DeepAMO: A Multi-slice, Multi-view Anthropomorphic Model Observer for Visual Detection Tasks Performed on Volume Images

Ye Li, Junyu Chen, Justin L. Brown, S. Ted Treves, Xinhua Cao, Senior *Member, IEEE,*
Frederic Fahey, George Sgouros, Wesley E. Bolch and Eric C. Frey, Senior *Member, IEEE*

***Abstract*—We have developed 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. 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 mimic human observer's defect detection performance. The main contributions of this paper are threefold. First, we propose a hypothetical model of the decision process of a reader performing a detection task using a 3D volume. Second, we propose a projection-based defect confirmation network architecture to confirm defect present in two different slicing orientations. Third, we propose a novel calibration method that 'learns' the underlying distribution of observer ratings from the human observer rating data (thus modeling inter- or intra- observer variability) using a Mixture Density Network. We implemented and evaluated the DeepAMO in the context of ^{99m}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. 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. The results show that the DeepAMO has the potential to mimic human observer defect detection task performance in a clinically realistic diagnostic task.**

***Index Terms*—Deep learning, model observer, task-based image quality assessment.**

I. Introduction

OFTEN, the quality of a medical image is measured in terms of physical properties of the image, such as image contrast, spatial resolution, and noise level [1]. Alternatively, fidelity-based measures such as root mean squared error (RMSE), peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM), which evaluate image quality in terms of similarity of the image with respect to truth, have 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 will be performed with the images and thus may not be clinically relevant. To be clinically relevant, image quality should be assessed with respect to the task that will 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) can be too time-consuming, inconvenient and expensive. Thus, a great deal of effort 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 has been advocated by a number of 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, despite their rigorous theoretical foundation, task-based measures are often not used as an image quality metric by researchers in the medical imaging community. This is partly due to the fact that (1) model observers typically require much more complicated computations than fidelity-

Manuscript submitted April, 2020. This work was supported by National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under grant number R01-EB013558.

Y. Li, J. Chen and E. C. Frey are with Department of Electrical and Computer Engineering and the Radiological Physics Division, Department of Radiology and Radiological Science, Johns Hopkins University, Baltimore MD, 21218, USA (e-mail: yli192@jhu.edu).

G. Sgouros is with the Radiological Physics Division, Department of Radiology and Radiological Science, Johns Hopkins University, Baltimore MD, 21218, USA

J. Brown and W. E. Bolch are with J. Crayton Pruitt Family Department of Biomedical Engineering, University of Florida, Gainesville, FL 32611, USA

S. T. Treves, X. Cao, F. Fahey are with Department of Radiology, Boston Children's Hospital and Harvard Medical School, Boston, MA 02115, USA

based image quality metrics and (2) existing model observers are often not directly applicable to diagnostic tasks [18]. For example, as described below, common model observers are strictly valid only for signal-location-known (exactly and statistically) tasks. In addition, while 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 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 [20, 21], SKE-background known statistically (BKS) (e.g., lumpy backgrounds) tasks [22], and SKE-realistic anatomical backgrounds tasks [23-25]. However, in those tasks the observer is only asked to decide whether the defect is present or not at a specified location. A more clinically realistic detection task is the signal-known-statistically (SKS)/BKS task, where variability can be present in both the signal and background. Here, signal variability is present in the form of variations in signal/defect shape, size, orientation, or topology/texture or combinations of the above. Background variability can come from two sources: quantum noise and anatomical variability. Modeling the latter is important in order to model clinical task where patients can vary greatly in size, shape, uptake, etc. 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 [25, 26], reconstruction methods and phantom populations [27], imaging systems [28], compensation methods, and post-filter cutoff frequencies[29]. 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 (ROI) as a potential defect site [30]. However, for SKS tasks, training the scanning CHO can be computationally expensive as it requires computing covariance matrices at every ROI location, which makes the scanning observers impractical for use in large-scale studies of 3D image volumes.

In addition to the above limitations, existing model observers often predict rankings but not the absolute performance of human observers [31]. For imaging system optimization or comparison studies, this can be sufficient, but for other applications, such as selecting imaging time, administered activity 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 [31]. This requires a calibration process that is not generalizable across applications. Basically, 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 largely designed for analyzing 2D images. By contrast, many modern clinical tasks require interpretation of 3D images. This often involves reviewing sequences of 2D slices in 3 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 a 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 (i.e., the internal noise) using a Mixture Density Network. 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 mimic 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 the results section.

II. MATERIALS AND METHODS

Image quality in this work was measured in terms of performance on the task of detecting renal functional defects in ^{99m}Tc-DMSA SPECT. The images used were simulated based on an anthropomorphic digital phantom of 5-year-old (a common age in DMSA imaging). The phantom and simulation methods were previously 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 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 curve (AUC) of the ROC analysis served as a figure of merit for task performance.

A. Data Simulation

The projection data for this study were generated using the Advanced Laboratory for Radiation Dosimetry Studies (ALRADS) 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 hours post injection). 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 [44, 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 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 macro locations 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. From the phantom, 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 ^{99m}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 47 sets of patient data acquired at BCH. Parameters for the distributions can be found in [45]. Each individual organ projection was scaled by the product of administered activity (AA), acquisition duration, and scanner sensitivity. The projections were generated using an analytic projection code that modeled attenuation, 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 (LEUHR) collimator used routinely at BCH for pediatric DMSA studies. Each individual organ projection dataset was generated at 120 projection views over a 360° 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 cortex to medulla plus pelvis activity concentration ratios were randomly sampled.

We followed the clinical reconstruction protocol routinely used by BCH in their clinical practice. Projection images were reconstructed using the OS-EM iterative reconstruction algorithm with compensation for the geometric collimator-detector response and post-filtered with a Gaussian filter with a 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 between 0 to 255. A sample of BCH's volumetric display image is shown in Fig. 1.

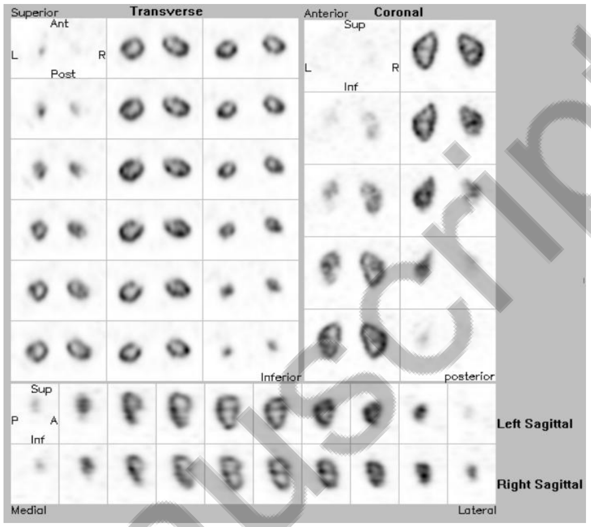

Fig. 1. A sample 48-slice image shown in the volumetric display format routinely used in clinical practice at the Boston Children's Hospital.

B. Proposed Model Observer: Theory

The DeepAMO is designed based on a hypothetical model of the image interpretation process of a human observer. We hypothesize that when an observer interprets an image, they would first scan over the slices to look for any suspicious abnormalities in single slices. If a defect is suspected to be present in one slice (of a particular orientation), they would then confirm that on adjacent slices. If positive, the observer would then confirm that defect is present using slices in the other two 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 a two-stage model observer to implement this decision-making process. Specifically, we propose to use a segmentation network as an abnormality search engine to perform the "first scan" over the slices. In that process, a nuclear medicine physician would normally consider adjacent slices when scanning for a defect in a particular slice. Thus, we propose to use a three-slice set (triad) of adjacent slices as the input to the segmentation network.

Within the network, the input image is first subdivided into multiple triads. Each triad is subsequently sent to a segmentation network to generate a segmentation mask, serving as a "first scan" over the slices in that triad. The segmentation masks along each orientation are then summed to form a summed segmentation mask in order to enhance the defect signal(s) that is/are present in that orientation. The summed segmentation masks are sent to a defect confirmation network to generate a low-dimensional feature vector. At last, a set feature vectors and the corresponding human observer rating values are sent to a Mixture Density Network to learn the

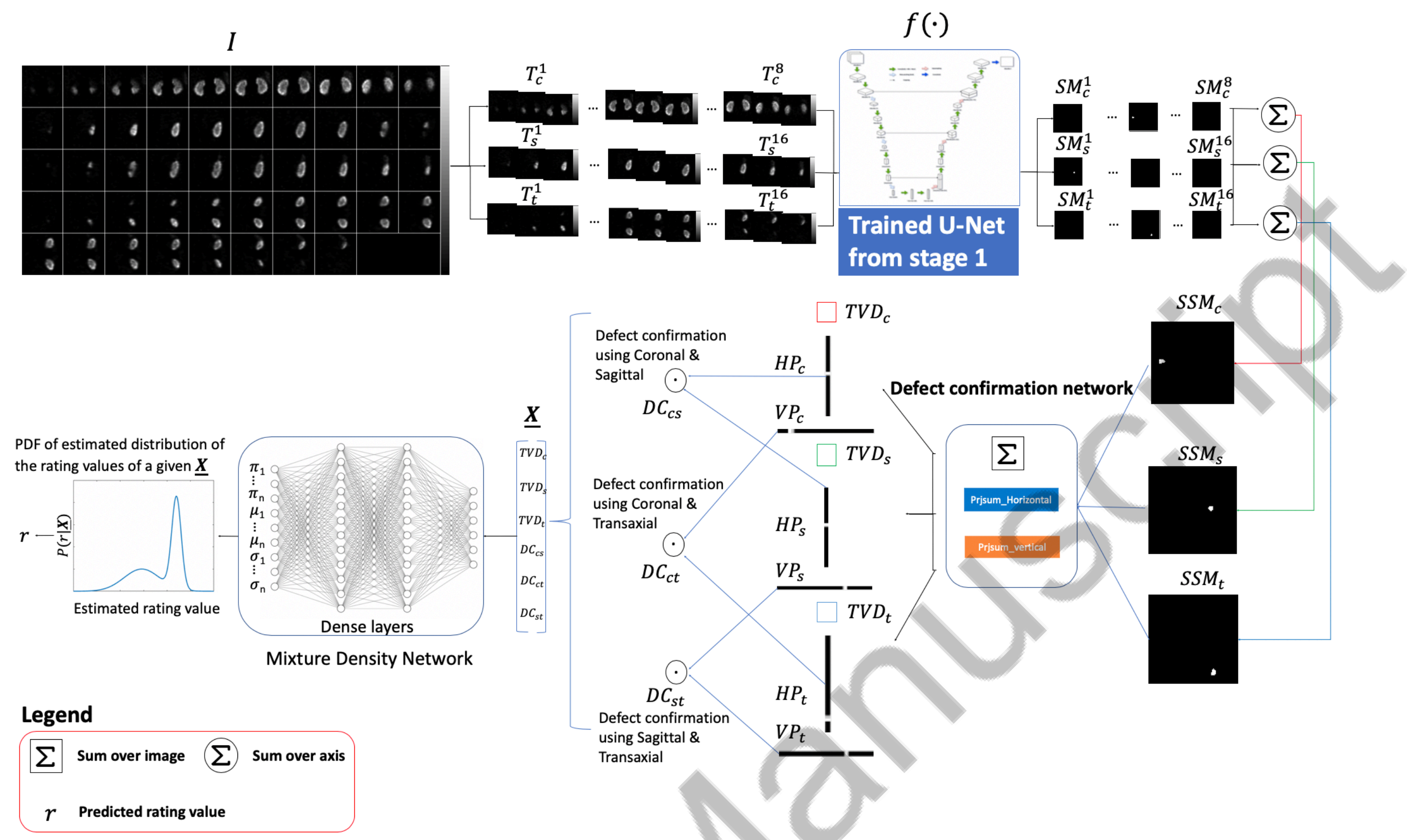

Fig. 2. A schematic of the proposed model observer: DeepAMO.

mapping between them, calibrating the DeepAMO to human performance.

C. 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 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

$$I(m,n,q_k^i) = S_k^i(m,n). \qquad (1)$$

In (1), q_k^i is the index number for the ith slice in the slicing direction $k \in (c,s,t)$ and m, n, and q are pixel indices for the x-, y-, and z-axis, respectively.

For each orientation, N-2 (N = the number of slices in each orientation) triads are generated: the first and last slices cannot act as the central slice for a triad.

$$T_k^j(m,n,q) = \{S_k^{i-1}(m,n), S_k^i(m,n), S_k^{i+1}(m,n)\},$$
$$i \in [0,N], j \in [1,N-1]. \qquad (2)$$

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.

$$SM_k^j(m,n) = f\left(T_k^j(m,n,q)\right). \qquad (3)$$

$$SSM_k(m,n) = \sum_{j=1}^{n_k} SM_k^j(m,n). \qquad (4)$$

with j the triad number and k the slicing direction. $T_k^j(m,n,q)$ and n_k represent the jth 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 defect confirmation 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:

$$HP_k(n) = \sum_{m=0}^{M-1} SSM_k(m,n), \text{ and} \qquad (5)$$
$$VP_k(m) = \sum_{n=0}^{N-1} SSM_k(m,n), \qquad (6)$$

with M and N being the number of pixels in x- and y-axis, 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 x 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

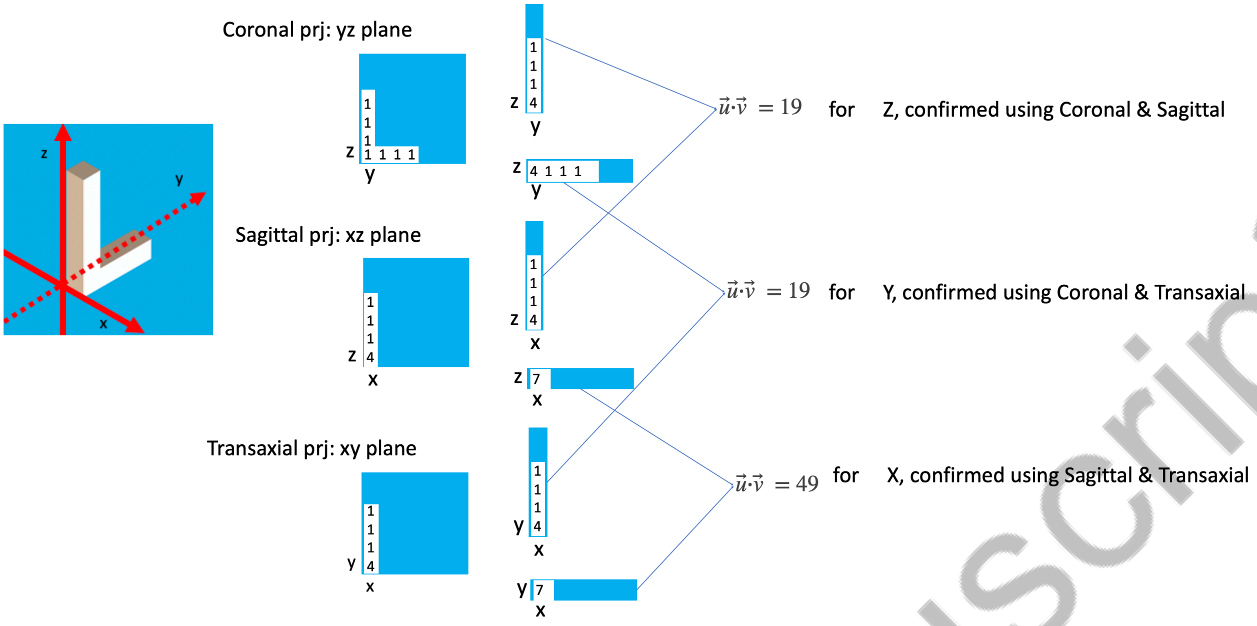
Coronal prj: yz plane
Sagittal prj: xz plane
Transaxial prj: xy plane
z
y
x
1 1 1 1
1 1 1 4
4 1 1 1
7
$\vec{u}\cdot\vec{v}$ = 19 for Z, confirmed using Coronal & Sagittal
$\vec{u}\cdot\vec{v}$ = 19 for Y, confirmed using Coronal & Transaxial
$\vec{u}\cdot\vec{v}$ = 49 for X, confirmed using Sagittal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

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

and dot product operations. This process yields 3 scalar values, representing the defect agreement along the x, y, z-axis, respectively. We named these 3 scalar values as defect confirmation (DC) scalars. They are derived from the HPs and VPs from different slicing directions as follows

$$DC_{cs} = HP_c(n) \cdot VP_s(m), \quad (7)$$

$$DC_{ct} = HP_t(n) \cdot VP_c(m), \text{ and} \quad (8)$$

$$DC_{st} = VP_t(m) \cdot VP_s(m). \quad (9)$$

The DC scalers 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

$$TVD_k = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} SSM_k(m,n). \quad (10)$$

The resulting 6-element concatenated feature vector is then sent to a Mixture Density Network (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.

D. Calibration to human observer via Mixture Density Network

For defect detection tasks, the observer performance is usually measured by the AUC of the ROC analysis 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 score, we propose to directly learn the distribution of the rating values of that observer. 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 of the human observers.

A mixture density network (MDN) was chosen for the task of turning 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, which 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 is $\underline{\boldsymbol{X}} = [x_{1\ldots}x_6]$ a 6-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 Gaussian mixture model as the conditional probability density function, which can be represented as a linear combination of kernel functions in the form

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

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

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

with π_i the output from the last dense layer, as shown in Fig. 3. The kernel functions, $\{\phi_i(r|\underline{\boldsymbol{X}})\}$, are in the form of Gaussian distributions

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

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 straight from the output of the last dense layer. Note that the $\{\pi_i(\underline{\boldsymbol{X}})\}$is a function of $\underline{\boldsymbol{X}}$. So, the $\{\pi_i(\underline{\boldsymbol{X}})\}$ can also be regarded as 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

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

In testing, a rating value is predicted by first non-uniformly sampling the mixing coefficients and then randomly sampling from the Gaussian distribution corresponding to that sampled mixing coefficient with its corresponding mean and standard deviation.

E. 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 with the goal being to sufficiently match the distribution and AUC values between the proposed model and human observer. We assumed the elements of the feature vectors and the rating values follow a (unimodal or multi-modal) Gaussian distribution.

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. 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.

$$\mu_{cs} = TVD_c \times TVD_s \tag{15}$$

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

$$\mu_{ct} = TVD_c \times TVD_t \tag{17}$$

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

$$\mu_{st} = TVD_s \times TVD_t \tag{19}$$

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

The rating values of each feature vector were sampled from an assumed multi- or uni- modal Gaussian distributions. The distribution parameters for these simulated rating values were derived qualitatively from distributions of rating values from human observer studies. The means and standard deviations of these assumed Gaussian distributions are shown in Table I. For each feature vector, we then sampled N rating values from the assumed distribution to simulate the appropriate level of inter- or intra- observer variability in the data. Specifically, in this work, we sampled 2 rating values for each feature vector. So, there were 15,000 (2,500 x 3 feature vector types x 2 repeated samples) feature vectors/rating values in total for the case that had 2,500 samples/feature vector type.

In the simulation experiment, we generated 3 types of feature vectors for each class (defect-present and defect-absent): definitely-present, equivocal, and definitely-absent, 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 3 large values for the 3 TVD_ks, modeling a high level of success of the segmentation network in detecting the defect in slices from all 3 orientations. The other two types (equivocal and definitely-absent, respectively) contained 2 and 1 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.

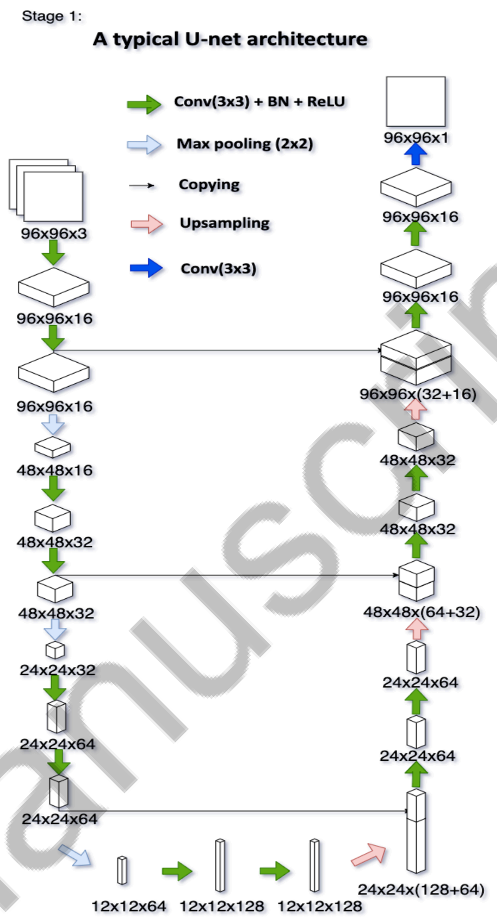

Fig. 4. Segmentation network architecture used in this study

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

Defect-present feature vector type	Definitely-yes		Not-sure		Definitely-no	
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	Definitely-yes		Not-sure		Definitely-no	
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

F. 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 segmentation masks. 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 is much smaller than the defect-absent ones, making this a highly imbalanced

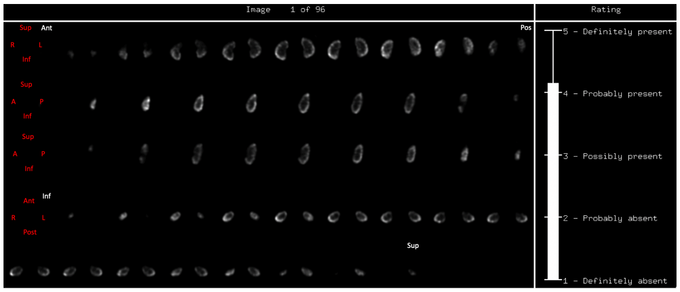

Fig. 5. A sample image of the GUI used in the human observer study

dataset. Thus, we adopted data augmentation on 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 degrees and flipping them and the original dataset upside down. The exponential logarithmic loss in [58] was adopted to emphasize segmentation of small structures with the best-performing weights ($\omega_{cross} = 0.2$ and $\omega_{Dice} = 0.8$).

For the segmentation network, we adopted a shallow version of the U-Net [59]. We used a shallow (depth) network due to the relatively small amount of training data available in this study; a deeper network might be needed when the number of signal and anatomical variations increases, as they will when applied to a larger population of phantoms. Gaussian noise with a standard deviation of 1.0 was added to the renormalized input image (ranges 0-255) to prevent overfitting. We searched for the optimal network capacity (depth) for the segmentation network. There was a tradeoff 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 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 dataset had 12,288 and 3,072 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 training took about 12 hours 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.

G. 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 to 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 Ph.D. students participated in the human observer study. A total of 384 of the composite images described in section A 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 II. 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 was collected from each observer.

Table II. 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	384

H. 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 is statically equivalent on a defect detection task. The null hypothesis and alternative hypothesis are expressed as follows:

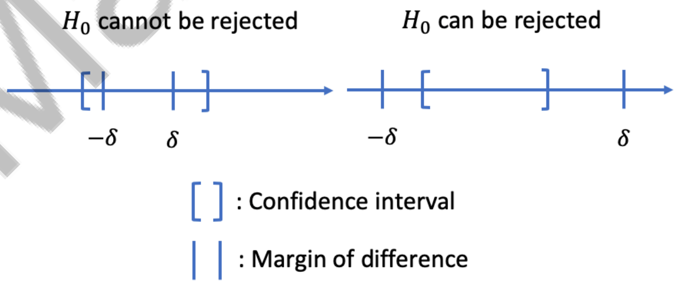

Fig. 6. A pictorial illustration of the rejectable and unrejectable case in equivalence hypothesis testing.

$$H_0: |AUC_{HO} - AUC_{MO}| = \delta \text{ and} \quad (21)$$
$$H_1: |AUC_{HO} - AUC_{MO}| < \delta.$$

where AUC_{HO} and AUC_{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 AUC_{HO} and AUC_{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 of the difference of the AUCs must lie within the interval defined by the margin of difference parameter, as described in [61] and illustrated in Fig. 6.

In order to calculate the confidence intervals for the difference in the AUCs (ΔAUC), 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 was 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 1st and 2nd 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.

III. RESULTS

A. *DeepAMO on Simulated Data*

The results (Fig. 7) show the degree of similarity between the histograms (distributions) of the simulated test data (simulated unseen data); the degree of similarity increases as the total number of samples increases, indicating that the MDN is capable of handling complex distributions of observer's rating values. This result agrees with the hypothesis that the MDN requires a minimum amount of training data in order to learn the underlying behavior of the observer on unseen data. Here, we assume that the underlying behavior of the observer is encoded in the distribution of that observer's rating values (training data).

The results also demonstrate that there is a tradeoff between ΔAUC and the total number of samples in the dataset. Bootstrapping was used to calculate the non-parametric confidence intervals on the ΔAUC. The ΔAUCs and 95% confidence intervals on the ΔAUCs are summarized in Table III. The results show that the 100, 500, and 2,500 samples/feature vector type cases had decreasing widths of the confidence intervals of ΔAUC, indicating that 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 bound of the 95% confidence interval on the ΔAUCs.

Table III. 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% C.I. on ΔAUC	C.I. 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

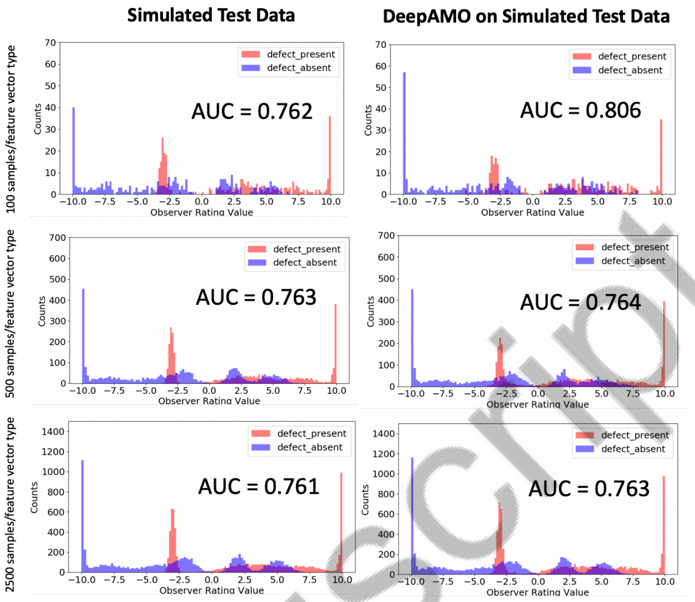

Fig. 7. Plots of histograms of the rating values of the simulated feature vectors (test data only) and predicted rating values on these data given by the DeepAMO. The plots show the class 0 and 1(defect present and absent, respectively) as well as the calculated AUC value.

B. *DeepAMO Test Results (stage II)*

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.

For stage II, The AUC values for the human observers and the corresponding DeepAMOs for the 5 × 2-fold cross validation experiment are summarized in Table IV. The mean and standard deviation of the ΔAUC were 0.03 and 0.0204, respectively. The 95% confidence interval 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) greater than 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. 8. 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.

Table IV. Summary of stage II training results

	1st fold		2nd fold				
Trial #	AUC HO	AUC Deep AMO	AUC HO	AUC Deep AMO	ΔAUC 1st fold	ΔAUC 2nd 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

IV. CONCLUSIONS

We propose a general framework for using deep convolution neural networks as an anthropomorphic model observer for the task of interpreting 3D image volumes and reproducing human

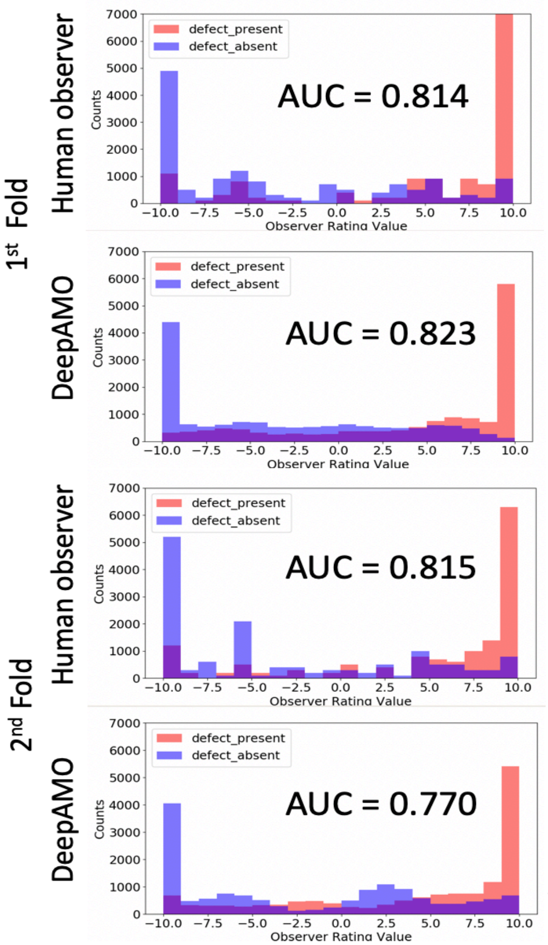

Fig. 8 Histograms of predicted rating values given by DeepAMO on unseen human observer data from the 3rd trial of the 5 x 2-fold cross validation experiment (other trials have similar patterns). Note that multiple predicted rating values were generated for each test image during testing of the DeepAMO 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.

observer performance. We applied this framework in the context of a renal functional defect detection task in nuclear medicine imaging using realistic simulated images. The results show that the proposed model observer and human observer's performance on unseen images can be 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. In addition, the results from the simulation experiment demonstrate that the proposed model observer is capable of precisely replicating a human observer's task performance on unseen data, as measured by the ΔAUC. 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.

References

[1] X. He, and S. Park, "Model observers in medical imaging research," *Theranostics*, vol. 3, no. 10, pp. 774-86, Oct 04, 2013.

[2] H. H. Barrett, J. L. Denny, R. F. Wagner *et al.*, "Objective Assessment of Image Quality .2. Fisher Information, Fourier Crosstalk, and Figures of Merit for Task-Performance," *Journal of the Optical Society of America a-Optics Image Science and Vision*, vol. 12, no. 5, pp. 834-852, May, 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," *Journal of the Optical Society of America a-Optics Image Science and Vision*, vol. 15, no. 6, pp. 1520-1535, Jun, 1998.

[4] H. H. Barrett, "Objective Assessment of Image Quality - Effects of Quantum Noise and Object Variability," *Journal of the Optical Society of America a-Optics Image Science and Vision*, vol. 7, no. 7, pp. 1266-1278, Jul, 1990.

[5] H. H. Barrett, K. J. Myers, N. Devaney *et al.*, "Objective assessment of image quality. IV. Application to adaptive optics," *Journal of the Optical Society of America a-Optics Image Science and Vision*, vol. 23, no. 12, pp. 3080-3105, Dec, 2006.

[6] H. H. Barrett, M. A. Kupinski, S. Mueller *et al.*, "Objective assessment of image quality VI: imaging in radiation therapy," *Phys Med Biol*, vol. 58, no. 22, pp. 8197-213, Nov 21, 2013.

[7] H. H. Barrett, and K. J. Myers, *Foundations of image science*, Hoboken, NJ: Wiley-Interscience, 2004.

[8] H. H. Barrett, K. J. Myers, C. Hoeschen *et al.*, "Task-based measures of image quality and their relation to radiation dose and patient risk," *Physics in Medicine and Biology*, vol. 60, no. 2, pp. R1-R75, Jan 21, 2015.

[9] K. J. Myers, and H. H. Barrett, "Addition of a Channel Mechanism to the Ideal-Observer Model," *Journal of the Optical Society of America a-Optics Image Science and Vision*, vol. 4, no. 12, pp. 2447-2457, Dec, 1987.

[10] M. B. Sachs, J. Nachmias, and J. G. Robson, "Spatial-frequency channels in human vision," *J Opt Soc Am*, vol. 61, no. 9, pp. 1176-86, Sep, 1971.

[11] S. Park, H. H. Barrett, E. Clarkson *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 Opt Image Sci Vis*, vol. 24, no. 12, pp. B136-50, Dec, 2007.

[12] A. E. Burgess, "Visual Perception Studies and Observer Models in Medical Imaging," *Seminars in Nuclear Medicine*, vol. 41, no. 6, pp. 419-436, Nov, 2011.

[13] J. L. Harris, "Resolving Power + Decision Theory," *Journal of the Optical Society of America*, vol. 54, no. 5, pp. 606-&, 1964.

[14] K. M. Hanson, and K. J. Myers, "Rayleigh Task-Performance as a Method to Evaluate Image-Reconstruction Algorithms," *Maximum Entropy and Bayesian Methods //*, vol. 43, pp. 303-312, 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," *Medical Imaging Vi : Image Processing*, vol. 1652, pp. 352-362, 1992.

[16] P. F. Judy, R. G. Swensson, and M. Szulc, "Lesion Detection and Signal-to-Noise Ratio in Ct Images," *Medical Physics*, vol. 8, no. 1, pp. 13-23, 1981.

[17] K. J. Myers, H. H. Barrett, M. C. Borgstrom *et al.*, "Effect of Noise Correlation on Detectability of Disk Signals in Medical Imaging," *Journal of the Optical Society of America a-Optics Image Science and Vision*, vol. 2, no. 10, pp. 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*, vol. 63, no. 3, pp. 1426-1434, Jun, 2016.

[19] H. H. Barrett, J. Yao, J. P. Rolland *et al.*, "Model observers for assessment of image quality," *Proc Natl Acad Sci U S A*, vol. 90, no. 21, pp. 9758-65, Nov 1, 1993.

[20] H. C. Gifford, M. A. King, D. J. de Vries *et al.*, "Channelized hotelling and human observer correlation for lesion detection in hepatic SPECT imaging," *J Nucl Med*, vol. 41, no. 3, pp. 514-21, Mar, 2000.

[21] K. J. Myers, and H. H. Barrett, "Addition of a channel mechanism to the ideal-observer model," *J Opt Soc Am A*, vol. 4, no. 12, pp. 2447-57, Dec, 1987.

[22] J. Yao, and H. H. Barrett, "Predicting Human-Performance by a Channelized Hotelling Observer Model," *Mathematical Methods in Medical Imaging*, vol. 1768, pp. 161-168, 1992.

[23] S. Sankaran, E. C. Frey, K. L. Gilland *et al.*, "Optimum compensation method and filter cutoff frequency in myocardial SPECT: a human observer study," *J Nucl Med*, vol. 43, no. 3, pp. 432-8, Mar, 2002.

[24] 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,* vol. 21, no. 9, pp. 1040-50, Sep, 2002.

[25] 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," *Physics in Medicine and Biology,* vol. 55, no. 17, pp. 4949-4961, Sep 7, 2010.

[26] M. P. Eckstein, C. K. Abbey, and J. S. Whiting, "Human vs. model observers in anatomic backgrounds," *Image Perception,* vol. 3340, pp. 16-26, 1998.

[27] S. D. Wollenweber, B. M. W. Tsui, D. S. Lalush *et al.*, "Comparison of hotelling observer models and human observers in defect detection from myocardial SPECT imaging," *Ieee Transactions on Nuclear Science,* vol. 46, no. 6, pp. 2098-2103, Dec, 1999.

[28] S. Park, E. Clarkson, M. A. Kupinski *et al.*, "Efficiency of the human observer detecting random signals in random backgrounds," *Journal of the Optical Society of America a-Optics Image Science and Vision,* vol. 22, no. 1, pp. 3-16, Jan, 2005.

[29] S. Sankaran, E. C. Frey, K. L. Gilland *et al.*, "Optimum compensation method and filter cutoff frequency in myocardial SPECT: A human observer study," *Journal of Nuclear Medicine,* vol. 43, no. 3, pp. 432-438, Mar, 2002.

[30] A. Sen, F. Kalantari, and H. C. Gifford, "Task Equivalence for Model and Human-Observer Comparisons in SPECT Localization Studies," *Ieee Transactions on Nuclear Science,* vol. 63, no. 3, pp. 1426-1434, Jun, 2016.

[31] Y. Zhang, B. T. Pham, and M. P. Eckstein, "Evaluation of internal noise methods for Hotelling observer models," *Medical Physics,* vol. 34, no. 8, pp. 3312-3322, Aug, 2007.

[32] L. Zhang, C. Cavaro-Menard, P. Le Callet *et al.*, "A Multi-Slice Model Observer for Medical Image Quality Assessment," *2015 Ieee International Conference on Acoustics, Speech, and Signal Processing (Icassp)*, 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," *Medical Physics,* vol. 46, no. 8, pp. 3431-3441, Aug, 2019.

[34] J. S. Kim, P. E. Kinahan, C. Lartizien *et al.*, "A comparison of planar versus volumetric numerical observers for detection task performance in whole-body PET imaging," *Ieee Transactions on Nuclear Science,* vol. 51, no. 1, pp. 34-40, Feb, 2004.

[35] H. Y. Liang, S. Park, B. D. Gallas *et al.*, "Image browsing in slow medical liquid crystal displays," *Academic Radiology,* vol. 15, no. 3, pp. 370-382, Mar, 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," *Academic Radiology,* vol. 11, no. 6, pp. 637-648, Jun, 2004.

[37] M. Chen, J. E. Bowsher, A. H. Baydush *et al.*, "Using the Hotelling observer on multislice and multiview simulated SPECT myocardial images," *Ieee Transactions on Nuclear Science,* vol. 49, no. 3, pp. 661-667, Jun, 2002.

[38] H. C. Gifford, M. A. King, P. H. Pretorius *et al.*, "A comparison of human and model observers in multislice LROC studies," *Ieee Transactions on Medical Imaging,* vol. 24, no. 2, pp. 160-169, Feb, 2005.

[39] Y. Li, S. O'Reilly, D. Plyku *et al.*, "A projection image database to investigate factors affecting image quality in weight-based dosing: application to pediatric renal SPECT," *Phys Med Biol,* vol. 63, no. 14, pp. 145004, Jul 9, 2018.

[40] S. T. Treves, M. J. Gelfand, A. Goodkind *et al.*, "Standardization of pediatric nuclear medicine administered radiopharmaceutical activities: the SNMMI/EANM Joint Working Group," *Clinical and Translational Imaging,* vol. 4, no. 3, pp. 203-209, Jun, 2016.

[41] C. E. Metz, "Basic principles of ROC analysis," *Semin Nucl Med,* vol. 8, no. 4, pp. 283-98, Oct, 1978.

[42] S.-S. B. Brown JL, Li Y, Frey EC, Treves ST, Fahey FH, Plyku D, Sgouros G, and Bolch WE, "A pediatric library of phantoms for renal imaging incorporating waist circumference, renal volume, and renal depth," in Annual Meeting of the European Association of Nuclear Medicine, Düsseldorf, Germany, 2018.

[43] S. E. O'Reilly, D. Plyku, G. Sgouros *et al.*, "A risk index for pediatric patients undergoing diagnostic imaging with (99m)Tc-dimercaptosuccinic acid that accounts for body habitus," *Phys Med Biol,* vol. 61, no. 6, pp. 2319-32, Mar 21, 2016.

[44] Y. Li, S. O'Reilly, D. Plyku *et al.*, "; Development of a Defect Model for Renal Pediatric SPECT Imaging Research," *2015 Ieee Nuclear Science Symposium and Medical Imaging Conference (Nss/Mic)*, 2015.

[45] Y. Li, S. O'Reilly, D. Plyku *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,* vol. 46, no. 11, pp. 4847-4856, Nov, 2019.

[46] E. C. Frey, Z. W. Ju, and B. M. W. Tsui, "A Fast Projector-Backprojector Pair Modeling the Asymmetric, Spatially Varying Scatter Response Function for Scatter Compensation in Spect Imaging," *Ieee Transactions on Nuclear Science,* vol. 40, no. 4, pp. 1192-1197, Aug, 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," *1996 Ieee Nuclear Science Symposium - Conference Record, Vols 1-3*, pp. 1082-1086, 1997.

[48] Y. Du, E. C. Frey, W. T. Wang *et al.*, "Combination of MCNP and SimSET for Monte Carlo simulation of SPECT with medium- and high-energy photons," *IEEE Transactions on Nuclear Science,* vol. 49, no. 3, pp. 668-674, JUN, 2002.

[49] Y. Du, B. M. W. Tsui, and E. C. Frey, "Model-based compensation for quantitative I-123 brain SPECT imaging," *Physics in Medicine and Biology,* vol. 51, no. 5, pp. 1269-1282, Mar, 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," *Medical Physics,* vol. 34, no. 9, pp. 3530-3543, Sep, 2007.

[51] B. He, Y. Du, X. Y. Song *et al.*, "A Monte Carlo and physical phantom evaluation of quantitative In-111SPECT," *Physics in Medicine and Biology,* vol. 50, no. 17, pp. 4169-4185, Sep, 2005.

[52] G. S. P. Mok, Y. Du, Y. C. Wang *et al.*, "Development and Validation of a Monte Carlo Simulation Tool for Multi-Pinhole SPECT," *Molecular Imaging and Biology,* vol. 12, no. 3, pp. 295-304, Jun, 2010.

[53] X. Rong, Y. Du, M. Ljungberg *et al.*, "Development and evaluation of an improved quantitative (90)Y bremsstrahlung SPECT method," *Med Phys,* vol. 39, no. 5, pp. 2346-58, May, 2012.

[54] N. Song, Y. Du, B. He *et al.*, "Development and evaluation of a model-based downscatter compensation method for quantitative I-131 SPECT," *Med Phys,* vol. 38, no. 6, pp. 3193-204, Jun, 2011.

[55] N. Song, B. He, R. L. Wahl *et al.*, "EQPlanar: a maximum-likelihood method for accurate organ activity estimation from whole body planar projections," *Phys Med Biol,* vol. 56, no. 17, pp. 5503-24, Sep 7, 2011.

[56] W. T. Wang, E. C. Frey, B. M. W. Tsui *et al.*, "Parameterization of Pb X-ray contamination in simultaneous Tl-201 and Tc-99m dual-isotope imaging," *IEEE Transactions on Nuclear Science,* vol. 49, no. 3, pp. 680-692, JUN, 2002.

[57] C. Bishop, "Mixture density networks," Aston University, Neural Computing Research Group, 1994.

[58] K. C. L. Wong, M. Moradi, H. Tang *et al.*, "3D Segmentation with Exponential Logarithmic Loss for Highly Unbalanced Object Sizes," *Medical Image Computing and Computer Assisted Intervention, Pt Iii,* vol. 11072, pp. 612-619, 2018.

[59] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Medical Image Computing and Computer-Assisted Intervention, Pt Iii,* vol. 9351, pp. 234-241, 2015.

[60] D. P. K. a. J. Ba, "Adam: A Method for Stochastic Optimization," *ArXiv e-prints,* vol. 1412.6980, 2014.

[61] W. J. Chen, N. A. Petrick, and B. Sahiner, "Hypothesis Testing in Noninferiority and Equivalence MRMC ROC Studies," *Academic Radiology,* vol. 19, no. 9, pp. 1158-1165, Sep, 2012.