DeepAMO: a multi-slice, multi-view anthropomorphic model observer for visual detection tasks performed on volume images

Ye Li,[a,b,*] Junyu Chen,[a,b] Justin L. Brown,[c] S. Ted Treves,[d,e] Xinhua Cao,[e,f] Frederic H. Fahey,[e,f] George Sgouros,[b] Wesley E. Bolch,[c] and Eric C. Frey[a,b]

[a]Johns Hopkins University, Whiting School of Engineering, Department of Electrical and Computer Engineering, Baltimore, Maryland, United States
[b]Johns Hopkins University, School of Medicine, Russell H. Morgan Department of Radiology and Radiological Science, Baltimore, Maryland, United States
[c]University of Florida, J. Crayton Pruitt Family Department of Biomedical Engineering, Gainesville, Florida, United States
[d]Brigham and Women's Hospital, Department of Radiology, Boston, Massachusetts, United States
[e]Harvard Medical School, Department of Radiology, Boston, Massachusetts, United States
[f]Boston Children's Hospital, Department of Radiology, Boston, Massachusetts, United States

Abstract

Purpose: We propose a deep learning-based anthropomorphic model observer (DeepAMO) for image quality evaluation of multi-orientation, multi-slice image sets with respect to a clinically realistic 3D defect detection task.

Approach: The DeepAMO is developed based on a hypothetical model of the decision process of a human reader performing a detection task using a 3D volume. The DeepAMO is comprised of three sequential stages: defect segmentation, defect confirmation (DC), and rating value inference. The input to the DeepAMO is a composite image, typical of that used to view 3D volumes in clinical practice. The output is a rating value designed to reproduce a human observer's defect detection performance. In stages 2 and 3, we propose: (1) a projection-based DC block that confirms defect presence in two 2D orthogonal orientations and (2) a calibration method that "learns" the mapping from the features of stage 2 to the distribution of observer ratings from the human observer rating data (thus modeling inter- or intraobserver variability) using a mixture density network. We implemented and evaluated the DeepAMO in the context of $^{99\mathrm{m}}$Tc-DMSA SPECT imaging. A human observer study was conducted, with two medical imaging physics graduate students serving as observers. A 5×2-fold cross-validation experiment was conducted to test the statistical equivalence in defect detection performance between the DeepAMO and the human observer. We also compared the performance of the DeepAMO to an unoptimized implementation of a scanning linear discriminant observer (SLDO).

Results: The results show that the DeepAMO's and human observer's performances on unseen images were statistically equivalent with a margin of difference (ΔAUC) of 0.0426 at $p < 0.05$, using 288 training images. A limited implementation of an SLDO had a substantially higher AUC (0.99) compared to the DeepAMO and human observer.

Conclusion: The results show that the DeepAMO has the potential to reproduce the absolute performance, and not just the relative ranking of human observers on a clinically realistic defect detection task, and that building conceptual components of the human reading process into deep learning-based models can allow training of these models in settings where limited training images are available.

© *2021 Society of Photo-Optical Instrumentation Engineers (SPIE)* [DOI: 10.1117/1.JMI.8.4.041204]

Keywords: model observer; deep learning; task-based image quality assessment.

[*]Address all correspondence to Ye Li, bettergary@gmail.com

Paper 20150SSRR received Jun. 2, 2020; accepted for publication Dec. 31, 2020; published online Jan. 28, 2021.

1 Introduction

Often, the quality of a medical image is measured in terms of the physical properties of the image, such as image contrast, spatial resolution, and noise level.[1] Fidelity-based measures, such as root mean squared error, peak signal-to-noise ratio, and structural similarity index, have also been widely used in the medical imaging community. These measures are appealing because they are relatively easy to compute, have straightforward physical interpretations, and can provide objective quantitative measures of image quality. However, they are not directly related to the diagnostic task that is performed with the images and thus may not be clinically relevant. Clinically relevant image quality assessment should be with respect to the task that is to be performed.[2–8] Ideally, the observers would be drawn from the population of people performing the task, i.e., for medical images, a radiologist or nuclear medicine physician. However, in practice, especially in large-scale developmental research studies, the use of human observers (and especially physicians) is too time-consuming, inconvenient, and expensive. Thus a great deal of work has gone into the development of anthropomorphic model observers that predict human observer performance.[9–12]

Task-based measures of image quality based on model observers have been advocated by several investigators over the years, starting from Harris,[13] and including Hanson and Myers,[14] Wager et al.,[15] Judy et al.,[16] and Myers et al.[9,17] However, existing model observers are often not directly applicable to diagnostic tasks of clinical relevance.[18] For example, as described below, commonly used model observers are strictly valid only for signal-location-known (exactly and statistically) tasks. In addition, although these observers predict rankings of human observer performance, they often require the use of concepts such as internal noise to match the absolute performance of human observers.

Of the existing anthropomorphic observer models, the channelized Hotelling observer (CHO) has been the most widely used as a substitute for human observers in signal-location-known tasks in nuclear medicine imaging research.[19] The CHO has been shown to correlate well with human observer performance on signal-known-exactly/background-known-exactly (SKE/BKE) tasks,[9,20] SKE/background-known-statistically (BKS) (e.g., lumpy backgrounds) tasks,[21] and SKE-realistic anatomical backgrounds tasks.[22–24] However, in those tasks, the observer is asked to decide whether the defect is present or not at a specified location. A more clinically relevant detection task is the signal-known-statistically (SKS)/BKS task, where variability can be present in both the signal and background. Here signal variability refers to differences in signal (defect) shape, size, orientation, or texture. Background variability can come from two sources: quantum noise and anatomical variability. Modeling the latter is important in order to model clinical tasks where patients can vary substantially in size, shape, and uptake pattern. It is important to model these image features, especially in studies such as virtual clinical trials, in order to accurately model performance on images from patient populations. For these clinically more realistic SKE/BKS and SKS/BKS tasks, there is evidence that rankings or ranking trends of human observers and the CHO are correlated for different noise levels,[24,25] reconstruction methods and phantom populations,[26] imaging systems,[27] compensation methods, and postfilter cutoff frequencies.[22] Scanning forms of the CHO can be applied for the clinically more realistic SKS/BKS tasks to analyze each location within a particular region of interest as a potential defect site.[18] However, images and defects investigated in this work did not have closed form expressions for the linear discriminant and ensemble methods were thus used in estimating the observer. This necessitates the use of a relatively large number of images. Since the defects and background were not spatially invariant, the linear discriminant would have to be evaluated at each potential defect location. Thus for the case of 3D images and a large number of potential defect sites, scanning observers based on the linear discriminant and CHO can be quite

2329-4302/2021/$28.00 © 2021 SPIE

computationally intensive. The use of channels reduces the computational demands, but they do remain significant.[28] Partly for these reasons, previous attempts at using scanning observers on multi-orientation, multi-slice images have focused on reducing the search region, i.e., use a front-end search process[28] to obtain a subset of the original search location set, and simplifying the defect confirmation (DC) process by simulating a simpler SKE/BKE detection task, etc.[29,30]

In addition to the above limitations, existing model observers often predict rankings but not the absolute performance of human observers.[28–31] For imaging system optimization or comparison studies, this can be sufficient, but for other applications, such as selecting imaging time, administered activity (AA), or radiation dose, prediction of absolute performance measures is required.[8] Obtaining absolute agreement for these model observers typically is done with the addition of observer internal noise.[28] The calibration process is a parameter search exercise where the goal is to find the value of an internal noise parameter that matches performance between the model and human observers. Note that the calibration process is often performed for one specific combination of signal (shape, size, and orientation) and noise level, and it is unclear the degree to which the calibration generalizes to other situations.

Another gap between current anthropomorphic observers and the real clinical task is that current model observers have been primarily designed for analyzing 2D images. By contrast, many clinical tasks require the interpretation of 3D datasets. This often involves reviewing sequences of 2D slices in three orthogonal orientations (coronal, sagittal, and transaxial). Existing multi-slice[32,33] or 3D model observers[34–38] are either for SKE tasks only or single-orientation SKS tasks.[32]

In this paper, we propose a novel deep learning-based anthropomorphic model observer (DeepAMO) that evaluates multi-orientation, multi-slice image sets to model the clinical diagnostic process of a radiologist or nuclear medicine physician in a clinically realistic 3D defect detection task. The DeepAMO was evaluated on an SKS/BKS tasks using a realistic anatomical background with variation in organ uptake and defect position (and thus orientation and shape). We also propose a novel calibration method that “learns” the underlying distribution of the human observer rating values (including the internal noise) using a mixture density network (MDN). Note that in this context a rating value is the raw data from human observer study and is a numeric value expressing the observer’s level of confidence that a defect is present or absent in a given image. The entire network is trained using human observer rating values so that the output, when applied to an input image volume, is a rating value designed to reproduce the performance of human observers.

A human observer study was conducted using the volumetric display format routinely used at Boston Children’s Hospital (BCH) for clinical interpretation. Quantitative comparisons of the performance between the DeepAMO and human observer are provided in Sec. 3.

2 Materials and Methods

Image quality in this work was measured in terms of performance on the task of detecting renal functional defects in $^{99\mathrm{m}}$Tc-DMSA SPECT. The images used were simulated based on an anthropomorphic digital phantom of 5-year-old (a typical age in DMSA imaging). The phantom and simulation methods are described in Ref. 39. The simulation modeled administered activities (and thus noise levels) based on the North America Consensus Guidelines.[40] Task performance was evaluated using both human observers and the proposed DeepAMO. Both of these observers produced a set of rating values for images where the true defect status was known. These rating values were analyzed using receiver operating characteristic (ROC) analysis methods.[41] The area under the ROC curve (AUC) served as a figure of merit for task performance.

2.1 *Data Simulation*

The projection data for this study were generated using the Advanced Laboratory for Radiation Dosimetry Studies UF NHANES-based phantom.[42] The pediatric phantom used was developed at the University of Florida based on demographic data from the CDC’s National Health and Nutrition Examination Survey (NHANES) data.[43] For this study, we used a 5-year-old male phantom with average girth and kidney size. The phantom was digitized using 0.1-cm cubic voxels. Activity uptake in the kidneys was modeled using data from a single imaging time point (3 h postinjection). A dataset of 47 patients acquired at the BCH was used to estimate the means and standard deviations of kidney uptake in units of activity.

The model previously described in Refs. 44 and 45 was used to simulate defects in the cortical wall of the right kidney consisting of volumes of reduced uptake consistent with focal, acute pyelonephritis. The defects were created at random locations (excluding the area close to the renal pelvis) along the cortical wall. Based on input from an experienced pediatric nuclear medicine specialist, we selected a defect volume of $0.5\ \mathrm{cm}^3$ as a defect size that is clinically relevant for the 5-year-old.

Using this model, we created four randomly located focal transmural renal defects at each of the following macrolocations on the right kidney cortex: upper pole, lower pole, and lateral. There was a total of 12 random locations for the defects generated in this study, modeling an SKS task. We simulated noise-free projection data for the renal cortex, medulla, pelvis, liver, spleen, and background (including all other organs), modeling the physics and acquisition parameters appropriate for $^{99\mathrm{m}}$Tc renal SPECT. The renal activity and relative activity concentrations for structures inside the kidney (the renal cortex, medulla, and pelvis) were randomly sampled from truncated Gaussian distributions with the means, standard deviations, minima, and maxima derived from the patient data acquired at BCH. Parameters for the distributions can be found in Ref. 45. Each single-organ projection was scaled by the product of AA, acquisition duration, and scanner sensitivity. The projections were generated using an analytic projection code that modeled attenuation, the spatially varying collimator-to-detector response,[46] and object-dependent scatter.[47] The code has been previously validated by comparison to Monte Carlo and experimental projection data for imaging of a variety of radionuclides.[48–56]

In this study, the projections were simulated to model a Siemens low-energy, ultra-high-resolution collimator used routinely at BCH for pediatric DMSA studies. Each single-organ projection dataset was generated at 120 projection views over a 360-deg body-contouring orbit with a 0.1-cm projection bin size and then collapsed to a bin size of 0.2 cm. A model of the patient bed obtained from a CT scan of the bed of a Siemens Symbia SPECT/CT system was added to the attenuation map of each computational phantom. Noise-free projection images of the entire phantom were obtained by summing the individual sets of scaled organ projections. Noisy projections were created by simulating Poisson noise using a Poisson pseudo-random generator.

A total of 384 projection images were thus generated, comprised of 16 uptake realizations × 12 defect locations × 2 defect statuses (present or absent). The mean (noise-free) activity distribution was statistically independent for each of these 384 projection images since the kidney uptake and the activity concentration ratio of the cortex to the medulla plus pelvis activity were randomly sampled.

We followed the clinical reconstruction protocol routinely used at BCH. Projection images were reconstructed using the OS-EM iterative reconstruction algorithm with compensation for the geometric collimator-detector response and postfiltered with a Gaussian filter with an FWHM of 5 mm. The reconstructed images were then interpolated and formatted to match the volumetric image display used at the BCH. In this display, 10 coronal, 20 sagittal, and 18 transaxial images with sizes of 96×96 pixels were generated. These composite images were used for training and testing of the proposed model observer and the human observers. Windowing was used to map the image pixel values to a range of 0 to 255. A sample of BCH's volumetric image display is shown in Fig. 1.

2.2 *Proposed Model Observer: Overview*

The DeepAMO is designed based on a hypothetical model of the image interpretation process of a human observer. One alternative of this approach would be to let the neural network "learn" how humans interpret 3D image volumes from the data. For example, the most direct approach would be to input the 3D image volume data into a fully connected network and then to train that network directly with human observer rating values. Such a network would have a large number of parameters. Since each trial (reading of a set of images by a human observer) provides a single scalar rating value, it provides relatively little information for training the network. A very large number of input rating values would thus be required. Since the rating value data are very

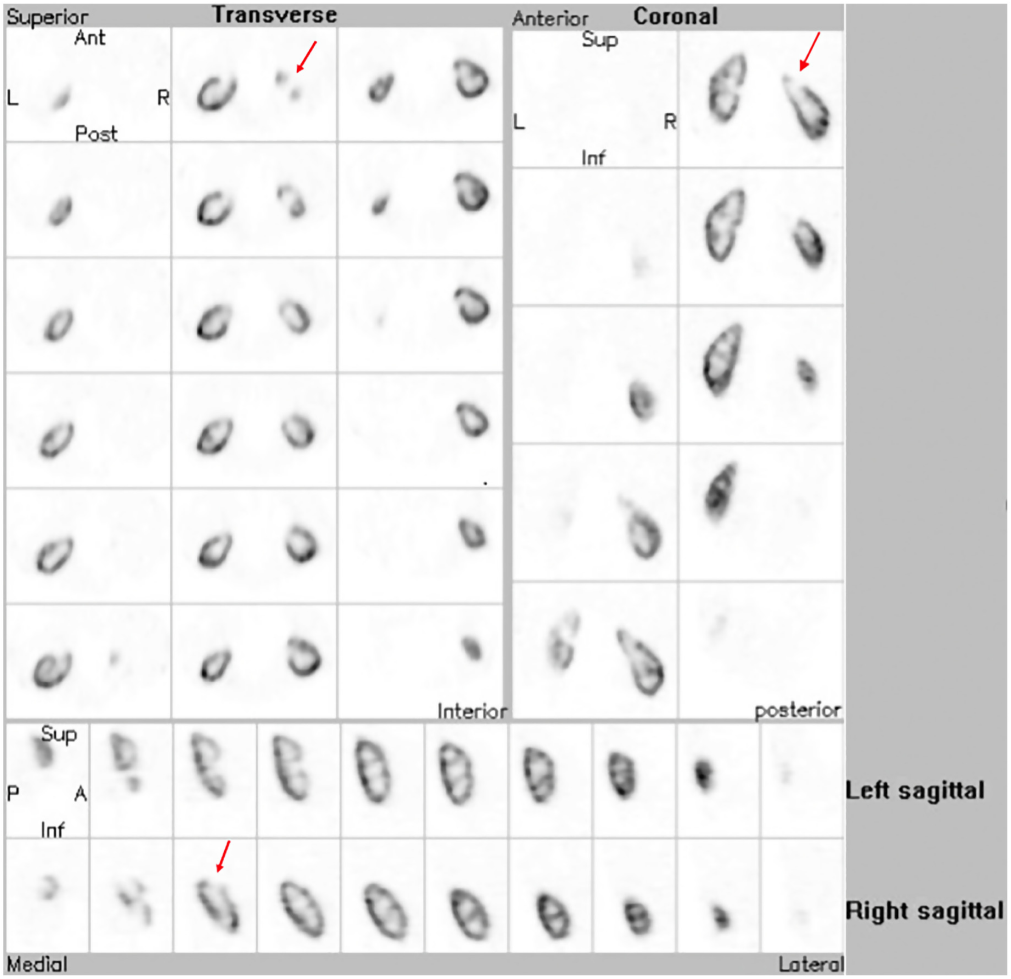

Fig. 1 A sample 48-slice image shown in the volumetric display format routinely used in clinical practice at BCH. The red arrow indicates the location of the functional defect.

expensive to obtain, we have divided the network into stages that are designed to require less human-observer training data. The division of the model is based on how humans interpret the images, as will be described below. The first two stages do not require human observer training data, and the third one maps a low-dimensional feature vector to a scalar rating value.

We hypothesize that a human observer interpreting an image first scans over the orthogonal slices to identify suspicious abnormalities in single slices. If a defect is suspected to be present in one slice (of a particular orientation), the observer confirms that on adjacent slices. The observer would confirm that a defect in one orientation is seen in the other two orthogonal orientations. We suppose that the observer would have more confidence in the presence of a defect if it is found in at least one other orientation.

Thus we propose to implement this decision-making process in three sequential stages. In stage 1, we use a segmentation network to identify defects in three orthogonal slice views. The segmentation is performed using groups of three adjacent slices. In stage 2, we use a deterministic algorithm that confirms the presence of defects in the three orthogonal views and generates a low-dimensional feature vector. In stage 3, we use a MDN to learn the mapping of feature vector to rating value, thus calibrating the DeepAMO to reproduce human observer performance.

2.3 *Proposed Model Observer: Architecture*

A schematic of the proposed DeepAMO is shown in Fig. 2. The input to the segmentation network was the same set of slices used in the previously described volume display used in clinical

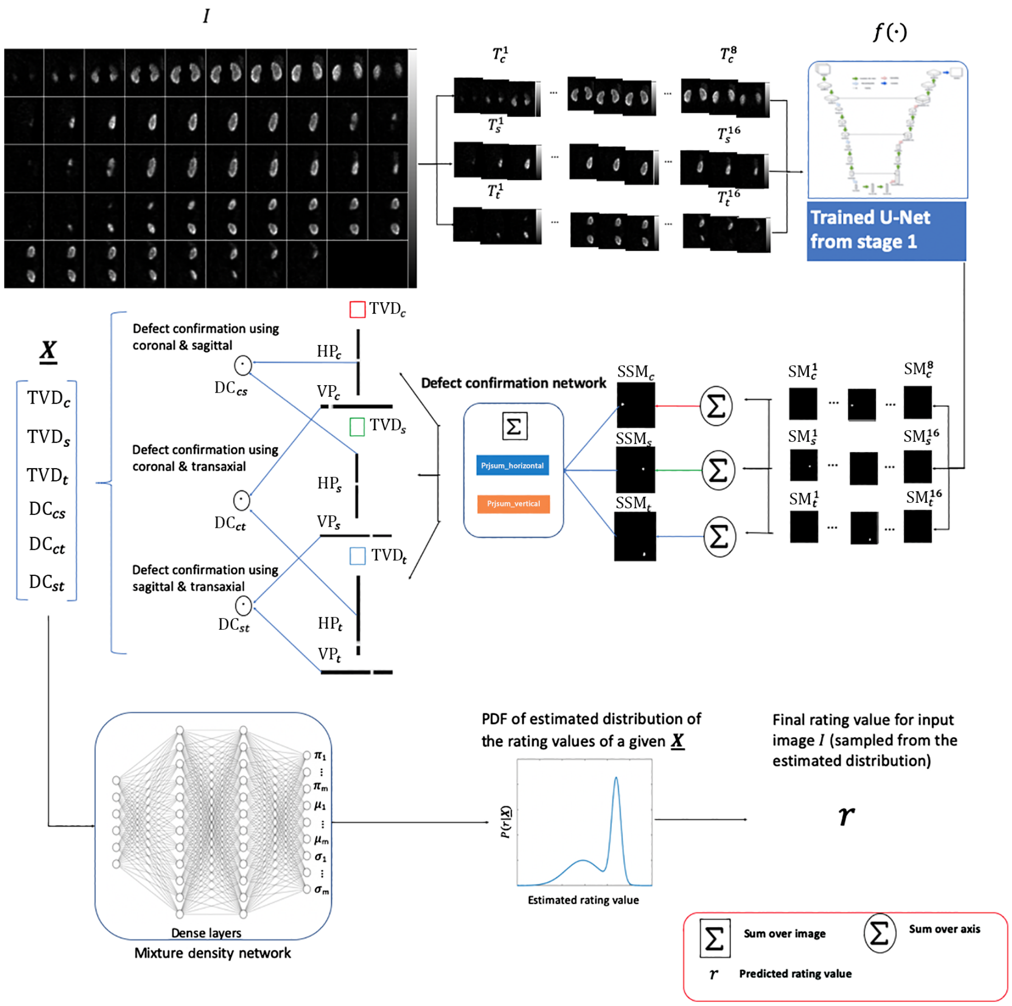
I
T_c^1
T_c^8
T_s^1
T_s^{16}
T_t^1
T_t^{16}
$f(\cdot)$
Trained U-Net from stage 1
$\underline{\boldsymbol{X}}$
TVD$_c$
TVD$_s$
TVD$_t$
DC$_{cs}$
DC$_{ct}$
DC$_{st}$
Defect confirmation using coronal & sagittal
Defect confirmation using coronal & transaxial
Defect confirmation using sagittal & transaxial
HP$_c$
VP$_c$
HP$_s$
VP$_s$
HP$_t$
VP$_t$
Defect confirmation network
Prjsum_horizontal
Prjsum_vertical
SSM$_c$
SSM$_s$
SSM$_t$
SM$_c^1$
SM$_c^8$
SM$_s^1$
SM$_s^{16}$
SM$_t^1$
SM$_t^{16}$
π_1
π_m
μ_1
μ_m
σ_1
σ_m
Dense layers
Mixture density network
PDF of estimated distribution of the rating values of a given $\underline{\boldsymbol{X}}$
$P(r|\underline{\boldsymbol{X}})$
Estimated rating value
Final rating value for input image I (sampled from the estimated distribution)
r
Sum over image
Sum over axis
r Predicted rating value

Fig. 2 A schematic of the proposed model observer, DeepAMO. I is the multi-slice, multi-view input image, T_k^j is the triad, where $k \in (c, s, t)$ represents the slicing direction and $j \in [1, N - 1]$, where N is the number of slices in each orientation. SM$_k^j$ is the output segmentation mask for each triad T_k^j. TVD$_k$ is the TVD seen in each slicing direction computed by summing the corresponding SSM$_k$. SSM$_k$ is the summed segmentation mask along each slicing direction k. HP$_k$ and VP$_k$ are horizontal and VP of the corresponding SSM$_k$. DC$_{\mathrm{cs}}$, DC$_{\mathrm{ct}}$, and DC$_{\mathrm{st}}$ are the three defect confirmation scalars from the defect confirmation network. Note that one triad is fed to the segmentation at a time.

practice, which consists of multiple slices in each of the three orientations: coronal, sagittal, and transaxial. Mathematically, the slice, $S_k^i(m, n)$, and input composite image, $I(m, n, q)$, are related as follows:

$$I(m, n, q_k^i) = S_k^i(m, n). \tag{1}$$

In Eq. (1), q_k^i is the index number for the i'th slice in the slicing direction $k \in (c, s, t)$, and m, n, and q are pixel indices for the x, y, and z axes, respectively.

For each orientation, $N_k - 2$ (N_k slices in each orientation) triads are generated: the first and last slices cannot act as the central slice for a triad. The j'th triad in the slicing direction k is

$$T_k^j(m, n, q) = \{S_k^{j-1}(m, n), S_k^j(m, n), S_k^{j+1}(m, n)\}, \quad j \in [1, N_k - 2]. \tag{2}$$

The output segmentation mask (SM) of each triad is a 2D binary mask of pixels thought to be in the defect. The SMs along each orientation are summed to form a summed segmentation mask (SSM) in order to enhance the defect signal(s) that is (are) present in that orientation. That is:

$$\mathrm{SM}_k^j(m,n) = f[T_k^j(m,n,q)], \tag{3}$$

$$\mathrm{SSM}_k(m,n) = \sum_{j=1}^{n_k} \mathrm{SM}_k^j(m,n), \tag{4}$$

where j is the triad number and k is the slicing direction. $T_k^j(m,n,q)$ and n_k represent the j'th triad and the number of triads in slicing direction k, respectively. Here $f(\cdot)$ denotes the segmentation network.

We propose to implement the process of confirming defect presence in other slicing directions, by projecting and comparing defect information from different slicing directions, through a DC network. Specifically, this is implemented by projecting (i.e., summing) each SSM_k vertically and horizontally and calculating the dot products between the resulting horizontal projections (HP) and vertical projections (VP) from different slicing directions. The HPs and VPs are derived as follows:

$$\mathrm{HP}_k(n) = \sum_{m=0}^{M-1} \mathrm{SSM}_k(m,n), \tag{5}$$

$$\mathrm{VP}_k(m) = \sum_{n=0}^{N-1} \mathrm{SSM}_k(m,n), \tag{6}$$

where M and N are the number of pixels in the x and y axes directions, respectively.

The projection is constructed so that the projections from the different slicing directions are along the same direction in space. To understand this, consider that any two views always share a common axis, and, by projecting the two views onto this common axis, we can confirm information about defect location that is compatible. For example, consider an L-shape object in a 3D space (Fig. 3). By projecting the sagittal and transaxial views vertically, we get two 1D vectors that both contain information about the object's maximum length along the horizontal axis. If the dot product between the two 1D vectors is large, then the object is present at the same location in that direction for both slicing directions. Likewise, we can confirm the object's location along the other two directions via the same projection and dot product operations. This process yields three scalar values, representing the defect agreement along the x, y, and z axis, respectively. We named these three scalar values the DC scalars. They are derived from the HPs and VPs from different slicing directions as follows:

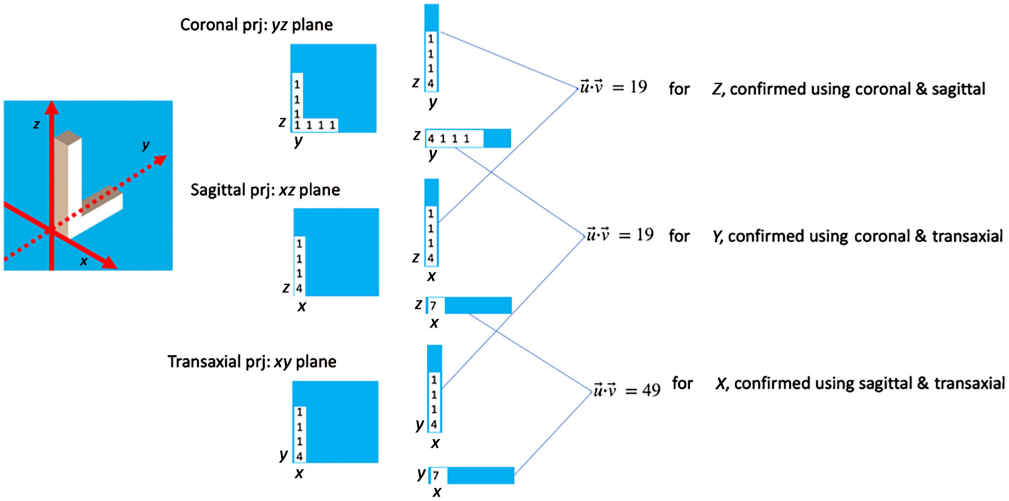
Coronal prj: yz plane
Sagittal prj: xz plane
Transaxial prj: xy plane
z
y
x
1
1
1
1 1 1 1
4 1 1 1
4
7
$\vec{u}\cdot\vec{v}$ = 19 for Z, confirmed using coronal & sagittal
$\vec{u}\cdot\vec{v}$ = 19 for Y, confirmed using coronal & transaxial
$\vec{u}\cdot\vec{v}$ = 49 for X, confirmed using sagittal & transaxial

Fig. 3 An illustration of the process of confirming the defect from different views using projection and dot product in 3D space.

$$\mathrm{DC}_{cs} = \mathrm{HP}_c(n) \cdot \mathrm{VP}_s(m), \tag{7}$$

$$\mathrm{DC}_{ct} = \mathrm{HP}_t(n) \cdot \mathrm{VP}_c(m), \tag{8}$$

$$\mathrm{DC}_{st} = \mathrm{VP}_t(m) \cdot \mathrm{VP}_s(m). \tag{9}$$

The DC scalars are concatenated with the total volume of the defect (TVD) seen in each slicing direction to form a single feature vector. The TVD from each slicing direction is computed as follows:

$$\mathrm{TVD}_k = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \mathrm{SSM}_k(m, n). \tag{10}$$

The resulting six-element concatenated feature vector is then sent to an MDN[57] to generate the rating (test statistic) value. The dense layers in the MDN are meant to model the process of a human making the final decision using combined information from the different directions. The output of the MDN is the set of parameters of a statistical distribution, in this case, a Gaussian mixture model (GMM), as described below.

2.4 *Calibration to Human Observer Data via a Mixture Density Network*

For defect detection tasks, the observer performance is usually measured by the AUC, which ultimately depends on the underlying distribution of the rating values given by the observer. Thus for the purposes of replicating an observer's AUC, we propose to directly learn the mapping of feature vectors to the distribution of the rating values. We hypothesize that more training and testing samples would help better capture the underlying rating value's distribution. However, demonstrating the equivalence of the distributions is a task requiring a large number of rating values. In addition, it is unclear what level of agreement between the true and modeled distribution is required. Thus we are focusing in this work on verifying that the model observer can replicate the AUC values obtained from the set of rating values resulting from an observer study.

A MDN was chosen for the task of mapping the input feature vector into a rating value in order to model the fact that a human observer will give a different rating value for the same input image. The MDN provides parameters of a distribution that can then be sampled to provide multiple, continuously valued ratings from a single set of input feature vectors. This can be useful during testing of the DeepAMO to reduce sampling error.

Typically, an MDN learns an entire probability distribution for the output by modeling the conditional probability distribution of the target data conditioned on the input data. In our case, the desired conditional probability distribution is $P(r|\underline{\boldsymbol{X}})$, where $\underline{\boldsymbol{X}} = [x_{1,\ldots,}x_6]$ is a six-element feature vector and r is a (continuous) human observer rating value for a given input feature vector. For the purpose of modeling any arbitrary probability distribution, the MDN uses a GMM as the conditional probability density function, which can be represented as a linear combination of kernel functions in the form:

$$P(r|\underline{\boldsymbol{X}}) = \sum_{i=1}^{m} \pi_i(\underline{\boldsymbol{X}})\phi_i(r|\underline{\boldsymbol{X}}), \tag{11}$$

where m is the number of components in the mixture and $\{\pi_i(\underline{\boldsymbol{X}})\}$ is the set of mixture coefficients for the kernel functions, which sum to 1. The set $\{\pi_i(\underline{\boldsymbol{X}})\}$ is derived from the output of the MDN and is converted to a set of probabilities as follows:

$$\pi_i(\underline{\boldsymbol{X}}) = \frac{\pi_i}{\sum_{i=1}^{m} \pi_i}, \tag{12}$$

where π_i is the output from the last dense layer, as shown in Fig. 2. The kernel functions $\{\phi_i(r|\underline{\boldsymbol{X}})\}$ are in the form of Gaussian distributions:

$$\phi_i(r|\underline{\boldsymbol{X}}) = \frac{1}{\sigma_i(\underline{\boldsymbol{X}})\sqrt{2\pi}} \exp\left(-\frac{[r-\mu_i(\underline{\boldsymbol{X}})]^2}{2\sigma_i(\underline{\boldsymbol{X}})^2}\right), \tag{13}$$

where $\sigma_i(\underline{\boldsymbol{X}})$ and $\mu_i(\underline{\boldsymbol{X}})$ are the estimated standard deviation and mean for the input feature vector $\underline{\boldsymbol{X}}$ and they come from the output of the last dense layer. Each mean and standard deviation is for a particular mode [with the mixing coefficient $\pi_i(\underline{\boldsymbol{X}})$] in the GMM. Note that $\{\pi_i(\underline{\boldsymbol{X}})\}$ is a function of $\underline{\boldsymbol{X}}$. So $\{\pi_i(\underline{\boldsymbol{X}})\}$ can also be regarded as a set of prior probabilities of the target data.

In training, the loss is computed using the human observer rating value r_{true} and the predicted mixture distribution $P(r|\underline{\boldsymbol{X}})$ from the MDN as follows:

$$L = -\log P(r_{\text{true}}|\underline{\boldsymbol{X}}). \tag{14}$$

Each mean $\mu_i(\underline{\boldsymbol{X}})$ in the predicted mixture distribution was trained to minimize the loss during training, which should follow the relative prevalence of the rating values given by the human observer(s). Each standard deviation $\sigma_i(\underline{\boldsymbol{X}})$ was trained to model the variance of each mean $\mu_i(\underline{\boldsymbol{X}})$.

In testing, a rating value is predicted by first randomly sampling the mixing coefficients and then sampling from the Gaussian distribution corresponding to that sampled mixing coefficient with the corresponding estimated mean and standard deviation. Multiple sample rating values can be generated to improve the uncertainty in AUC values calcluated from the testing data.

2.5 *DeepAMO Performance on Unseen Images*

To estimate the number of images needed to train the DeepAMO, we used simulated feature vectors and rating values to train and test the MDN. The criterion for judging the number of images to be sufficient is the statistical confidence level needed in comparing AUC values between the proposed model and human observer. We assumed that each element of the feature vectors followed a Gaussian distribution (with dependencies introduced between the $\text{TVD}_k s$ and DC scalars) and the rating values described by a mixture of Gaussians.

The feature vectors were simulated by first generating values for the TVD_k, one for each orientation. Each TVD_k was assumed to be mutually independent and was generated by sampling from independent Gaussian distributions [N ($\mu = 25, \sigma = 5$) for large TVD_k and N ($\mu = 5$, $\sigma = 1$) for small TVD_k]. The sampled TVD_k values were then used to calculate the means and standard deviations of the DC scalars, which were also assumed to follow a Gaussian distribution:

$$\mu_{cs} = \text{TVD}_c \times \text{TVD}_s, \tag{15}$$

$$\sigma_{cs} = \frac{\mu_{cs}}{3}, \tag{16}$$

$$\mu_{ct} = \text{TVD}_c \times \text{TVD}_t, \tag{17}$$

$$\sigma_{cs} = \frac{\mu_{cs}}{3}, \tag{18}$$

$$\mu_{st} = \text{TVD}_s \times \text{TVD}_t, \tag{19}$$

$$\sigma_{st} = \frac{\mu_{st}}{3}. \tag{20}$$

The rating values of each feature vector were sampled from multi- or unimodal Gaussian distributions. The distribution parameters for these simulated rating values were derived qualitatively from distributions of rating values from human observer studies and are shown in Table 1. For each feature vector, we then sampled N rating values from the assumed distribution to simulate the appropriate level of inter- or intraobserver variability in the data. Specifically, in this work, we sampled two rating values for each feature vector. So there were 15,000 (2500 × 3 feature vector types × 2 repeated samples) feature vector and rating value pairs for the case

Table 1 Summary of distribution parameters for the simulated rating values.

	Definitely-yes		Not-sure		Definitely-no	
Defect-present feature vector type						
Rating value means	7	10	2	4	−3	
Standard deviation	1.2	0.2	1.2	1.2	0.2	
Component weight	0.5	0.5	0.5	0.5	1	
Defect-absent feature vector type						
Rating value means	−10	−8	−2	−4	2	5
Standard deviation	0.2	1.2	0.7	1.2	0.5	0.8
Component weight	0.5	0.5	0.5	0.5	0.5	0.5

that had 2500 samples/feature vector type, and 30,000 in total for both the defect-present and defect-absent cases.

In the simulation experiment, we generated three types of feature vectors for each class (defect-present and defect-absent): definitely-present (three large $\mathrm{TVD}_k s$), equivocal (two large $\mathrm{TVD}_k s$ and one small TVD_k), and definitely-absent (three small $\mathrm{TVD}_k s$), reflecting different levels of user confidence in making the decision. For example, the feature vectors that belong to the definitely-present type in the defect-present class were generated by sampling three large values for the three TVD_ks, modeling a high level of success of the segmentation network in detecting the defect in slices from all three orientations. The other two types (equivocal and definitely-absent, respectively) contained two and one large values (assigned randomly to any of the three orientations) in the TVD_ks to simulate different degrees of success in detecting the defect in the three orientations.

2.6 *Training and Testing of DeepAMO*

The proposed model observer was trained in two stages. First, the segmentation network was trained given the ground-truth defect SMs. Next, the MDN was trained using the output from the trained segmentation network and the human observer rating values.

The segmentation network was trained with triad images and their corresponding binary defect segmentation labels. Since each defect only contained about 0.5% of the kidney cortex volume, the number of defect-present triads was much smaller than the defect-absent ones, making this a highly imbalanced dataset. Thus we adopted data augmentation of the defect-present triads to balance the training data. We enriched the data by forming an additional seven sets of raw images and their labels by rotating each original defect-present triad image by 90, 180, and 270 deg and flipping them and the original dataset upside down. The exponential logarithmic loss in Ref. 58 was adopted to emphasize segmentation of small structures with the best-performing weights ($\omega_{\text{cross}} = 0.2$ and $\omega_{\text{dice}} = 0.8$) suggested in this paper. The mixed exponential logarithmic loss in Ref. 58 was adopted in order to aid segmentation of small structures:

$$L = \omega_{\text{cross}} L_{\text{cross}} + \omega_{\text{dice}} L_{\text{dice}}, \tag{21}$$

where ω_{cross} and ω_{dice} are the mixing coefficients of the exponential categorical cross-entropy loss and the exponential logarithmic dice loss:

$$L_{\text{cross}} = E(\{-\ln[p(x)]\}^{\beta_{\text{cross}}}), \tag{22}$$

$$L_{\text{dice}} = E\left[w_l\left(-\ln\left\{\frac{2[\sum_x y(x)p(x)]}{[\sum_x y(x) + p(x)]}\right\}\right)^{\beta_{\text{dice}}}\right], \tag{23}$$

where x is the pixel position and $p(x)$ and $y(x)$ are the network's predicted value and the ground truth segmentation at position x, respectively. $E(\cdot)$ is the expectation value with respect to x. w_l is the label weight, which was used to increase the importance of objects comprised of a small number of voxels such as the defect:

$$w_l = \left(\frac{\sum_k f_k}{f_l}\right). \tag{24}$$

For the segmentation network, we adopted a shallow version of the U-Net.[59] We used a shallow (in depth) network due to the relatively small amount of training data available in this study. A deeper network might be needed for a larger number of signal and anatomical variations. The architecture of the segmentation network used in this study is shown in Fig. 4. Gaussian noise with a standard deviation of 1.0 was added to the renormalized input image (ranges 0 to 255) to prevent overfitting. We searched for the optimal network capacity (depth) for the segmentation network. There was a trade-off between producing the highest dice score and using the smallest number of parameters. However, it was observed that there was a relatively small increase in dice score with increased number of parameters in the tested network architectures, and the dice

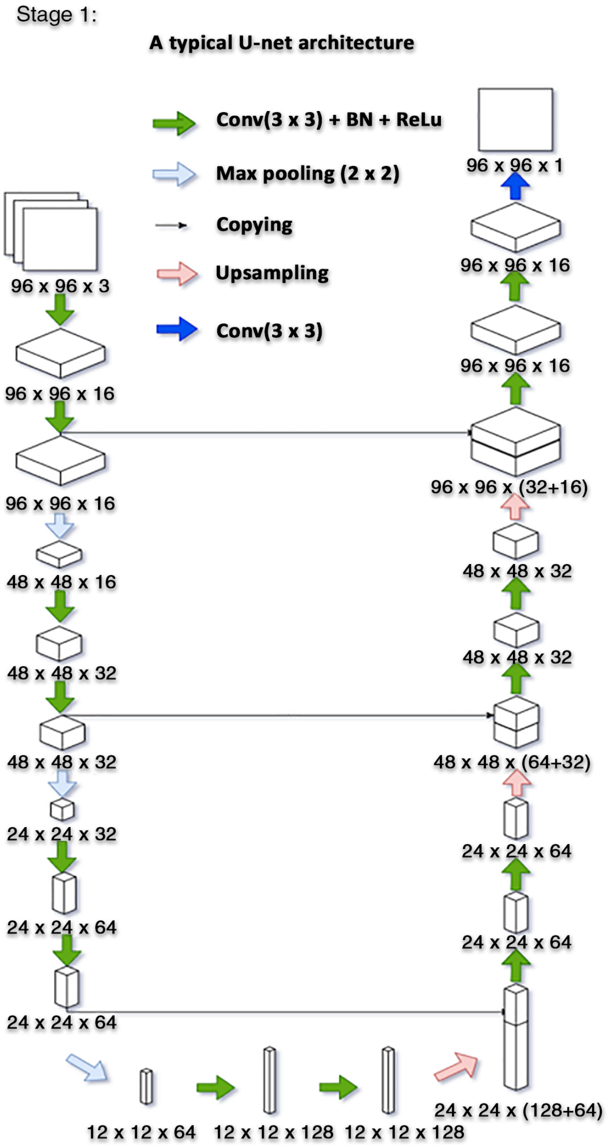

Fig. 4 Segmentation network architecture used in this study.

scores were all reasonably high. So we adopted the network architecture that had the smallest number of parameters and yet gave a reasonably high dice score (0.97). The train and test datasets had 12,288 and 3072 triads, respectively. Data augmentation was done on-the-fly. We used an Adam[60] optimizer with a learning rate of 0.001 and a batch size of 200. The segmentation network was trained for 500 epochs to investigate convergence, but very near convergence was achieved in about 100 epochs. The training took about 2 h to converge (∼100 epochs) on a single Tesla K40 GPU.

For the MDN, the number of mixtures was chosen by visually inspecting the distribution of the target human observer's rating values. The number of mixtures was selected to be equal or greater than the number of modes observed in the distribution of the observer's rating values. For this study, we used an MDN with three fully connected dense blocks each with 128 dense units and a dropout rate of 0.5. Each dense block contained a dense layer with the above-mentioned dense units and a batch normalization layer, followed by the rectified linear unit activation and dropout layer. The outputs from the last dense block were then connected to three dense layers that, respectively, output the mixing coefficients $\pi_i(\underline{\boldsymbol{X}})$, means $\mu_i(\underline{\boldsymbol{X}})$, and sigmas $\sigma_i(\underline{\boldsymbol{X}})$ for the estimated distribution. The number of mixing coefficient was set to 5 since we obsevered about 5 modes in the distribution of huamn observers' rating vlaues.

2.7 *Human Observer Study*

The same image display format shown in Fig. 1 was used in the human and model observer studies. A sample display of the human observer GUI is shown in Fig. 5. In the study, the observer was asked to rate their confidence that a defect was present on a continuous scale ranging between 1 and 5 (later mapped to −10 to 10), with the highest number representing the greatest confidence that a defect was present. To familiarize themselves with the display program and the nature of the clinical defect detection task, all observers participated in an initial training session comprised of 24 images. In the training session, phantom images of the kidney cortex were provided as ground truth to the observers once their rating value was recorded. Additional training was done as described below. Rating values from the training study were not used in training the network.

Two senior medical imaging physics PhD students participated in the human observer study. A total of 384 of the composite images described in Sec. 2.1 were used. To simulate an SKS detection task, the train and test datasets were created without requiring a balance of defect locations. Thus the test dataset could contain defect locations that were not present in the initial training dataset. The images were divided into an initial training set and three test blocks. The block layout for each observer is shown in Table 2. In each test block, a refresher set of 24 images was provided to refresh the observer's memory about the task. A total of 288 rating values were collected from each observer.

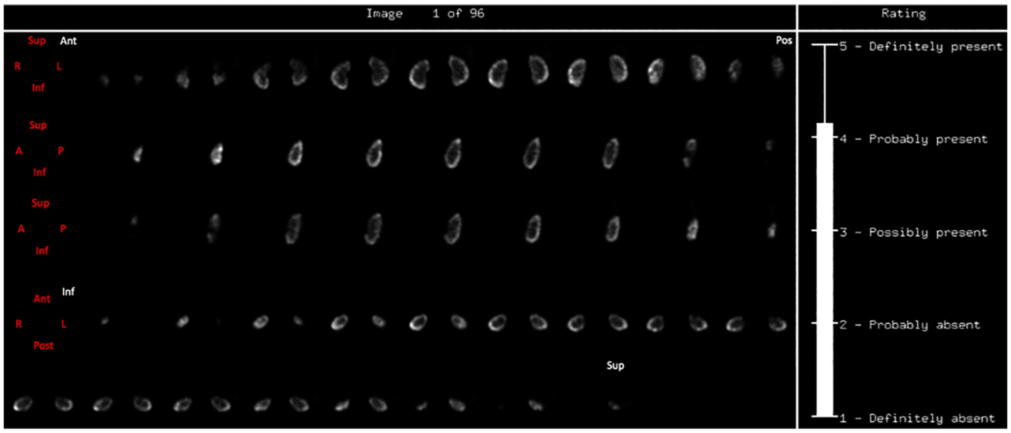

Fig. 5 A sample image of the GUI used in the human observer study for DeepAMO.

Table 2 Summary of human observer study block partition.

Session	Initial training images	Blocks	Image/block	Total images
	24	1	24 training	24
1	0	1	24 training/96 test	120
2	0	1	24 training/96 test	120
3	0	1	24 training/96 test	120

2.8 *Equivalence Hypothesis Testing*

An equivalence statistical hypothesis test[61] was conducted to test whether the performance (as measured by the AUC) of the human observer and the proposed model observer was statistically equivalent on a defect detection task. The null hypothesis and alternative hypothesis are expressed as follows:

$$H_0\colon\ |\mathrm{AUC}_{\mathrm{HO}} - \mathrm{AUC}_{\mathrm{MO}}| = \delta \quad H_1\colon\ |\mathrm{AUC}_{\mathrm{HO}} - \mathrm{AUC}_{\mathrm{MO}}| < \delta, \tag{25}$$

where $\mathrm{AUC}_{\mathrm{HO}}$ and $\mathrm{AUC}_{\mathrm{MO}}$, respectively, are the AUC values for the human and proposed model observer; δ is a threshold for an important difference (margin of difference) between $\mathrm{AUC}_{\mathrm{HO}}$ and $\mathrm{AUC}_{\mathrm{MO}}$. The difference parameter was used as it is very difficult, if not impossible, to show statistically that two quantities are exactly equal. In addition, small differences are not practically important. The difference parameter was prespecified and is a determinant of sample size: in order to prove better equivalence (smaller δ), a larger sample size is required. In order to reject the null hypothesis, the confidence intervals (CIs) of the difference of the AUCs must lie within the interval defined by the margin of difference parameter, as described in Ref. 61 and illustrated in Fig. 6. For this study, we set the δ to be 0.05 (equivalent to a 95% confidence level). That is, as long as the CIs of the ΔAUC are found to be smaller than 0.05, the null hypothesis can be rejected and equivalence of the human and model observer can be claimed.

In order to calculate the CIs for the differences in the AUCs (ΔAUCs), we conducted a 5×2-fold cross-validation experiment using data generated by the two human observers. A total of 576 rating values (288 images × 2 observers) were used in training and testing of the proposed model observer. The data were partitioned randomly for each of the five trials, and a 50-50 train-to-test fraction was adopted. Within each trial, the train and test data were switched between the first and second fold. We used a 50-50 split strategy to divide the data, as we assumed that the number of images in the test dataset should not be too small, otherwise, the distribution of rating values produced would be too coarse to represent the observer's true performance, thus resulting in unfair AUC comparisons. However, we have not investigated whether the 50-50 splitting is optimal.

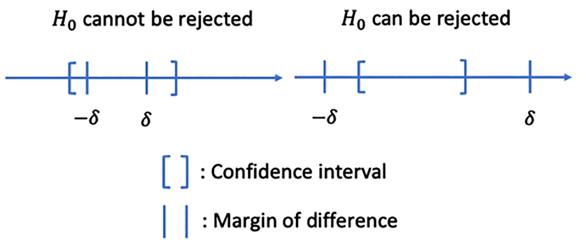

Fig. 6 A pictorial illustration of the rejectable and unrejectable case in equivalence hypothesis testing.

2.9 *Comparison of DeepAMO to an Unoptimized Scanning Linear Observer*

The performance of a scanning linear discriminant observer (SLDO) study was evaluated using the same reconstructed images as described in Sec. 2.1. However, since the background and defects did not have closed form expressions for the observer and the observer had to be trained at each potential defect location, we limited scan range for the SLDO to only slices that could actually contain a defect. This somewhat reduces the difficulty of the task by eliminating the chance of making a mistake, e.g., due to the presence of a noise artifact in those slices, as described in Sec. 2.2.

In the SLDO study, we applied the observer to a 3-slice composite image. The composite image was formed by extracting the coronal, transaxial, and sagittal slices containing the defect centroid from the 3D reconstructed image. The SLDO was applied to the defect at the intersection point of those slices. Thus the SLDO did not have to perform the defect confirmation in the three directions. All slices had a size of 128 × 128 pixels and their defect centroid shifted to the center of the image. Samples of the defect-present and defect-absent composite image are shown in Fig. 7. We used seven non-overlapping rotationally symmetric difference-of-mesa channels. The starting frequency and the width of the first channel were 0.5 cycles per pixel, and subsequent channels had widths that doubled and abutted the previous channel. The frequency domain channels and corresponding spatial templates are shown in Fig. 8.

Each of the 7 spatial domain templates was applied to each of the 3 images (transaxial, sagittal, and coronal) to give a 21-element feature vector. Each element in the resulting feature vector was obtained by taking the dot product of a spatial domain template with an input composite image. These feature vectors served as inputs to train and test the SLDO as described below.

To apply the SLDO on a test image, we first generated N feature vectors of each test image, corresponding the $N = 12$ different defect locations. Then we trained a different SLDO on the feature vectors at each of the 12 potential defect locations. For each test image, we applied each

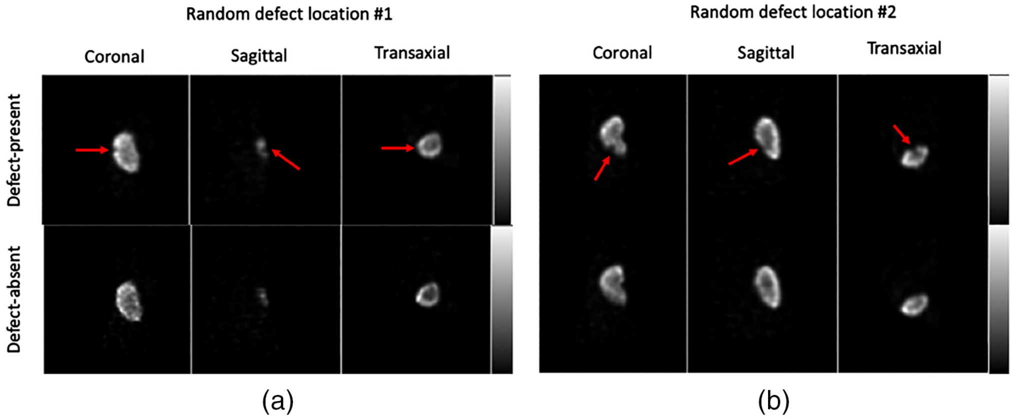

Fig. 7 (a), (b) The defect-present and defect-absent composite image at two different randomly sampled defect locations, respectively. The red arrows mark the exact location of the defect inside each slice.

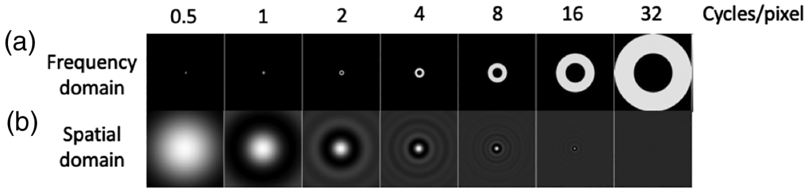

Fig. 8 Images of the seven anthropomorphic DOM channels used in this work. (a) The frequency channels and (b) the spatial domain templates. From left to right, the start frequencies and widths of the channels were 0.5, 1, 2, 4, 8, 16, and 32 cycles/pixel. The spatial templates are the analytic inverse Fourier transforms of the frequency channels sampled at the image pixel size.

of the 12 SLDOs to the feature vectors from each of the potential defect locations, producing a set of 12 test statistics and selected the largest to serve as the test statistic for that test image. We used a leave-one-out training–testing strategy. In this strategy, one feature vector was left-out (i.e., not used in the training), and the remaining vectors were used to train the observer. In our case, the feature vector corresponding to the ground-truth defect location of the test image was left out in training the SLDO for that defect location. The trained SLDO was then applied to the left-out vector to produce a test statistic for that defect location. ROC analysis was performed on the test statistics using the LABROC4 code (68), and the AUC calculated. Bootstrapping and non-parametric analysis were used to compute 95% CIs for the AUC value.

3 Results

3.1 *DeepAMO on Simulated Data*

The results (Fig. 9) show the degree of similarity between the histograms (distributions) of the simulated test data (simulated unseen data). The degree of similarity increased as the total number of samples increased, indicating that the MDN was capable of handling complex distributions of observer's rating values. This result agrees with the hypothesis that the MDN requires a modest amount of training data in order to learn the underlying behavior of the observer on unseen data. Here we assumed that the underlying behavior of the observer was encoded in the distribution of that observer's rating values (training data).

The results also demonstrated that there is a trade-off between the CI width of the ΔAUC and the total number of samples in the dataset. Bootstrapping was used to calculate the non-parametric CIs on the ΔAUC. The ΔAUCs and 95% CIs on the AUCs are summarized in

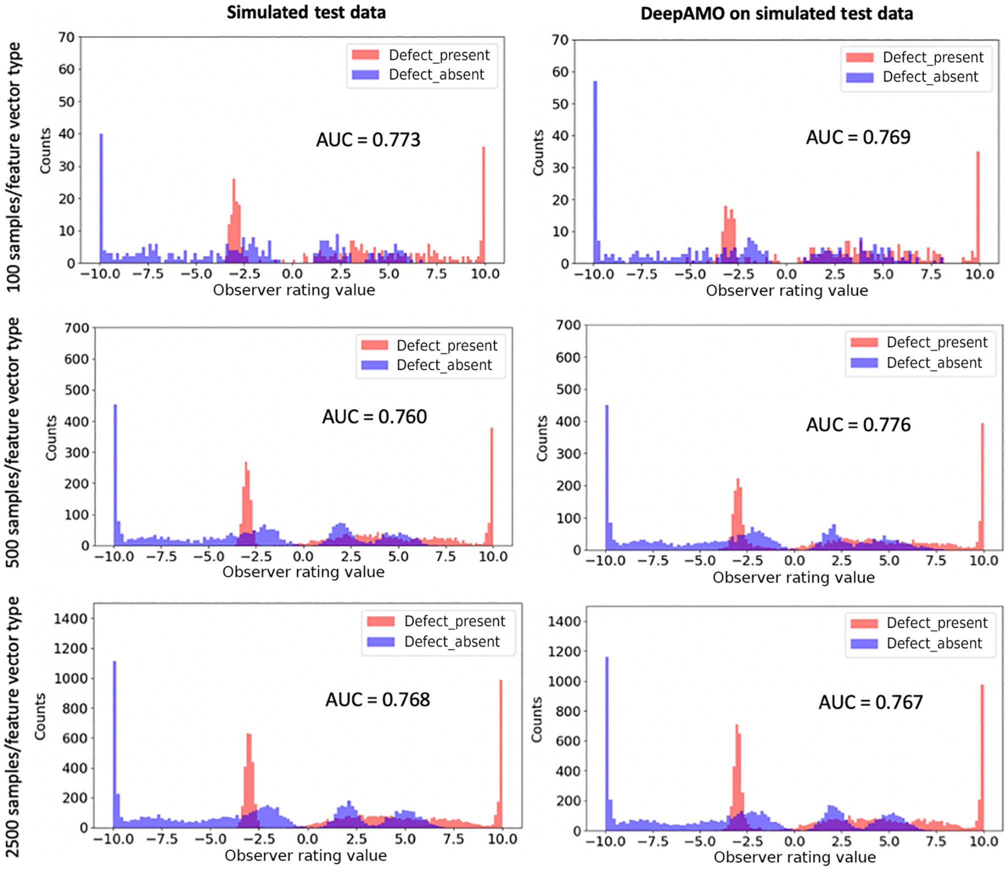

Fig. 9 Plots of histograms of the rating values of the simulated feature vectors (test data only) and predicted rating values on these data given by the DeepAMO. The plots show the class 0 and 1 (defect present and absent, respectively) as well as the calculated AUC value.

Table 3 Summary of simulation results

Number of samples per feature vector type	AUC of DeepAMO on simulated test data	AUC of simulated test data (ground truth)	ΔAUC	95% CI on ΔAUC	CI width
100	0.773	0.769	0.004	$[-0.0502, 0.0477]$	0.0979
500	0.760	0.776	–0.015	$[-0.0352, 0.0261]$	0.0613
2500	0.768	0.767	0.001	$[-0.0074, 0.0089]$	0.0163

Table 3. The results show that the 100, 500, and 2500 samples/feature vector type cases had decreasing widths of the CIs of ΔAUC, indicating that, as expected, more samples are needed to demonstrate greater equivalence (smaller δ) between the human and proposed model observer. The data also suggest that training set size is an important parameter in determining the bounds of the 95% CI on the ΔAUCs.

3.2 *DeepAMO Test Results*

For stage I, the highest dice score achieved on the validation data for the best segmentation network was 0.975. The validation was done on a balanced dataset with 50% of the triads containing a defect.

The AUC values for the human observers and the corresponding DeepAMOs for the 5×2-fold cross-validation experiment are summarized in Table 4. The mean and standard deviation of the ΔAUC were 0.03 and 0.0204, respectively. The 95% CI for the ΔAUC was $[-0.0174, 0.0426]$, under the assumption that ΔAUC was normally distributed. The results of the study show that the null hypothesis with a margin of difference (δ) >0.0426 can be rejected at a confidence level of 95%, with this training set comprised of 288 samples. The histograms of the rating values from the human observers and the DeepAMOs for the 5×2-fold cross-validation experiment are shown in Fig. 10. The AUC value is given at the top of each plot in that figure. The distributions of the rating values for the human and model observer are qualitatively similar.

3.3 *Scanning-Linear Observer Test Results*

The mean AUC for the scanning-linear discriminant observer and its 95% confidence internal was 0.992 and [1.00, 0.986], respectively. Without an implementation of internal noise, the SLDO had a substantially higher AUC (0.99) compared to the DeepAMO and human observer.

Table 4 Summary of stage II training results

	First fold		Second fold				
Trial#	AUC HO	AUC DeepAMO	AUC HO	AUC DeepAMO	ΔAUC first fold	ΔAUC second fold	Mean ΔAUC per trial
1	0.829	0.79	0.797	0.75	0.039	0.05	0.045
2	0.814	0.77	0.816	0.78	0.044	0.036	0.04
3	0.814	0.82	0.815	0.77	–0.01	0.045	0.018
4	0.82	0.77	0.809	0.8	0.046	0.007	0.027
5	0.826	0.82	0.806	0.77	0.008	0.035	0.022

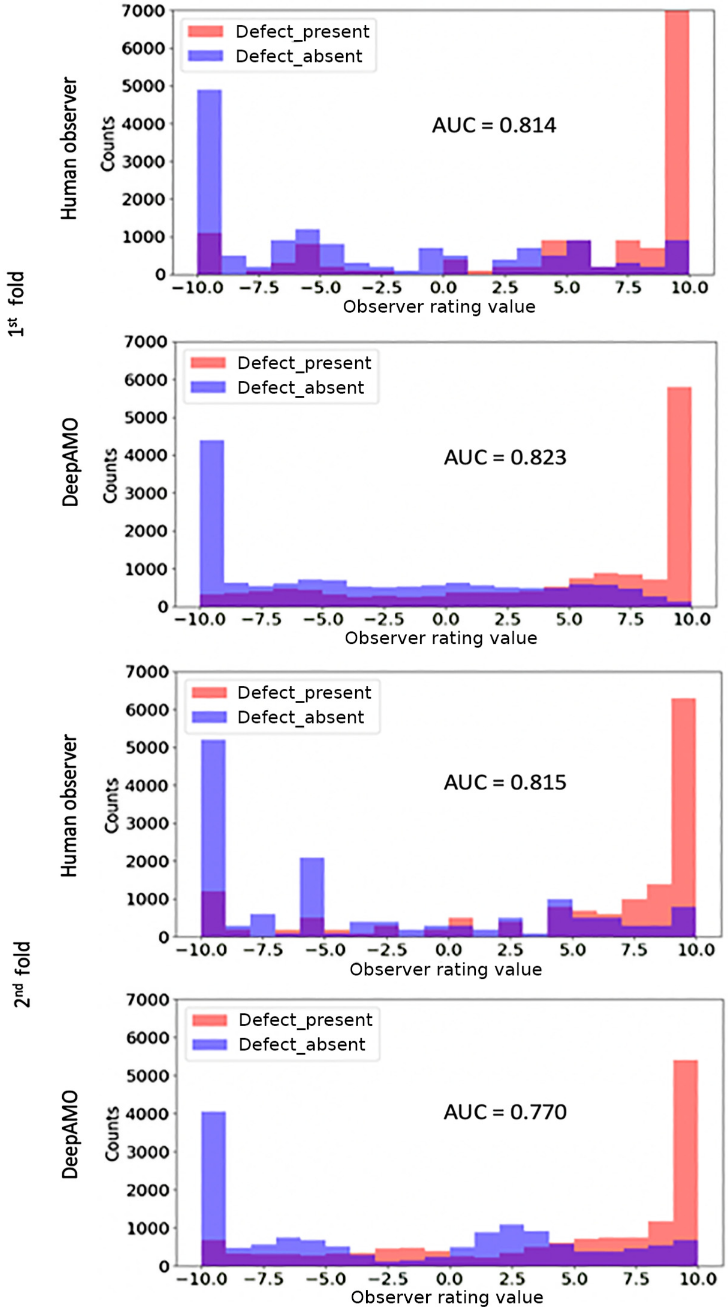

Fig. 10 Histograms of predicted rating values given by DeepAMO on unseen human observer data from the third trial of the 5 × 2-fold cross validation experiment (other trials have similar patterns). Note that multiple predicted rating values were generated for each test image during testing of the DeepAMO to reduce sampling error. The histograms of the other half of human observer data used for training the DeepAMO are not shown in the plot.

4 Discussion

One limitation of this paper is that the simulated dataset has limited background (anatomical) and signal (shape and size) variation. However, we believe that this limitation does not detract from this paper's demonstration that the proposed network architecture can model human observer performance. A dataset with greater anatomical and signal variations might require a different architecture for the segmentation network. However, as long as the segmentation network produced results that distinguish between the defect-present and absent cases at least as well as a human observer, the subsequent stages could still match that performance to human observer performance.

A second limitation of this paper is the use of non-physician observers. Non-physicians were used because of the difficulty of recruiting physician observers to perform a study of this nature. Although the lack of physician observers would clearly affect the clinical diagnostic task, the task that the observers performed in this study was limited to identifying defects in images. We believe that well-trained non-physicians, with sufficient training, can perform well on this more limited task. In addition and more importantly, the purpose of this paper was to validate the ability of the proposed model observer to reproduce human observer defect detection performance and not to generate data on performance that impacts a clinical task. So even if the human observers used performed poorly compared to physicians, the data demonstrate that the model can reproduce their performance. The limitations of the human visual system that degrade performance on defect detection are present even for the non-physician observers, and this work demonstrates the ability of the proposed observer to model these limitations. Therefore, we believe that the data from the observers used in this study demonstrate the utility of the proposed method.

A third limitation of this paper is that the performance comparison study between the DeepAMO and the SLDO. We only compared the performance of the DeepAMO to a limited implementation of an SLDO modeling a simplified defect detection task (using single-slice, multi-orientation images). The SLDO implementation used rotationally symmetric channels for a non-symmetric signal and did not include an internal noise model. In addition, we limited the scan range for the SLDO to only slices that could actually contain a defect in this particular dataset. That restriction could have reduced the difficulty of the defect detection task as explained in Sec. 2.9. Thus we cannot conclude definitively that the DeepAMO is better than an SLDO. A comprehensive study would require calibrating the noise model prior to applying it and was beyond the scope of this work.

A potential concern for the DeepAMO is the training time (∼2 h) required by the segmentation network. However, computational cost for a scanning CHO to read a 3D image can be quite computationally intensive as well. Additionally, substantially faster GPUs than the one used in this study are now available and could reduce this time substantially.

5 Conclusions

The DeepAMO model developed in this paper was able to reproduce human observer ratings for the task of interpreting 3D image volumes based on realistic simulations of pediatric renal disease. This indicates that building conceptual components of the reading process (segmentation, confirmation, and feature synthesis into a score) into network models can allow training of these models in settings where limited training images are available. The results show that the performances of the proposed model and human observers on unseen images were equivalent with respect to a margin of difference in the AUC (ΔAUC) of 0.0426 at $p < 0.05$, for a training set of 288 samples. The proposed framework could be readily adapted to model human observer performance on detection tasks for other imaging modalities such as PET, CT, or MRI.

Disclosures

No potential conflicts of interest relevant to this article exist.

Acknowledgments

This work was supported by NIBIB under Grant No. R01-EB013558. 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. X. He and S. Park, “Model observers in medical imaging research,” *Theranostics* **3**(10), 774–786 (2013).

2. H. H. Barrett et al., "Objective assessment of image quality. 2. Fisher information, Fourier crosstalk, and figures of merit for task-performance," *J. Opt. Soc. Am. A* **12**(5), 834–852 (1995).
3. H. H. Barrett, C. K. Abbey, and E. Clarkson, "Objective assessment of image quality. III. ROC metrics, ideal observers, and likelihood-generating functions," *J. Opt. Soc. Am. A* **15**(6), 1520–1535 (1998).
4. H. H. Barrett, "Objective assessment of image quality—effects of quantum noise and object variability," *J. Opt. Soc. Am. A* **7**(7), 1266–1278 (1990).
5. H. H. Barrett et al., "Objective assessment of image quality. IV. Application to adaptive optics," *J. Opt. Soc. Am. A* **23**(12), 3080–3105 (2006).
6. H. H. Barrett et al., "Objective assessment of image quality. VI Imaging in radiation therapy," *Phys. Med. Biol.* **58**(22), 8197–8213 (2013).
7. H. H. Barrett and K. J. Myers, *Foundations of Image Science*, Wiley-Interscience, Hoboken, NJ (2004).
8. H. H. Barrett et al., "Task-based measures of image quality and their relation to radiation dose and patient risk," *Phys. Med. Biol.* **60**(2), R1–R75 (2015).
9. K. J. Myers and H. H. Barrett, "Addition of a channel mechanism to the ideal-observer model," *J. Opt. Soc. Am. A* **4**(12), 2447–2457 (1987).
10. M. B. Sachs, J. Nachmias, and J. G. Robson, "Spatial-frequency channels in human vision," *J. Opt. Soc. Am.* **61**(9), 1176–1186 (1971).
11. S. Park et al., "Channelized-ideal observer using Laguerre–Gauss channels in detection tasks involving non-Gaussian distributed lumpy backgrounds and a Gaussian signal," *J. Opt. Soc. Am. A* **24**(12), B136–150 (2007).
12. A. E. Burgess, "Visual perception studies and observer models in medical imaging," *Semin. Nucl. Med.* **41**(6), 419–436 (2011).
13. J. L. Harris, "Resolving power and decision theory," *J. Opt. Soc. Am.* **54**(5), 606–611 (1964).
14. K. M. Hanson and K. J. Myers, "Rayleigh task-performance as a method to evaluate image-reconstruction algorithms," in *Maximum Entropy and Bayesian Methods*, Fundamental Theories of Physics, W. T. Grandy and L. H. Schick, Eds., Vol. 43, pp. 303–312, Springer, Dordrecht (1991).
15. R. F. Wagner, K. J. Myers, and K. M. Hanson, "Task-performance on constrained reconstructions—human observer performance compared with suboptimal Bayesian performance," *Proc. SPIE* **1652**, 352–362 (1992).
16. P. F. Judy, R. G. Swensson, and M. Szulc, "Lesion detection and signal-to-noise ratio in CT images," *Med. Phys.* **8**(1), 13–23 (1981).
17. K. J. Myers et al., "Effect of noise correlation on detectability of disk signals in medical imaging," *J. Opt. Soc. Am. A* **2**(10), 1752–1759 (1985).
18. A. Sen, F. Kalantari, and H. C. Gifford, "Task equivalence for model and human-observer comparisons in SPECT localization studies," *IEEE Trans. Nucl. Sci.* **63**(3), 1426–1434 (2016).
19. H. H. Barrett et al., "Model observers for assessment of image quality," *Proc. Natl. Acad. Sci. U. S. A.* **90**(21), 9758–9765 (1993).
20. H. C. Gifford et al., "Channelized Hotelling and human observer correlation for lesion detection in hepatic SPECT imaging," *J. Nucl. Med.* **41**(3), 514–521 (2000).
21. J. Yao and H. H. Barrett, "Predicting human-performance by a channelized Hotelling observer model," *Proc. SPIE* **1768**, 161–168 (1992).
22. S. Sankaran et al., "Optimum compensation method and filter cutoff frequency in myocardial SPECT: a human observer study," *J. Nucl. Med.* **43**(3), 432–438 (2002).
23. E. C. Frey, K. L. Gilland, and B. M. Tsui, "Application of task-based measures of image quality to optimization and evaluation of three-dimensional reconstruction-based compensation methods in myocardial perfusion SPECT," *IEEE Trans. Med. Imaging* **21**(9), 1040–1050 (2002).
24. X. He, J. M. Links, and E. C. Frey, "An investigation of the trade-off between the count level and image quality in myocardial perfusion SPECT using simulated images: the effects of statistical noise and object variability on defect detectability," *Phys. Med. Biol.* **55**(17), 4949–4961 (2010).

25. M. P. Eckstein, C. K. Abbey, and J. S. Whiting, "Human vs. model observers in anatomic backgrounds," *Image Perception* **3340**, 16–26 (1998).
26. S. D. Wollenweber et al., "Comparison of Hotelling observer models and human observers in defect detection from myocardial SPECT imaging," *IEEE Trans. Nucl. Sci.* **46**(6), 2098–2103 (1999).
27. S. Park et al., "Efficiency of the human observer detecting random signals in random backgrounds," *J. Opt. Soc. Am. A* **22**(1), 3–16 (2005).
28. Y. Zhang, B. T. Pham, and M. P. Eckstein, "Evaluation of internal noise methods for Hotelling observer models," *Med. Phys.* **34**(8), 3312–3322 (2007).
29. J. G. Brankov, "Evaluation of the channelized Hotelling observer with an internal-noise model in a train-test paradigm for cardiac SPECT defect detection," *Phys. Med. Biol.* **58**(20), 7159–7182 (2013).
30. J. G. Brankov, "Optimization of the internal noise models for channelized Hotelling observer," in *IEEE Int. Symp. Biomed. Imaging: From Nano to Macro* (2011).
31. J. G. Brankov, "Comparison of the internal noise models for channelized Hotelling observer," in *IEEE Nucl. Sci. Symp. Conf. Record* (2011).
32. L. Zhang et al., "A multi-slice model observer for medical image quality assessment," in *IEEE Int. Conf. Acoust., Speech, and Signal Process.*, pp. 1667–1671 (2015).
33. M. Han and J. Baek, "A performance comparison of anthropomorphic model observers for breast cone beam CT images: a single-slice and multislice study," *Med. Phys.* **46**(8), 3431–3441 (2019).
34. J. S. Kim et al., "A comparison of planar versus volumetric numerical observers for detection task performance in whole-body PET imaging," *IEEE Trans. Nucl. Sci.* **51**(1), 34–40 (2004).
35. H. Y. Liang et al., "Image browsing in slow medical liquid crystal displays," *Acad. Radiol.* **15**(3), 370–382 (2008).
36. C. Lartizien, P. E. Kinahan, and C. Comtat, "Volumetric model and human observer comparisons of tumor detection for whole-body positron emission tomography," *Acad. Radiol.* **11**(6), 637–648 (2004).
37. M. Chen et al., "Using the Hotelling observer on multislice and multiview simulated SPECT myocardial images," *IEEE Trans. Nucl. Sci.* **49**(3), 661–667 (2002).
38. H. C. Gifford et al., "A comparison of human and model observers in multislice LROC studies," *IEEE Trans. Med. Imaging* **24**(2), 160–169 (2005).
39. Y. Li et al., "A projection image database to investigate factors affecting image quality in weight-based dosing: application to pediatric renal SPECT," *Phys. Med. Biol.* **63**(14), 145004 (2018).
40. S. T. Treves et al., "Standardization of pediatric nuclear medicine administered radiopharmaceutical activities: the SNMMI/EANM joint working group," *Clin. Transl. Imaging* **4**(3), 203–209 (2016).
41. C. E. Metz, "Basic principles of ROC analysis," *Semin. Nucl. Med.* **8**(4), 283–298 (1978).
42. J. L. Brown et al., "A pediatric library of phantoms for renal imaging incorporating waist circumference, renal volume, and renal depth," in *Annu. Meeting Eur. Assoc. Nucl. Med.*, Düsseldorf, Germany (2018).
43. S. E. O'Reilly et al., "A risk index for pediatric patients undergoing diagnostic imaging with (99m)Tc-dimercaptosuccinic acid that accounts for body habitus," *Phys. Med. Biol.* **61**(6), 2319–2332 (2016).
44. Y. Li et al., "Development of a defect model for renal pediatric SPECT imaging research," in *IEEE Nucl. Sci. Symp. and Med. Imaging Conf.* (2015).
45. Y. Li et al., "Current pediatric administered activity guidelines for (99m)Tc-DMSA SPECT based on patient weight do not provide the same task-based image quality," *Med. Phys.* **46**(11), 4847–4856 (2019).
46. E. C. Frey et al., "A fast projector-backprojector pair modeling the asymmetric, spatially varying scatter response function for scatter compensation in SPECT imaging," *IEEE Trans. Nucl. Sci.* **40**(4), 1192–1197 (1993).
47. E. C. Frey and B. M. W. Tsui, "A new method for modeling the spatially-variant, object-dependent scatter response function in SPECT," in *IEEE Nucl. Sci. Conf. Rec.*, pp. 1082–1086 (1997).

48. Y. Du et al., “Combination of MCNP and SimSET for Monte Carlo simulation of SPECT with medium- and high-energy photons,” *IEEE Trans. Nucl. Sci.* **49**(3), 668–674 (2002).
49. Y. Du, B. M. W. Tsui, and E. C. Frey, “Model-based compensation for quantitative I-123 brain SPECT imaging,” *Phys. Med. Biol.* **51**(5), 1269–1282 (2006).
50. Y. Du, B. M. W. Tsui, and E. C. Frey, “Model-based crosstalk compensation for simultaneous Tc-99m/I-123 dual-isotope brain SPECT imaging,” *Med. Phys.* **34**(9), 3530–3543 (2007).
51. B. He et al., “A Monte Carlo and physical phantom evaluation of quantitative In-111SPECT,” *Phys. Med. Biol.* **50**(17), 4169–4185 (2005).
52. G. S. P. Mok et al., “Development and validation of a Monte Carlo simulation tool for multi-pinhole SPECT,” *Mol. Imaging Biol.* **12**(3), 295–304 (2010).
53. X. Rong et al., “Development and evaluation of an improved quantitative (90)Y bremsstrahlung SPECT method,” *Med. Phys.* **39**(5), 2346–2358 (2012).
54. N. Song et al., “Development and evaluation of a model-based downscatter compensation method for quantitative I-131 SPECT,” *Med. Phys.* **38**(6), 3193–3204 (2011).
55. N. Song et al., “EQPlanar: a maximum-likelihood method for accurate organ activity estimation from whole body planar projections,” *Phys. Med. Biol.* **56**(17), 5503–5524 (2011).
56. W. T. Wang et al., “Parameterization of Pb x-ray contamination in simultaneous Tl-201 and Tc-99m dual-isotope imaging,” *IEEE Trans. Nucl. Sci.* **49**(3), 680–692 (2002).
57. C. Bishop, “Mixture density networks,” Aston University, Neural Computing Research Group (1994).
58. K. C. L. Wong et al., “3D segmentation with exponential logarithmic loss for highly unbalanced object sizes,” *Lect. Notes Comput. Sci.* **11072**, 612–619 (2018).
59. O. Ronneberger, P. Fischer, and T. Brox, “U-Net: convolutional networks for biomedical image segmentation,” *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
60. D. P. Kingma and J. Ba, “Adam: a method for stochastic optimization,” in *3rd Int. Conf. Learn. Represent.*, San Diego, California (2015).
61. W. J. Chen, N. A. Petrick, and B. Sahiner, “Hypothesis testing in noninferiority and equivalence MRMC ROC studies,” *Acad. Radiol.* **19**(9), 1158–1165 (2012).

Ye Li is completing his PhD in electrical engineering at the Johns Hopkins University, under the supervision of Professor Eric Frey. His PhD research focused on task-based optimization of imaging systems. His research interests span the fields of computer-aided diagnosis, medical image computing, and medical imaging physics, especially the intersection of them. He also worked at IBM Watson Research, Brigham and Women’s Hospital, and Oak Ridge National Laboratory.

Junyu Chen received his BSc degree in computer engineering and electrical engineering from North Carolina State University in 2017 and his MSE degree in electrical and computer engineering from Johns Hopkins University in 2019. He is currently pursuing his PhD at Johns Hopkins University. He has held intern positions in PET reconstruction development at Canon Medical Research USA, Inc. His research interests include quantitative SPECT, medical image analysis, and deep learning.

Justin L. Brown is a diagnostic imaging medical physics resident at the University of Florida. He holds a PhD from the University of Florida where his research focused on computational human phantom development and Monte Carlo radiation dosimetry.

S. Ted Treves is a professor of radiology at Harvard Medical School and Brigham and Women’s Hospital and a fellow of the Society of Nuclear Medicine and Molecular Imaging. He is the founder and former director of the Division of Nuclear Medicine and Molecular Imaging and the Small Animal Imaging Laboratory at Boston Children’s Hospital. His research interests include physiologic imaging, diagnostic image optimization with radiation dose reduction. He trained at McGill and Yale.

Xinhua Cao is a computer specialist in the Division of Nuclear Medicine and Molecular Imaging at Boston Children’s Hospital. He is also an instructor of radiology at Harvard Medical School.

Frederic H. Fahey practiced nuclear medicine physics for 35 years and is a professor emeritus of radiology at Harvard Medical School. He served as the director of Nuclear Medicine/PET Physics at Boston Children's Hospital from 2003 to 2020. He received his DSc from the Harvard School of Public Health. He served as president of the Society of Nuclear Medicine and Molecular Imaging (SNMMI) in 2012 to 2013.

George Sgouros is a professor and director of the Radiological Physics Division in the Department of Radiology at the Johns Hopkins University School of Medicine. The focus of his research is on modeling and dosimetry of internally administered radionuclides with a particular emphasis on patient-specific dosimetry, alpha-particle dosimetry and mathematical modeling of radionuclide therapy. He is author on more than 200 peer-reviewed articles, as well as several book chapters and review articles.

Wesley E. Bolch is Distinguish Professor of Biomedical Engineering and Medical Physics at the University of Florida. He has been a member of the MIRD committee since 1993, a member of the NCRP since 2005, a member of ICRP since 2005, and a member of the US delegation of the United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) since 2015.

Eric C. Frey is a professor in the Radiological Physics Division of the Department of Radiology at Johns Hopkins University. He has over 30 years' experience in nuclear medicine physics, with contributions in scatter compensation and quantitative reconstruction in SPECT and task-based image quality assessment. He has been an advisor to 19 PhD students, has more than 150 journal publications, and is the cofounder and chief operating officer of Radiopharmaceutical Imaging and Dosimetry.