readers across 12 experimental conditions. These conditions involved three lung nodule sizes, two lung nodule types, four radiation dose levels, and three image reconstruction types (Table 1). The procedure of lesion-/noise-insertion is illustrated in Fig. 2. Poisson noise was inserted into patient raw projection data to simulate chest CT exams acquired at additional lower radiation dose levels [10%, 25%, and 50% of routine dose (RD)], using our previously validated projection-domain noise insertion tool.[27] Moreover, the forward projections of lung nodule CT images were added to patient raw projection data to synthesize nodule-present cases, using our previously validated projection-based lesion insertion tool.[24,26] The real CT images of a ground-glass nodule (GGN, 5.4 mm, −660 HU) and a partially solid nodule (PSN, 3.4 mm, −442 HU) were selected as the base lung nodule models. The GGN images were numerically modified to generate two additional nodule sizes (3.4 and 7.4 mm). For each nodule-present case, one lung

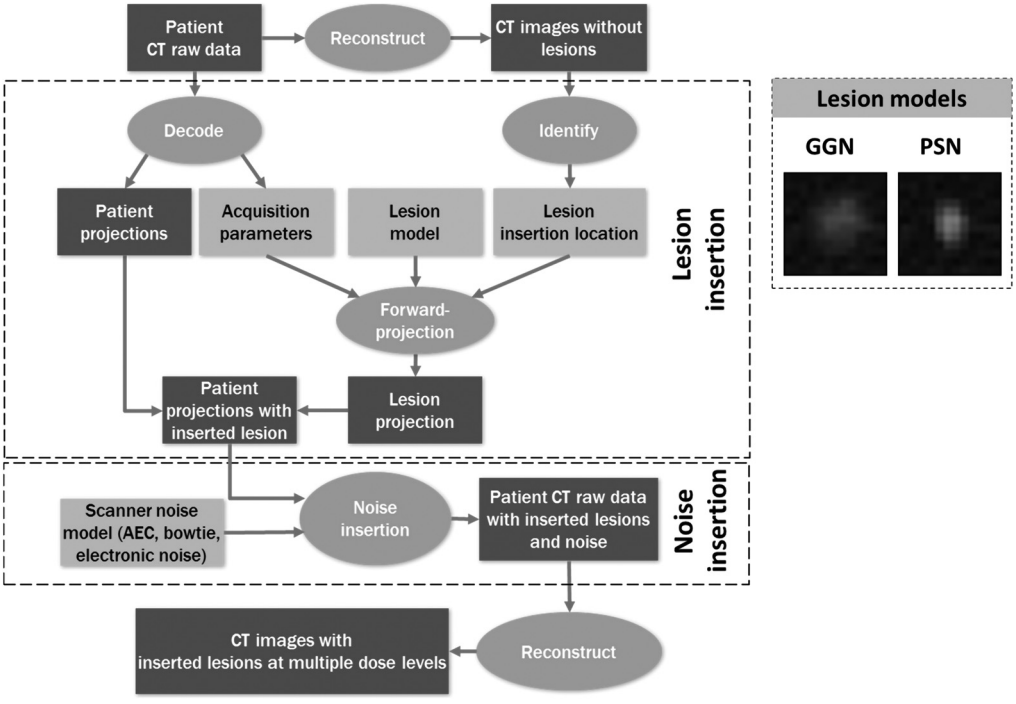

Fig. 2 The schematic illustration of the process of synthesizing lesion-present and lower-dose cases, using existing raw patient projection data. Of note, lesion model refers to nodule images that were numerically modified from real nodule CT images. The inset illustrates example CT images of two types of nodules. GGN, ground glass nodule and PSN, partially solid nodule.

nodule was inserted at a preselected location at each side of chest. The location selection was guided by a supervising radiologist who did not participate the reading sessions of virtual CT exams. The following principles were used to determine nodule location: (1) the nodule abutted a small blood vessel (solid, but fainter, white lines); (2) the nodule should not cross an airway or fissure; (3) a nodule should not be placed against the margins of the lung; and (4) if possible, the oval nodules were oriented so the long axis is oriented radially from the center of the chest.

After lesion- and noise-insertions were completed, the modified patient projection data were used to reconstruct CT images with respect to varying experimental conditions. Image reconstruction used filtered back-projection (FBP) and a commercial IR algorithm [Advanced Modeled Iterative Reconstruction (ADMIRE), Siemens Healthcare, Germany] with two strength settings (2 and 4). Each experimental trial included 30 CT images of the middle level of one nodule-present/-absent lung. Each experimental condition involved 70 nodule-present and 30 nodule-absent trials.

2.3 *DL-MO Study*

DL-MO was constructed separately for each experimental condition, using the CNN codes from the 26th convolutional layer and the most significant 20 PLS components. The repeated two-fold cross validation (R2f-CV) was used to estimate the generalizability of DL-MO. The R2f-CV method is briefly explained as follows. The stratified random sampling was first used to split the patient exams to two equal-sized subgroups (i.e., $n = 25$ cases per subgroup). One subgroup was used to train DL-MO while the other was used for validation. Then the two subgroups were swapped to retrain and revalidate DL-MO, respectively. In this study, this process was repeated twice, and then DL-MO performance was averaged across all validation subgroups to estimate the generalization performance of DL-MO. Of note, the training of DL-MO used the volume-of-interests (VOIs) extracted across the entire available image volume of those training cases (i.e., 30 slices per trial per condition). Further, the testing procedure was simplified by confining the nodule searching process within a 3-mm-thick volume (Fig. 3), to reduce the computational time. For nodule-present trials, DL-MO response was only calculated across three consecutive images that coincided with the central region of lung nodules (along the z dimension), instead of through the entire volume. For nodule-absent trials, DL-MO response was calculated over three

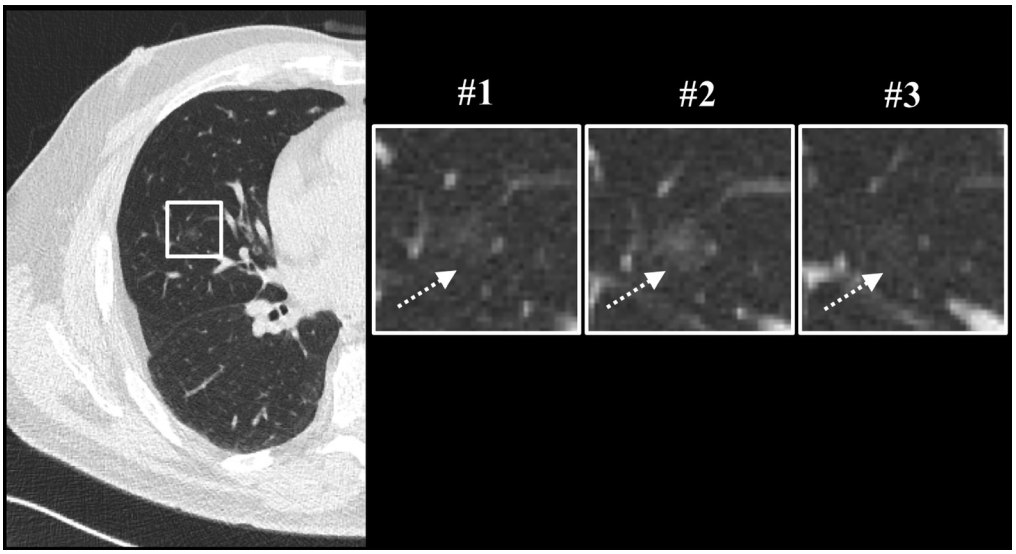

Fig. 3 An example of testing trial for deep-learning-based model observer. The square region-of-interest indicates the location of the zoom insets. The zoom insets (#1 to #3) illustrate the central three consecutive slices across a 7.4-mm ground glass nodule. The dashed arrows indicate the location of nodule. Of note, the arrows were placed slightly off the center of lesion, for illustration purpose.

consecutive images that were randomly selected from the middle lobe/lingual. Note that the random selection of nodule-absent images was repeated across all validation subgroups. The further discussion about the simplification of testing procedure is presented in Sec. 4.

Data augmentation strategies were applied to the training subgroups, to improve the performance of DL-MO. We selected similar data augmentation as used in Ref. 23. These strategies included image conversion, cropping, z-direction interpolation, and random rotation. In image conversion, a lung display window (W/L: $1600/-600$ HU) was applied to original CT images and then transformed to the grayscale range of [0, 255], to match the image dynamic range used in ResNet-50. As for cropping, additional multisized VOIs were extracted for training the DL-MO. The size of VOIs ranged from $10.4 \times 10.4 \times 3.0$ mm^3 to $19.3 \times 19.3 \times 3.0$ mm^3, i.e., $14 \times 14 \times 3$ voxels to $26 \times 26 \times 3$ voxels. The z-direction interpolation and random rotation were both used to generate more VOIs for training DL-MO. For each experimental condition, data augmentation generated approximately additional 150,000 training samples. One could refer to Ref. 23 for the further details of data augmentation strategies.

2.4 *Human Observer Study*

Four human readers (two board-certified radiologists and two radiology fellows) were recruited to perform a signal-known-exactly localization task. The two board-certified radiologists had more than 10 and 20 years' experience, respectively. All human observers were subspecialized in thoracic CT. A MATLAB (Mathwork, Inc.) based graphical user interface was developed to display each experimental trial and record human reader response (Fig. 4). The target lung nodule image with a clear background was displayed with the corresponding experimental trials. Each reader participated in no <4 reading sessions where the 12 conditions were randomly split across all sessions. The consecutive sessions were separated by >3 days to reduce the potential recall effects. In each session, the experimental conditions were viewed in the order of difficulty levels, i.e., the conditions with the lowest radiation dose level and/or smallest nodule size were viewed earlier than the others. For each condition, all experimental trials were displayed in a randomized order. Human readers were tasked to identify the most-likely nodule location.

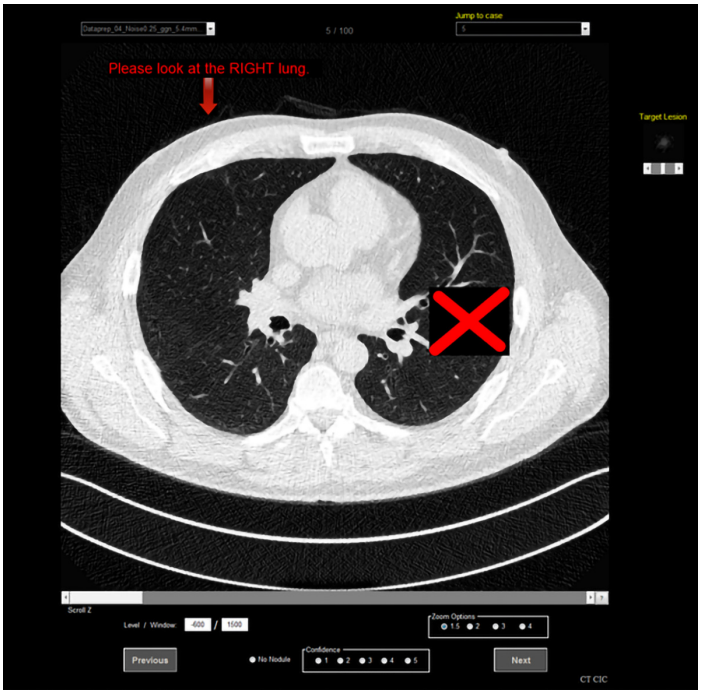

Fig. 4 The graphical user interface developed for human observer study. In each experimental trial, only one side of chest was viewed by readers, whereas the other side was blocked using a cross marker.

They were also requested to provide a six-scale detection confidence score (0: definitely no nodule; 1: the case is very likely nodule-absent, but it would be the most suspicious location if it is actually nodule-present; 2: the case is likely nodule-absent, but it would be the most suspicious location if there is a nodule; 3: the case is likely nodule-present, and this is the most suspicious location of nodule; 4: the case is very likely nodule-present, and this is the most suspicious location of nodule; and 5: the case is definitely present).

2.5 *Figure of Merit and Statistical Analyses*

Area under localization receiver operating characteristic curve (A_L) was used as the figure-of-merit for the performance of both DL-MO and radiologist readers. The LROC plots the true positive localization fraction (TPLF) as a function of false positive fraction, where TPLF is the joint proportion of true positive decision and correct localization.[32] The correct localization was defined as when the distance between the reference location and the reader location <3.7 mm (as was used in our preliminary reader study[29]), i.e., radius of the largest lung nodule used in this study. The mean A_L of all human readers was calculated for each experimental condition and used as the overall performance metric. To determine the variance of A_L, because all readers were reading the same 100 cases, correlation exists. A nonparametric approach for multireader multicase analysis of A_L was used to calculate the variance, taking into account the correlation.[33,34] Pearson's product moment correlation coefficient (denoted as Pearson's ρ) was used to gauge the strength of the correlation between DL-MO and human. Finally, Bland–Altman plot was used to quantify the degree of discrepancy between DL-MO performance and human reader performance across all experimental conditions.

3 Results

3.1 *Summary of Radiologist Performance*

We observed obvious inter-reader variability across all experimental conditions (Fig. 5), although Wilcoxon signed rank test indicated statistically insignificant difference (P value = 0.204 at 5% significance level) between the mean A_L of senior radiologists (i.e., readers #1 and #2) and that of fellows (readers #3 and #4). The strength of variability appeared to decrease as the difficulty level of experimental condition was reduced. The standard deviation of A_L per condition ranged from 3.1% to 11.8%, and the range of A_L (i.e., the difference between maximal and minimal values) per condition was 6.6% and 25.9%.

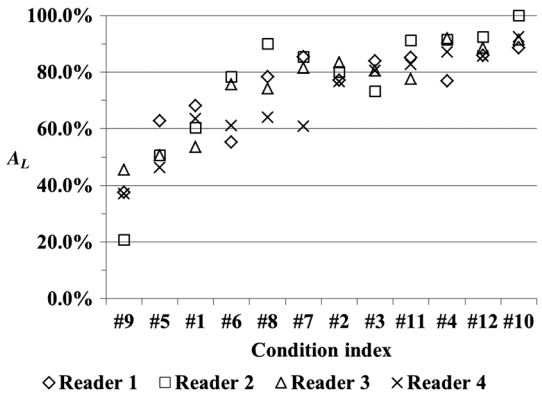

Fig. 5 The performance of each radiologist reader across 12 experimental conditions. A_L is the area under the localization receiver operating characteristic curve. Each value of A_L was calculated across all available trials ($n = 100$) per condition, which did not involve the repeated twofold cross validation. These conditions were sorted according to the ascending order of mean A_L.

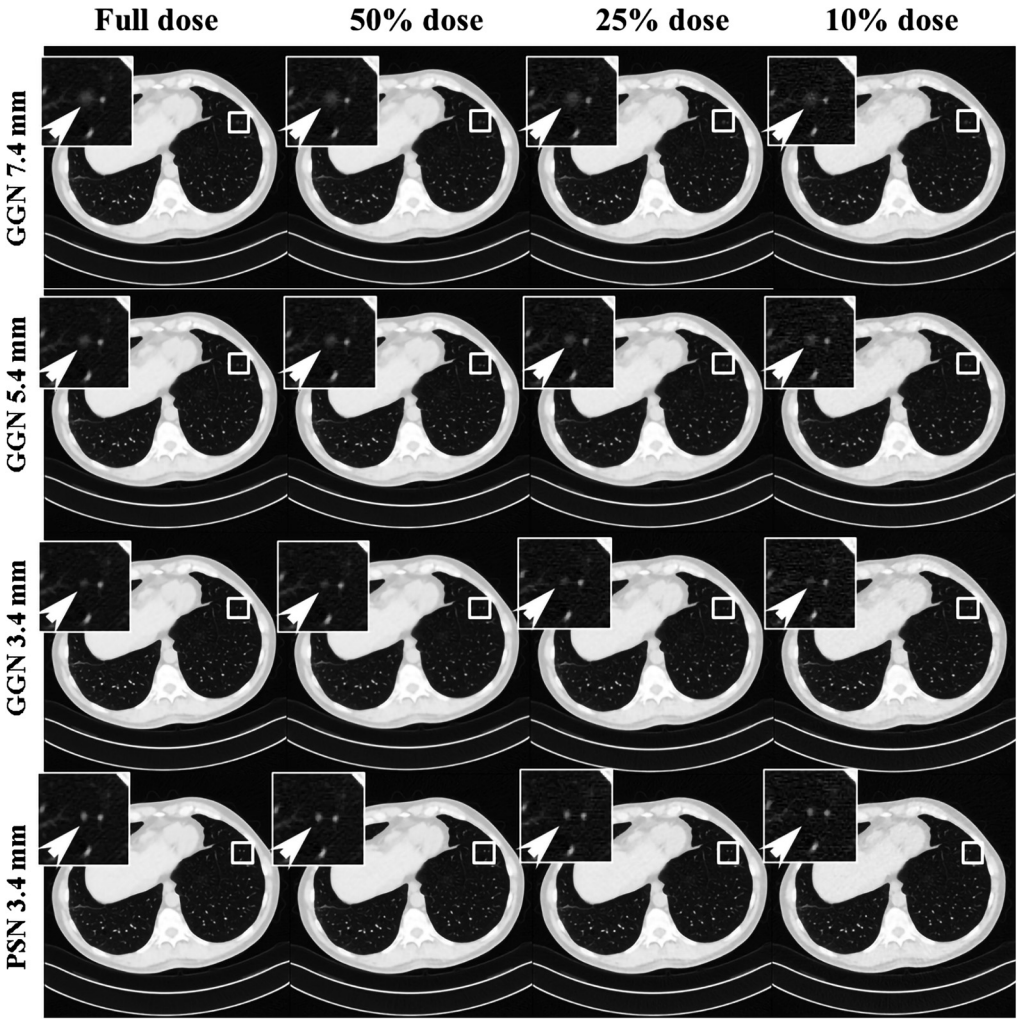

Fig. 6 Examples of synthesized patient images at different conditions with varying nodule types, nodule sizes, and radiation dose levels. The arrows indicate the nodule location. GGN, ground glass nodule and PSN, partially solid nodule.

3.2 *Virtual CT Exams and Heat Maps*

Figure 6 illustrates some examples of synthesized patient images at different experimental conditions with varying nodule types/sizes and radiation dose levels. Examples of heat maps from validation subgroups are illustrated in Fig. 7. These heat maps were regrouped to four scenarios for convenience of comparison. It can be seen that the voxel-wise response of DL-MO was generally stronger at the location of lung nodules than that of nodule-absent regions. However, the relative strength between the DL-MO response at nodule-present and nodule-absent regions tended to decrease with lower radiation dose level and smaller nodule size, which suggested a higher likelihood of detection error.

3.3 *Internal Noise Calibration*

The calibration of the weighting factor α of internal noise component (Sec. 2.1) is illustrated in Fig. 8. The DL-MO performance was calibrated to match the averaged radiologist reader performance ($A_L = 79.4\%$) at the experimental condition #2 (i.e., 25% RD, 5.4 mm GGN, with IR-2). After the calibration, the value of α was determined to be 1.8, i.e., the internal noise was 1.8 times of the noise of $\lambda_{0,\,\mathrm{bkg}}$. The same value of α was used for all the other experimental conditions with varying radiation dose levels, nodule sizes/types, and image reconstruction types.

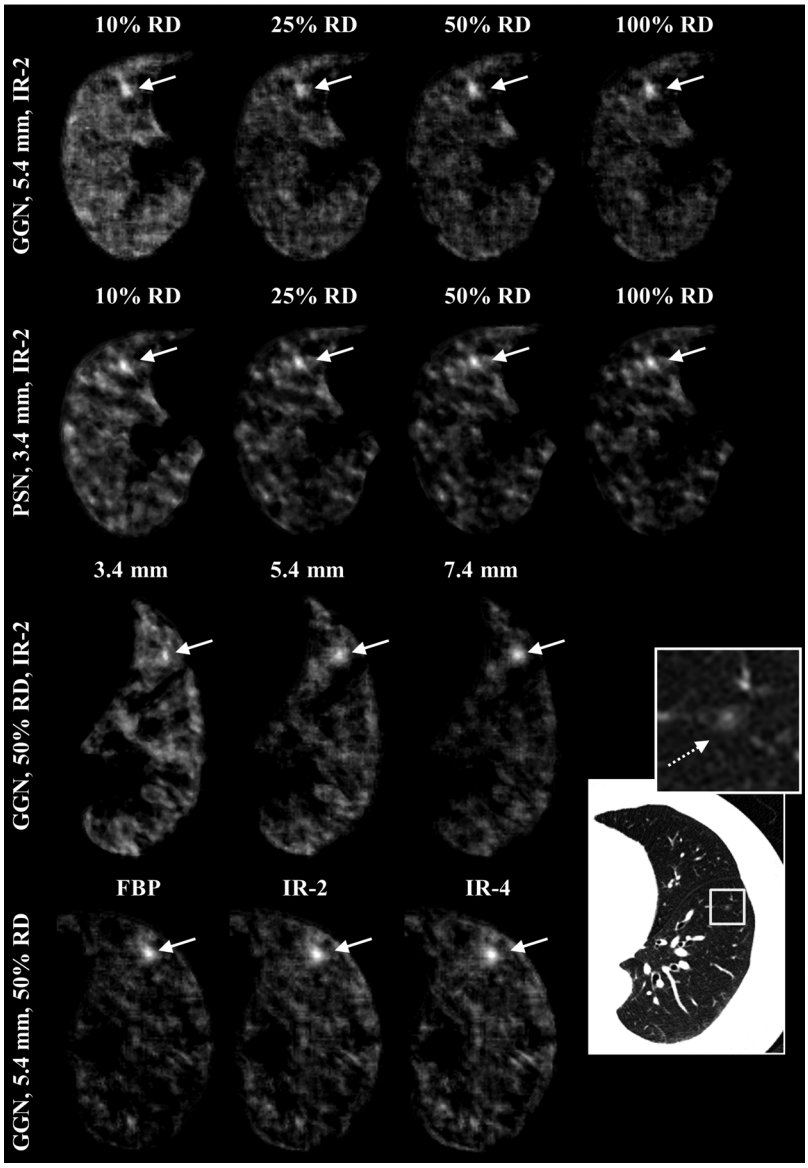

Fig. 7 Examples of heat maps from experimental trials across all experimental conditions at different nodule size, nodule types, radiation dose levels, and image reconstruction types. The heat maps were reorganized into four rows for convenience of comparison. The first and second rows: radiation dose levels were changed from 10% to 100% RD. The third row: the nodule size was changed from 3.4 to 7.4 mm. The fourth row: the image reconstruction types were changed across FBP, IR-2, and IR-4. The solid arrows indicate the location of lung nodules in the heat maps. The display window of heat map is W/L: 2/0. The inset image at the bottom right illustrates an example of experimental trials, and the dashed arrow indicates a GGN in the CT image. GGN, ground glass nodule; PSN, partially solid nodule; RD, routine dose level; FBP, filtered back projection; IR-2, iterative reconstruction with the strength setting of 2; and IR-4, iterative reconstruction with the strength setting of 4.

3.4 *Statistical Analyses*

For convenience of comparison, the experimental conditions were regrouped into four scenarios (Fig. 9). The A_L of the calibrated DL-MO was comparable to the averaged A_L of radiologist readers across all experimental conditions. A statistically significant correlation was observed between DL-MO and human observers. The value of Pearson's ρ was 0.980 with 95% confidence interval (CI) [0.924, 0.994]. Further, Bland–Altman plot indicated that there was no statistically significant discrepancy between DL-MO and radiologist readers (Fig. 10).

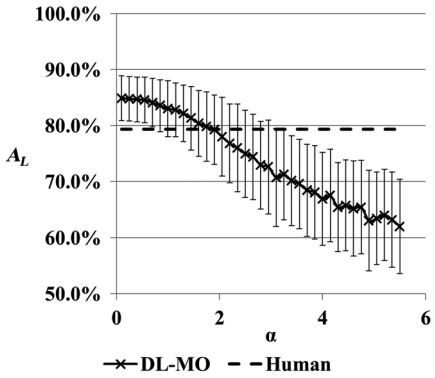

Fig. 8 The weighting factor α was adjusted to calibrate DL-MO performance to the averaged radiologist reader performance at the experimental condition #2, that is, lesion size was 5.4 mm, lesion type was ground glass nodule, radiation dose was 25% of routine dose level, and CT images were reconstructed by iterative algorithm. The value of α was varied from 0.1 to 5.5 with a uniform interval of 0.05. Every third sample was plotted out for the convenience of illustration. The error bars indicate the standard deviation of DL-MO performance. A_L is the area under the localization receiver operating characteristic curve and DL-MO, deep learning-based model observer.

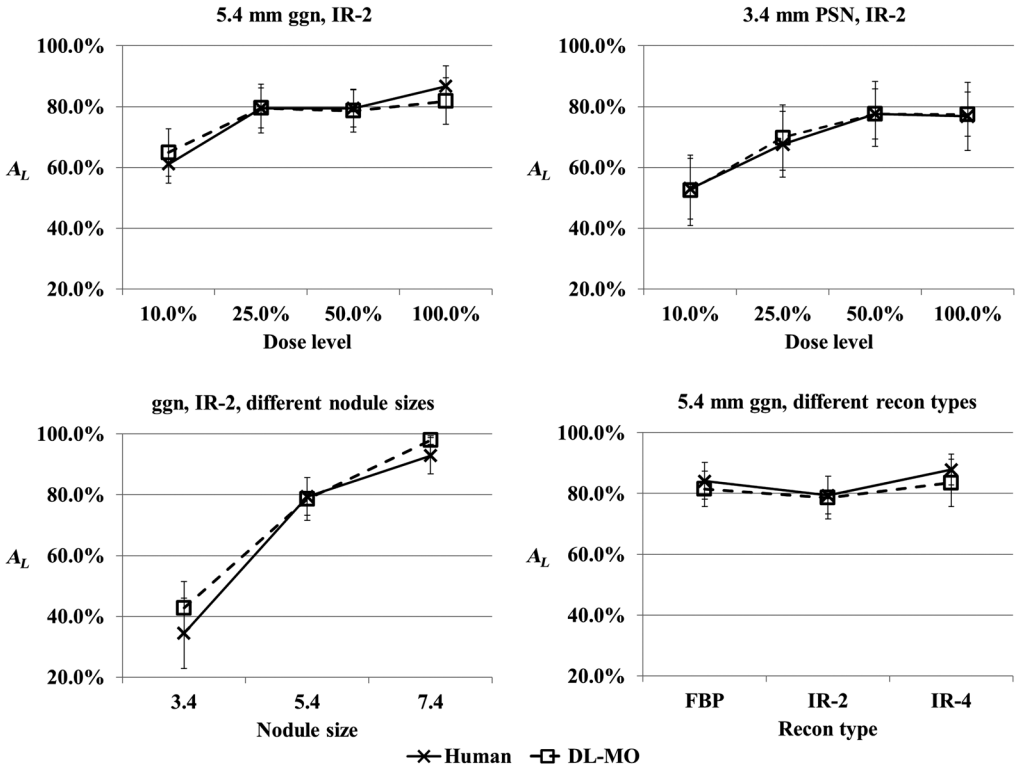

Fig. 9 Performance comparison between DL-MO and radiologist readers across all experimental conditions with varying lung nodule sizes (3.4 to 7.4 mm), lung nodule types (ground glass and partially solid), radiation dose levels (10% to 100% of routine dose), and image reconstruction types (FBP and IR). For convenience of illustration, these experimental conditions were regrouped to four charts. (a) Conditions #1 to #4, (b) conditions #5 to #8, (c) conditions #9, #3, and # 10, and (d) conditions #11, #3, and #12. The radiologist reader performance was averaged across all readers. The error bars indicate the standard deviation of the performance. DL-MO, deep-learning-based model observer; A_L, the area under the localization receiver operating characteristic curve; GGN, ground glass nodule, PSN, partially solid nodule; FBP, filtered back projection; and IR, iterative reconstruction.

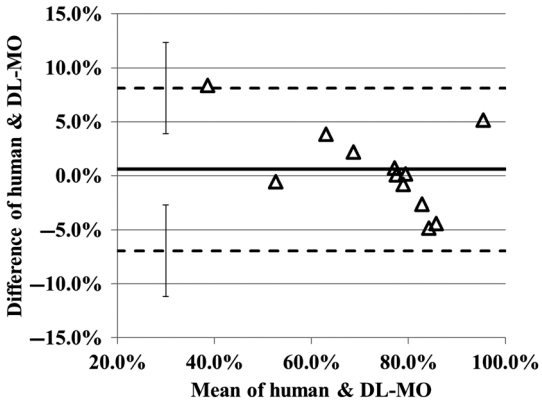

Fig. 10 Bland–Altman plots of the performance difference between DL-MO and radiologist readers across all experimental conditions. Human performance was averaged across all radiologist readers. The solid line denotes the bias. The two dashed lines denote the upper and the lower LOA, respectively. The error bars indicate the 95% CI of LOA. DL-MO, deep learning-based model observer and LOA, limit of agreement.

Table 2 Strength of correlation between the calibrated DL-MO and individual radiologist reader

Reader	#1	#2	#3	#4
Pearson's ρ	0.877	0.943	0.925	0.931
95% CI	[0.611, 0.965]	[0.806, 0.984]	[0.748, 0.979]	[0.766, 0.981]

The mean bias was 0.57%. The upper and lower limits of agreement were −6.83%, (with 95% CI [−11.0%, −2.68%]) and 7.96% (with 95% CI [3.81%, 12.1%]). In addition, the calibrated DL-MO performance also yielded statistically significant correlation to individual radiologist performance, despite the obvious inter-reader variability (Table 2). The corresponding Pearson's ρ ranged from 0.877 to 0.943. The corresponding 95% CI was expanded due to the large variance of human reader performance.

4 Discussion

In this study, we developed a DL-MO framework for lesion localization tasks by incorporating the sliding window strategy and nodule searching process into our previous deep-learning-based method. We validated this framework in a realistic lung nodule detection task in patient CT lung cancer screening exams by comparing the performance of DL-MO with that of radiologist readers on multiple experimental conditions at varying lung nodule sizes, lung nodule types, radiation dose levels, and image reconstruction types. Strong correlation and agreement between the proposed DL-MO and human readers were demonstrated.

The selection of CNN layer and PLS components was empirically determined to be the same as used in our prior work[22,23] that involved a liver metastases detection task in patient images. It is possible to fine-tune the selection of CNN layer and PLS components, to further enhance the efficiency of DL-MO for the lung nodule localization task or other types of clinically relevant diagnostic tasks. As was pointed out in Ref. 23, it is very challenging to theoretically predict the optimal configuration of CNN layer and PLS components for the proposed DL-MO, due to the extreme complexity for estimating the CNN feature transferability. Nevertheless, one could still employ the straightforward grid searching strategy as used in Ref. 23, to determine a reasonable configuration of CNN layer and PLS components.

The R2f-CV method was selected to evaluate the generalizability of DL-MO. Compared to the standard 2f-CV method, R2f-CV can reduce the variance of the estimated generalization performance. However, it still tends to yield a downward bias (i.e., the underestimation) on the

generalizability of a given machine learning model (e.g., the proposed DL-MO), especially on the small-sized datasets.[35] Other advanced cross-validation strategies, such as the bootstrap cross-validation method,[36] could be employed to provide more accurate estimation of the generalization accuracy of the proposed DL-MO, whereas the computational cost would be much higher than the R2f-CV method. Further study is warranted to investigate different cross-validation strategies, which is beyond the scope of current study.

Computational efficiency of the proposed DL-MO was limited, mainly due to the use of sliding window strategy and the absence of GPU-based acceleration techniques. For each experimental condition, the average training time was ~60 min. For each testing trial, the average time for calculating the DL-MO response per CT image was ~7.5 min, i.e., DL-MO would spend ~225 min per experimental trial if the calculation is carried over all 30 images of the middle lobe/lingula. Therefore, we were motivated to carry out the simplification of the testing procedure (Sec. 2.3) in DL-MO studies, to avoid the excessive computational cost. Such simplification may not significantly degrade the DL-MO performance over signal-present images since the most relevant nodule information was already included in the images that coincided with the nodule centroid. However, we acknowledge that this simplification may underestimate the potential false positives (i.e., overestimating the accuracy) over signal-absent images as the model has a limited number of slices to test on, and then the weight of internal noise component may need to be recalibrated for the testing over the entire image volume. Since the signal-absent images per case per condition were randomly selected for DL-MO testing (Sec. 2.3), a potential way to reduce such variance is to repeat the cross validation for more times so that DL-MO performance would be evaluated against more diverse anatomical background. Further, as was aforementioned, the R2f-CV tends to underestimate the generalization performance especially on the small-sized dataset.[35] This phenomenon could at least partially cancel out the effects of the potentially underestimated false positives. Despite this simplification, the agreement between the DL-MO and human observer performance suggested that the potential under-estimate of the false positives in signal-absent images could be statistically insignificant. The current method could be readily extended to enable nodule searching in a larger 3-D image volume: the sliding window can be moved across the entire volume to generate 3-D heat map; then the most-likely nodule location can be labeled as the voxel that coincided with the maximal λ_0; the remaining steps would be the same as described in Sec. 2. To improve computational efficiency, the sliding window strategy may be converted to a convolutional operation, without significantly compromising DL-MO performance. The computation can be further accelerated by GPU-based techniques. With proper acceleration techniques, it would be more practical to have the DL-MO to search the nodule throughout the entire image volume of the middle lobe/lingula, which could potentially further improve the strength of correlation and agreement between DL-MO and radiologist readers. These aspects shall be investigated in our follow-up study.

The proposed DL-MO was calibrated to match the averaged human observer performance at a given condition, by adjusting the weight of internal noise component instead of directly using human data as the ground truth of DL-MO training. The calibrated DL-MO also yielded strong correlation to individual human observer involved in calibration. We estimate that the calibrated DL-MO may maintain positive correlation to a new human observer who is not involved in calibration, when there is no statistically significant performance difference between new observer and the prior ones involved in calibration.

Finally, we acknowledge several limitations in the presented study. First, we only used two original lung nodules for mimicking different nodule sizes and contrast levels, whereas radiologists are frequently challenged to identify a large number of different types of lesions with varying radiological features. The reason why we chose only two original lung nodules was because we would like to evaluate the impact of nodule size and contrast levels on the reader performance, with other factors fixed. Second, the size of patient cohort and the number of radiologist readers may be limited, and the large inter-reader variability may result in non-negligible standard error to the averaged human reader performance. Third, although the current validation study is considered relatively comprehensive since it included four radiologist readers (both experienced radiologists and trainees) and both FBP and IR reconstruction methods, studies involving more human readers from multiple institutions and images obtained from other CT scanners may be needed to fully evaluate its generalizability.

5 Conclusion

A DL-MO framework was developed to evaluate diagnostic image quality in lesion localization tasks. This framework was validated in a realistic lung nodule detection task in CT by comparing the performance predicted by the DL-MO and that of radiologist readers. It was shown that the DL-MO performance was highly correlated with human observer performance in the lung nodule detection task that involved patient chest background, realistic lung nodule images, and complex visual searching process. The presented study demonstrated strong potential of using the proposed DL-MO for task-based image quality assessment and radiation dose optimization in realistic CT tasks.

Disclosures

Dr. Cynthia H McCollough is the recipient of a research grant from Siemens Healthcare. The other authors have no relevant conflicts of interest to disclose.

Acknowledgments

This study was sponsored by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award Nos. R01 EB017095 and U01 EB017185. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Y. Nagatani et al., "Lung nodule detection performance in five observers on computed tomography (CT) with adaptive iterative dose reduction using three-dimensional processing (AIDR 3D) in a Japanese multicenter study: comparison between ultra-low-dose CT and low-dose CT by receiver-operating characteristic analysis," *Eur. J. Radiol.* **84**(7), 1401–1412 (2015).
2. F. R. Verdun et al., "Image quality in CT: from physical measurements to model observers," *Physica Med.* **31**(8), 823–843 (2015).
3. C. H. McCollough et al., "Degradation of CT low-contrast spatial resolution due to the use of iterative reconstruction and reduced dose levels," *Radiology* **276**(2), 499–506 (2015).
4. J. M. Kofler et al., "Assessment of low-contrast resolution for the American College of Radiology Computed Tomographic Accreditation Program: what is the impact of iterative reconstruction?" *J. Comput. Assisted Tomogr.* **39**(4), 619–623 (2015).
5. O. Christianson et al., "An improved index of image quality for task-based performance of CT iterative reconstruction across three commercial implementations," *Radiology* **275**(3), 725–734 (2015).
6. L. Yu et al., "Correlation between a 2D channelized Hotelling observer and human observers in a low-contrast detection task with multi-slice reading in CT," *Med. Phys.* **44**(8), 3990–3999 (2017).
7. L. Yu et al., "Prediction of human observer performance in a 2-alternative forced choice low-contrast detection task using channelized Hotelling observer: impact of radiation dose and reconstruction algorithms," *Med. Phys.* **40**(4), 041908 (2013).
8. S. Leng et al., "Correlation between model observer and human observer performance in CT imaging when lesion location is uncertain," *Med. Phys.* **40**(8), 081908 (2013).
9. H. C. Gifford, "A visual-search model observer for multislice-multiview SPECT images," *Med. Phys.* **40**(9), 092505 (2013).
10. H. C. Gifford, "Efficient visual-search model observers for PET," *Br. J. Radiol.* **87**(1039), 20140017–20140017 (2014).
11. M. A. Lago, C. K. Abbey, and M. P. Eckstein, "Foveated model observers to predict human performance in 3D images," *Proc. SPIE* **10136**, 101360P (2017).

12. M. Lago et al., "Foveated model observer to predict human search performance on virtual digital breast tomosynthesis phantoms," *Proc. SPIE* **11316**, 113160V (2020).
13. E. Samei, M. J. Flynn, and W. R. Eyler, "Detection of subtle lung nodules: relative influence of quantum and anatomic noise on chest radiographs," *Radiology* **213**(3), 727–734 (1999).
14. F. O. Bochud et al., "Importance of anatomical noise in mammography," *Proc. SPIE* **3036**, 74–80 (1997).
15. H. L. Kundel et al., "Nodule detection with and without a chest image," *Invest. Radiol.* **20**(1), 94–99 (1985).
16. P. F. Judy et al., "Contrast-detail curves for liver CT," *Med. Phys.* **19**(5), 1167–1174 (1992).
17. M. Alnowami et al., "A deep learning model observer for use in alterative forced choice virtual clinical trials," *Proc. SPIE* **10577**, 105770Q (2018).
18. W. Murphy et al., "Using transfer learning for a deep learning model observer," *Proc. SPIE* **10952**, 109520E (2019).
19. F. Massanes and J. G. Brankov, "Evaluation of CNN as anthropomorphic model observer," *Proc. SPIE* **10136**, 101360Q (2017).
20. W. Zhou, H. Li, and M. A. Anastasio, "Approximating the ideal observer and Hotelling observer for binary signal detection tasks by use of supervised learning methods," *IEEE Trans. Med. Imaging* **38**(10), 2456–2468 (2019).
21. F. K. Kopp et al., "CNN as model observer in a liver lesion detection task for x-ray computed tomography: a phantom study," *Med. Phys.* **45**(10), 4439–4447 (2018).
22. H. Gong et al., "Best in physics (imaging): a deep learning based model observer for low-contrast object detection task in x-ray computed tomography," in *American Association of Medical Physics*, Vol. **45**(6), Wiley, New Jersey (2018).
23. H. Gong et al., "A deep learning- and partial least square regression-based model observer for a low-contrast lesion detection task in CT," *Med. Phys.* **46**(5), 2052–2063 (2019).
24. B. Chen et al., "Lesion insertion in projection domain for computed tomography image quality assessment," *Proc. SPIE* **9412**, 94121R (2015).
25. B. Chen et al., "Lesion insertion in the projection domain: methods and initial results," *Med. Phys.* **42**(12), 7034–7042 (2015).
26. C. Ma et al., "Evaluation of a projection-domain lung nodule insertion technique in thoracic CT," *Proc. SPIE* **9783**, 97835Y (2016).
27. L. Yu et al., "Development and validation of a practical lower-dose-simulation tool for optimizing computed tomography scan protocols," *J. Comput. Assisted Tomogr.* **36**(4), 477–487 (2012).
28. H. Gong et al., "Correlation between a deep-learning-based model observer and human observer for a realistic lung nodule localization task in chest CT," *Proc. SPIE* **10952**, 109520K (2019).
29. L. Yu et al., "A virtual clinical trial using projection-based nodule insertion to determine radiologist reader performance in lung cancer screening CT," *Proc. SPIE* **10132**, 101321R (2017).
30. K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 770–778 (2016).
31. O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vision* **115**, 211–252 (2015).
32. R. G. Swensson, "Unified measurement of observer performance in detecting and localizing target objects on images," *Med. Phys.* **23**(10), 1709–1725 (1996).
33. L. M. Popescu, "Nonparametric ROC and LROC analysis," *Med. Phys.* **34**(5), 1556–1564 (2007).
34. A. Wunderlich and F. Noo, "A nonparametric procedure for comparing the areas under correlated LROC curves," *IEEE Trans. Med. Imaging* **31**(11), 2050–2061 (2012).
35. R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. Artifi. Intell.*, Morgan Kaufmann Publishers Inc., Montreal, Quebec, Canada, Vol. **2**, pp. 1137–1143 (1995).

36. W. J. Fu, R. J. Carroll, and S. Wang, "Estimating misclassification error with small samples via bootstrap cross-validation," *Bioinformatics* **21**(9), 1979–1986 (2005).

Hao Gong received his MS degree from electrical and computer engineering at Purdue University, and his PhD from a joint doctoral program at School of Biomedical Engineering and Sciences between Virginia Tech and Wake Forest University. His research interests include but not limited to bio-medical image processing, computer vision, machine learning, and x-ray computed tomography image reconstruction/artifact correction.

Qiyuan Hu is a PhD student in medical physics at the University of Chicago. She received her BA degree in physics and mathematics from Carleton College in 2017. Her research interests include radiomics and deep learning methodologies for computer-aided diagnosis. She is a student member of SPIE and an officer of the University of Chicago SPIE Student Chapter.

Andrew Walther is a senior undergraduate at Creighton University majoring in mathematics and biomedical physics. His contributions to the CT Clinical Innovation Center at Mayo Clinic occurred while he was an undergraduate research fellow during the summer of 2017. Following graduation, Andrew plans to pursue graduate studies in biostatistics.

Chi Wan Koo is an associate professor of radiology at Mayo Clinic. She earned her undergraduate and medical degrees from New York University (NYU). She completed a fellowship in cardiothoracic imaging at the Hospital of University of Pennsylvania. Her current research interests include radiation dose reduction as well as quantitative and artificial intelligence assisted imaging of thoracic diseases.

Edwin A. Takahashi is an assistant professor of radiology at Mayo Clinic. He received his medical degree at the University of Hawaii in 2013. He completed his residency in diagnostic radiology in 2018 and fellowship in vascular and interventional radiology in 2019, both at Mayo Clinic. His research interests focus on the utilization of medical imaging to improve endovascular procedure planning and outcomes.

David L. Levin is a professor of radiology at the Mayo Clinic, specializing in thoracic imaging. He has a long-standing interest in the development of novel imaging techniques, especially for use in the noninvasive assessment of pulmonary physiology and pathophysiology.

Tucker F. Johnson is an instructor in radiology at the Mayo Clinic, specializing in thoracic radiology who has a longstanding interest in radiology informatics, especially for use in the detection and analysis of lung cancer and interstitial lung disease.

Megan J. Hora is a diagnostic radiologist for Avera Medical Center in Sioux Falls, South Dakota. She received her BS degree in biology in 2007 and her MD in 2011, both from the University of South Dakota. She completed her diagnostic radiology residency at Creighton University in 2016 and continued on to complete a cardiothoracic imaging fellowship at Mayo Clinic in 2017. Her main interests include interstitial lung disease and oncologic imaging, especially with PETCT.

Shuai Leng received his BS and MS degrees in engineering physics from Tsinghua University in 2001 and 2003, respectively, and his PhD in medical physics in 2008 from the University of Wisconsin Madison. He is an associate professor of medical physics at Mayo Clinic, Rochester, Minnesota, USA. He has authored over 100 peer-reviewed articles. His research interest is in technical development and clinical application of x-ray and CT imaging.

Joel G. Fletcher is an abdominal radiologist at Mayo Clinic. He established the CT Clinical Innovation Center at Mayo Clinic with Dr. Cynthia McCollough. His research interests include image acquisition/reconstruction, radiation dose reduction and observer performance studies. He is known for interdisciplinary collaborations in GI imaging, particularly the development and validation of radiologic methods for diagnosis and staging of Crohn's disease, and the use of these tests to guide clinical management.

Cynthia H. McCollough received her doctorate degree from the University of Wisconsin in 1991. She is a professor of radiological physics and biomedical engineering at Mayo Clinic, where she directs the CT Clinical Innovation Center. Her research interests include CT dosimetry, advanced CT technology, and new clinical applications, such as dual-energy and multispectral CT. She is an NIH-funded investigator and is active in numerous professional organizations. She is a fellow of the AAPM and ACR.

Lifeng Yu is a professor of medical physics at Mayo Clinic and a fellow of the AAPM. He received his BS degree in nuclear physics in 1997 and an MEng. degree in nuclear technology in 2000, both from Beijing University, and his PhD in medical physics from the University of Chicago in 2006. His research interests include CT physics, image quality assessment, radiation dose reduction, and spectral CT.