A Convolutional Neural Network-Based Anthropomorphic Model Observer for Signal Detection in Breast CT Images Without Human-Labeled Data

BYEONGJOON KIM, MINAH HAN, AND JONGDUK BAEK
School of Integrated Technology, Yonsei Institute of Convergence Technology, Yonsei University, Incheon 21983, South Korea
Corresponding author: Jongduk Baek (jongdukbaek@yonsei.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT under Grants 2019R1A2C2084936 and 2020R1A4A1016619.

ABSTRACT Various imaging parameters in X-ray computed tomography (CT) should be examined and optimized by task-based assessment of human observer performance. Recently, convolutional neural networks (CNNs) have been introduced as anthropomorphic model observers. However, when human-labeled data are not available or limited, CNNs with an existing training strategy do not produce good performance agreement with human observers. The purpose of this study is to propose new training strategies for a CNN-based anthropomorphic model observer without human-labeled data for signal-known-exactly and background-known-statistically detection tasks. We acquired cone-beam CT projection data of breast background volume and reconstructed the projection data using the Feldkamp-Davis-Kress algorithm with 12 different imaging conditions including viewing image planes. Training data for the CNN were labeled utilizing conventional model observers. We employed an early stopping rule to reflect internal noise during the CNN training. To examine the CNN performance, we used three different training–testing schemes. The performance agreement between the human and model observers was measured via a Bland-Altman plot, the root-mean-squared error (RMSE), and the Pearson’s correlation coefficient ($r$) of their proportion correct values. Throughout the three different training–testing schemes, CNNs with the proposed training strategies yielded narrower limits of agreements (with a bias lower than 0.03) and higher scores in both RMSE and $r$ than the conventional model observers. This indicates the proposed training strategies enable the CNN-based anthropomorphic model observer to have good performance agreement with human observers and generalize better to different imaging conditions than conventional anthropomorphic model observers.

INDEX TERMS Anthropomorphic model observer, breast cone-beam CT, convolutional neural network, data labeling, early stopping, signal detection.

I. INTRODUCTION Optimizing the imaging parameters in X-ray computed tomography (CT) imaging is essential for improving diagnostic performance. Task-based assessment is considered to be a thorough image quality evaluation method because it directly quantifies the diagnostic accuracy (e.g., lesion detectability) with the reconstructed images [1]–[3]. To maximize the detection performance, various imaging parameters should be examined for the given task, and the performance then evaluated by a human observer. However, conducting human observer studies for numerous imaging parameters is time-consuming and expensive.

To overcome this limitation, anthropomorphic model observers, such as the non-prewhitening observer with an eye filter (NPWE) and the channelized Hotelling observer (CHO) [4]–[6], have been developed to represent the human observer by mimicking the frequency selective sensitivity of the human visual system with frequency filters [7]–[9] and reflecting the inefficiency of a human observer with internal...
To predict human observer performance, proper filter design, filter parameter, internal noise model, and internal noise level should be determined for each imaging task (e.g., signal shape and background noise structure), because model observers use limited features such as a signal template and covariance matrix of images [4]–[6], [11].

In recent studies, convolutional neural networks (CNNs) have been used in the implementation of model observers. The main advantage of using CNNs is that we do not have to manually find task-specific design of filter or internal noise model because CNNs can be trained in an end-to-end fashion. CNNs also have strong representational power to learn non-linear functions, which makes it possible to develop a single model that generalizes to different imaging conditions.

It has been shown that CNNs can approximate an ideal model observer [12], [13] and generally perform better than human observers [14] when they are trained with ground truth label (e.g., 1 for all signal-present images and 0 for all signal-absent images). Therefore, to implement CNN-based anthropomorphic model observers, it is necessary to reflect the human inefficiency in CNNs. Most of the recent approaches are to manually label every training image with a binary decision or confidence level of a human observer for signal presence or absence [15]–[19]. However, since CNNs require a large amount of labeled training data [15], this approach involves extensive human observer studies, contradicting to the original purpose of anthropomorphic model observers. Another recent approach incorporates the transfer learning and an internal noise component to reduce the required amount of labeled training data and calibrate model observer performance, respectively [20]. However, the explicit use of the internal noise component might restrict the generalization performance.

In this study, as an extension of our previous work [21], we propose new training strategies that can replace human-labeled data in implementing CNN-based anthropomorphic model observers. There are two key components in our training strategies: data labeling and early stopping. We obtain proportion correct ($P_c$) values of human observers from a relatively small amount of images and automatically label a large amount of training images with decision variables of the conventional model observer of which parameters are optimized to match $P_c$ values of the human and model observers. Therefore, the binary classification problem is converted into the regression problem where CNNs approximate the non-linear template that covers one or multiple linear observer templates optimal for each task. Accordingly, we also employ the new early stopping rule to better prevent overfitting of CNNs. Here, several different model observers were used for data labeling to investigate generalizability of the proposed training strategies. Observer performance was evaluated on four alternative forced choice (4-AFC) signal-known-exactly (SKE) and background-known-statistically (BKS) detection tasks in cone-beam CT (CBCT) images with anatomical backgrounds. The performance agreement between the human and model observers was measured using the difference in and correlation between the $P_c$ values of the human and model observers. Considering that model observers are generally trained and tested on different imaging conditions for image quality assessment, the generalization performance of CNN-based anthropomorphic model observers from one image plane to the other image plane was also evaluated and compared to that of the conventional model observers.

II. MATERIALS AND METHODS

A schematic illustration of the proposed training strategies for a CNN-based anthropomorphic model observer without human-labeled data is presented in Figure 1.

A. DATA GENERATION

For the breast background volume, we generated 512 × 512 × 512 voxels of Gaussian random noise volume with a voxel size of 0.16 × 0.16 × 0.16 mm³ and filtered the volume using a $1/f^{3/2}$ kernel, where $f$ is a radial frequency. Note that we used breast anatomy structures in mammograms characterized by $P(f) = K/f^3$ [22], where $K$ is a scaling factor. The central spherical volume with a 128-voxel diameter was extracted to prevent the wrap-around effect. We set the upper 50% and the lower 50% voxel values to the attenuation coefficients of glandular and adipose tissues, respectively, to generate a 50% volume glandular fraction (VGF) of the breast. The attenuation coefficients of glandular and adipose tissues are 0.233 and 0.194 cm⁻¹ at 50 keV, respectively [23].

For the breast background volume containing a signal, we inserted a 2-mm-diameter sphere near the center of the generated breast background volume. We set the voxel values in the region of the sphere to the attenuation coefficient of breast mass (0.238 cm⁻¹ at 50 keV [23]).

We used Siddon’s algorithm [24] to acquire 200 CBCT projection data slices of the generated breast background volume with and without the signal, and then, added log-normalized uniform Poisson noise of 6914 photons/detector pixel. Note that we used a mean glandular dose equivalent to two-view mammography (i.e., 6.4 mGy for a 14-cm 50% VGF of the breast) and estimated photon fluence according to the study of Boone et al. [25]. The simulation parameters are summarized in Table 1.

We reconstructed the projection data using the Feldkamp-Davis-Kress algorithm [26], where Hanning, Shepp-Logan, and Ram-Lak weighted ramp filters were used as a reconstruction filter to generate different background noise structures.

We filtered projection data with and without Fourier interpolation and performed voxel-driven back-projection with linear interpolation. For the observer study, we used the central transverse (i.e., x-y plane) and longitudinal (i.e., x-z plane) images. Consequently, 12 types of background noise structures were generated, as summarized in Table 2.

For the datasets, we generated 4,000 pairs of signal-present and signal-absent images for each noise structure, leading to 48,000 image pairs in total. The central 129 × 129 pixels...
FIGURE 1. Schematic diagram of the proposed approach to CNN training. The dataset consists of signal-present and signal-absent CT images generated with different imaging conditions. To automatically label the dataset with pseudo human labels, the conventional model observer is optimized to match its performance to that of human observers for each imaging condition. The CNN-based anthropomorphic model observer is then trained to minimize the difference between its output and decision variable of the conventional model observer. The CNN has two convolution layers with rectified linear units (ReLUs) and one fully connected layer.

TABLE 1. The CBCT simulation parameters.

| Parameter                           | Value          |
|-------------------------------------|----------------|
| Source to iso-center distance       | 460 mm         |
| Source to detector distance         | 880 mm         |
| Detector pixel size                 | 0.388×0.388 mm²|
| Number of projection data           | 200            |
| Reconstructed voxel size            | 0.1014×0.1014 mm³|

TABLE 2. Background noise structures from different imaging conditions.

| Noise structure | Reconstruction filter | Interpolation method | Image plane |
|-----------------|-----------------------|----------------------|-------------|
| N1              | Hannig                |                      |             |
| N2              | Shepp-Logan          | Linear               | Transverse  |
| N3              | Ram-Lak              |                      |             |
| N4              | Hannig                |                      |             |
| N5              | Shepp-Logan          | Fourier              |             |
| N6              | Ram-Lak              |                      |             |
| N7              | Hannig                |                      |             |
| N8              | Shepp-Logan          | Linear               | Longitudinal|
| N9              | Ram-Lak              |                      |             |
| N10             | Hannig                |                      |             |
| N11             | Shepp-Logan          | Fourier              |             |
| N12             | Ram-Lak              |                      |             |

of each images were then extracted. Signal-present example images are shown in Figure 2.

B. DETECTION TASKS

We conducted 4-AFC detection tasks under SKE/BKS conditions, where a 2-mm signal was located near the center of the breast background images generated independently in each detection trial. Hypotheses for signal-absent (i.e., \( H_0 \)) and signal-present (i.e., \( H_1 \)) images are given by

\[ H_0 : g = f_b + f_n, \]  
\[ H_1 : g = f_s + f_n, \] 

where \( g \) is the observed image, \( f_b \) is the breast background, \( f_s \) is the breast background with the signal, and \( f_n \) is the reconstructed CT noise.

C. THE HUMAN OBSERVER STUDY

Seven human observers performed the 12 detection tasks with different background noise structures (Table 2). For each detection task, the training session consisted of 30 trials, while the testing session consisted of 100 trials [2], [27]. In each trial, four images were displayed on a Nio 3MP LED monitor (Barco, Kortrijk, Belgium) with 2048×1536 pixel resolution and the human observers were asked to choose the signal-present image among one signal-present and three signal-absent images. The human observer performance is quantified by

\[ P_c = \frac{1}{N_t} \sum_{j=1}^{N_t} o_j, \] 

where \( N_t \) is the number of trials in testing session and \( o_j \) is the score of the \( j-th \) trial (1 for correct answer and 0 for incorrect answer). We estimated the variance of \( P_c \) by bootstrapping the scores 1,000 times.
D. DATA LABELING USING THE ANTHROPOMORPHIC MODEL OBSERVERS

To label the data for network training, we propose new labeling strategies (light green box in Fig. 1) using the NPWE and dense difference-of-Gaussian (D-DOG) CHO for the generated signal-present and signal-absent images.

1) NPWE

The eye filter of NPWE is defined as

\[ E(f) = f^{1.3} \exp(-cf^2), \]

where \( c \) is an eye filter parameter. The NPWE template is defined as

\[ w = F^{-1}\{E \cdot F\{\Delta g\}\}, \]

where \( F\{\cdot\} \) and \( F^{-1}\{\cdot\} \) indicate discrete Fourier transform (DFT) and inverse DFT operators, respectively, and \( \Delta g = \mathbb{E}[g \cdot H_1] - \mathbb{E}[g \cdot H_0] \) (note that \( \mathbb{E}[\cdot] \) is an expectation operator). We estimated the template \( w \) using 400 image pairs. The decision variable for the test image is computed by

\[ t = w^T F^{-1}\{E \cdot F\{g\}\} + \epsilon, \]

where \( \epsilon \) is the internal noise term sampled from normal distribution \( N(0, b^2) \); the internal noise level is adjusted by the magnitude of the \( a \).

NPWE does not predict the exact human observer performance for all noise structures with the same value of \( a \) or \( c \). Therefore, we adjusted the values of \( a \) or \( c \) for each noise structure for data labeling. When the value of \( a \) was adjusted with \( c = 2 \), we denoted the model observer as NPWE4i, and when the value of \( c \) was adjusted with \( a = 0 \), we denoted the model observer as NPWEf. Note that \( c = 2 \) yielded the peak value of the eye filter at 4 cyc/deg for a 40-cm viewing distance, which is the most sensitive frequency of the human visual system [6], [7].

We used two new labeling strategies:

1. We adjusted the value of \( a \) to minimize the difference in \( P_r \) between NPWE4i and the human observers for each background noise structure and then assigned the decision variable of NPWE4i as the label value.

2. We adjusted the value of \( c \) to minimize the difference in \( P_r \) between NPWEf and the human observers for each background noise structure and then assigned the decision variable of NPWEf as the label value.

The parameter searching interval was 0.00005 within the range of [0, 0.003] for \( a \) and 0.01 within the range of [0, 1] for \( c \).

2) D-DOG CHO

Image \( g \) is transformed to the vector \( v \) for calculation of CHO performance:

\[ v = Tg, \]

where \( T \) is the channel matrix. The \( k - th \) D-DOG channel is defined as

\[ C_k(\rho) = \exp\left[-\frac{1}{2}(\frac{\rho}{Q\sigma_k})^2\right] - \exp\left[-\frac{1}{2}(\frac{\rho}{\sigma_k})^2\right], \]

where \( \rho \) is the radial frequency, \( Q \) is a multiplicative factor, and \( \sigma_k \) is the standard deviation \( (i.e., \sigma_k = \sigma_0 a^k) \). Ten channels \( (k = 1 \ldots 10) \) were used with parameters of \( Q = 1.67, \sigma_0 = 0.005, \) and \( a = 1.4 [8] \). Each row of \( T \) is the sampled values from the inverse Fourier transform of (8). The CHO template is computed by

\[ w_{CHO} = K_{v^{-1}} \Delta v, \]

where \( \Delta v = \mathbb{E}[v_1] - \mathbb{E}[v_0] \) and \( K_v = \mathbb{E}[(v_1 - \mathbb{E}[v_1])(v_1 - \mathbb{E}[v_1])^2]/2 + \mathbb{E}[(v_0 - \mathbb{E}[v_0])(v_0 - \mathbb{E}[v_0])^2]/2 \), and \( v_1 \) and \( v_0 \) are \( v \) from signal-present and signal-absent images, respectively. We estimated the template \( w_{CHO} \) using 400 image pairs. The decision variable for the test image is computed by

\[ t = w_{CHO}^T v + \epsilon, \]

where \( \epsilon \) is the internal noise term sampled from normal distribution \( N(0, b^2) \); the internal noise level is adjusted by the magnitude of the \( b \).

We adjusted the value of \( b \) to minimize the difference in \( P_r \) between D-DOG CHO and the human observers for each background noise structure and then assigned the decision variable of D-DOG CHO as the label value. The parameter searching interval was 0.1 with in the range of [0,5] for \( b \).

E. THE CNN-BASED ANTHROPOMORPHIC MODEL OBSERVER

NPWE and DDOG-CHO can be interpreted as one convolution layer (i.e., the eye filter or channel matrix) followed by one fully connected layer (i.e., the observer template) without any non-linear activation function. Inspired by this interpretation, we used a three-layer CNN (light blue box in Fig. 1) to learn a non-linear model observer with its additional convolution layer and non-linear activation function. The three-layer CNN comprises two convolution layers, one fully connected layer, and rectified linear units after every convolution layer. The first convolution layer extracts feature maps from the input image, and the second convolution layer further transforms the feature maps into a higher dimensional feature space to reflect several different eye filters or DDOG channels. The feature maps are then combined into a scalar decision variable in the fully connected layer. The convolution layers had 13 × 13 filters with a stride of 2, and the number of filters was set as two in the first layer and doubled when the feature map size was halved to preserve the time complexity per layer [28]. The size of the filters was chosen to match the receptive field seen at the last convolution layer to the signal size (i.e., 37 × 37).

The true objective function to be minimized was the error in \( P_r \) between the human and CNN-based model observers for a 4-AFC detection task. However, it involves the 0-1 loss function which is intractable [29]. Therefore, we defined our
problem as the least squares regression problem and used the mean-squared error in the decision variable as a surrogate loss function:

\[ L = \frac{1}{N} \sum_{i=1}^{N} (\hat{t}^{(i)} - t^{(i)})^2, \]

where \( N \) is the number of mini-batches and \( \hat{t}^{(i)} \) and \( t^{(i)} \) are the label and the output of the model from the \( i \)-th image, respectively.

F. TRAINING AND TESTING THE MODEL OBSERVERS

To examine the CNN performance, we used three different training–testing schemes.

1) THE FIRST TRAINING–TESTING SCHEME

We trained a CNN on images with one noise structure and tested it on unseen images with the same noise structure. Therefore, we had 12 different CNNs optimized for each noise structure. Note that this is the most generous condition where even a linear model observer can match its performance with that of human observers. This is also the condition given to the linear model observers when labeling the training datasets of CNN.

2) THE SECOND TRAINING–TESTING SCHEME

We trained a CNN on images in one image plane (i.e., six different noise structures) and tested it on unseen images in the same image plane. This led to two different CNNs optimized for the transverse plane and longitudinal plane, respectively. Through this training–testing scheme, we examined CNN’s ability to handle multiple tasks with a single model. For comparison, the performance of NPWE4i, NPWEf, and DDOG-CHO was evaluated by using fixed optimal values of \( a, c, \) and \( b \), respectively, which were searched to minimize average difference in \( P_c \) between the human and model observers over all noise structures in the given image plane.

3) THE THIRD TRAINING–TESTING SCHEME

We trained a CNN on one image plane and tested it on the other image plane. This training–testing scheme examines the generalization performance of the CNN from one image plane to the other image plane, thus better representing the practical use of the proposed CNN-based anthropomorphic model observer. For comparison, the performance of NPWE4i, NPWEf, and DDOG-CHO was evaluated by using fixed values of \( a, c, \) and \( b \), respectively, optimized for the image plane other than the given image plane.

4) SPLITTING DATASETS

The datasets illustrated in II-A were divided into training, validation, and testing sets in the ratio 2:1:1. Therefore, the training, validation, and testing sets for each CNN were composed of 2,000, 1,000, and 1,000 image pairs for the first training–testing scheme and 12,000, 6,000, and 6,000 image pairs for the second and third training–testing schemes.

5) TRAINING DETAILS

In all training–testing schemes, CNN was initialized using the Glorot Uniform method [30] and optimized by the Adam algorithm [31] with a learning rate of 0.001, \( \beta_1 \) of 0.9, \( \beta_2 \) of 0.999, and a batch size of 64. The maximum number of training epochs was 300 (18,900 iterations) for the first training–testing scheme and 200 (75,000 iterations) for the second and third training–testing schemes. The CNNs trained with the NPWE4i, NPWEf, and DDOG-CHO decision variables are denoted as CNN-NPWE4i, CNN-NPWEf, and CNN-DDOG-CHOi, respectively. All CNN models were implemented with Keras [32] library in Python.

G. EARLY STOPPING RULE

As a form of regularization, we propose an early stopping rule that directly monitors the RMSE in \( P_c \) on the validation set. The RMSE in \( P_c \) is computed by

\[ \text{RMSE} = \sqrt{\frac{1}{K} \sum_{k=1}^{K} (\hat{P}_c^{(k)} - P_c^{(k)})^2}. \]

where \( K \) is the number of different noise structures in the training set, and \( P_c^{(k)} \) and \( \hat{P}_c^{(k)} \) are \( P_c \) values of human and CNN observers, respectively, on the \( k \)-th noise structure. Here, we used the same \( P_c^{(k)} \) values used to label the data. For simplicity, we implemented this rule by measuring validation RMSE in \( P_c \) after every epoch and saving the model only if its validation RMSE was lower than the previous minimum value.

H. FIGURES OF MERIT

The model observers performed the 4-AFC detection tasks where they chose the image with the highest decision variable as a signal-present image among one signal-present and three signal-absent images in each trial. We quantified the model observer performance using (3) with 10,000 trials for each noise structure. We estimated the mean and variance of \( P_c \) by bootstrapping the scores 1,000 times. To evaluate the agreement between the human and model observers, We measured RMSE and the Pearson’s correlation coefficient (\( r \)) of the \( P_c \) values. The bias and the limits of agreement in \( P_c \) were also calculated through Bland-Altman plots.

III. RESULTS

A. THE FIRST TRAINING–TESTING SCHEME

Table 3 summarizes \( a, c, \) and \( b \) values used to label the data, which can match human observer and model observer \( P_c \) exactly for each noise structure.

Figure 3 compares \( P_c \) of human and model observers when the model observers were optimized and evaluated on the same noise structure. Note that the training datasets of CNN were labeled by NPWE4i, NPWEf, and DDOG-CHO in this training–testing scheme. It can be observed that the \( P_c \) values of NPWE4i, NPWEf, and DDOG-CHO are within the 95% confidence intervals of those of human observers. This provides the ground for using NPWE4i, NPWEf,
TABLE 3. The parameter values of NPWE4i, NPWEf, and DDOG-CHOi that were adjusted for labeling the data.

| Model       | Parameter | N1  | N2  | N3  | N4  | N5  | N6  | N7  | N8  | N9  | N10 | N11 | N12 |
|-------------|-----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| NPWE4i      | $a \times 100$ | 0.115 | 0.150 | 0.165 | 0.090 | 0.140 | 0.185 | 0.080 | 0.100 | 0.130 | 0.045 | 0.150 | 0.200 |
| NPWEf       | $c$       | 0.30 | 0.52 | 0.54 | 0.61 | 0.70 | 0.63 | 0.25 | 0.45 | 0.45 | 0.54 | 0.56 | 0.60 |
| DDOG-CHOi   | $b$       | 4.3  | 4.7  | 5.1  | 3.6  | 4.5  | 5.4  | 3.8  | 3.6  | 4.0  | 2.9  | 4.0  | 4.5  |

FIGURE 3. Averaged $P_c$ values of the human and model observers with 95% confidence intervals. The model observers were evaluated with the first training–testing scheme. The decision variables of NPWE4i, NPWEf, and DDOG-CHOi in this scheme were used as data labels.

TABLE 4. The parameter values of NPWE4i, NPWEf, and DDOG-CHOi that minimize the difference in $P_c$ between the human and model observers for a given image plane.

| Model       | Parameter | Transverse | Longitudinal |
|-------------|-----------|------------|--------------|
| NPWE4i      | $a \times 100$ | 0.135 | 0.105 |
| NPWEf       | $c$       | 0.56 | 0.52 |
| DDOG-CHOi   | $b$       | 4.4  | 3.7  |

and DDOG-CHOi in place of human observers to label the data. It is also evident that CNN can learn an anthropomorphic model observer with our training strategies, producing $P_c$ values within the 95% confidence intervals of human $P_c$ values.

B. THE SECOND TRAINING–TESTING SCHEME

Figure 4 compares the performances of the human and model observers when the model observers were optimized and evaluated on the same image planes. The NPWE4i, NPWEf, and DDOG-CHOi were evaluated using the optimal parameter values for each image plane as summarized in Table 6.

Figure 5 shows Bland-Altman plots of the $P_c$ values between the human and model observers. The mean difference of each model observer was between −0.01 and 0.02. The limits of agreement were reduced by CNN from [−0.1, 0.12] to [−0.05, 0.04] for NPWE4i, from [−0.05, 0.08] to [−0.05, 0.03] for NPWEf, and from [−0.08, 0.07] to [−0.03, 0.06] for DDOG-CHOi. This allows more strict clinical agreement limit for CNN-based anthropomorphic model observers to be used interchangeably in place of human observers.

Table 5 summarizes the RMSE and $r$ in $P_c$ between the human and model observers for the transverse, longitudinal, and both image planes. In terms of RMSE, the CNNs had significantly higher scores than the conventional model observers. The CNNs also achieved generally higher $r$ than conventional model observers. This suggests that the proposed training strategies are effective for reflecting task-dependent human inefficiency in one model, thus leading to better prediction of human observer performance for various tasks on an absolute scale. It is also evident that the model observer type used for data labeling had little impact on the performance of the CNN because every linear model observer could be optimized for each task, as already shown in Figure 3, to properly impose human inefficiency.

C. THE THIRD TRAINING–TESTING SCHEME

Figure 6 compares the performances of the human and model observers when the model observers were optimized on one image plane and evaluated on the other image plane. For example, NPWE4i was evaluated on the transverse plane using the $a$ value optimal for the longitudinal plane. It can be observed that the performance of all the model observers was less similar to that of the human observers compared to the second training–testing scheme. Among the model observers, the performance of CNNs were still in better agreement with that of the human observers.

Figure 7 shows Bland-Altman plots. The mean differences of model observers were between −0.02 and 0.01. The limits of agreement were reduced by CNN from [−0.13, 0.15] to [−0.07, 0.07] for NPWE4i, from [−0.06, 0.09] to [−0.07, 0.03] for NPWEf, and from [−0.1, 0.1] to [−0.08, 0.09] for DDOG-CHOi.

Table 6 summarizes the RMSE and $r$ in $P_c$ between the human and model observers. In terms of RMSE and $r$ values, CNNs did not always outperform the best among...
FIGURE 4. Averaged $P_c$ values of the human and model observers with 95% confidence intervals. The model observers were evaluated with the second training–testing scheme.

FIGURE 5. Bland-Altman plot of the difference of $P_c$ values in Figure 4 between the human and model observers across all 12 noise structures. The solid line denotes the mean difference and the dashed lines denote the limits of agreement.

TABLE 5. RMSE and $r$ of the $P_c$ values between the human and model observers presented in Figure 4.

| Observer            | Transverse      | Longitudinal    | All         |
|---------------------|-----------------|-----------------|-------------|
|                     | RMSE            | $r$             | p-value     | RMSE          | $r$             | p-value     | RMSE          | $r$             | p-value     |
| NPWE4i              | 0.0464          | 0.1011          | 0.0603      | 0.9760       | 0.0016        | 0.0538      | 0.4471        | 0.1450        |
| NPWEi               | 0.0400          | 0.7715          | 0.0724      | 0.9729       | 0.0011        | 0.0361      | 0.9236        | < 0.0001      |
| DDOG-CHOi           | 0.0378          | 0.4382          | 0.3848      | 0.0358       | 0.9760        | 0.0009      | 0.0368        | 0.7849        | 0.0025      |
| CNN-NPWE4i          | **0.0228**      | **0.8108**      | 0.0503      | **0.0215**   | 0.9684        | 0.0015      | **0.0221**    | **0.9351**    | < 0.0001    |
| CNN-NPWEi           | 0.0276          | 0.7907          | 0.0611      | 0.0205       | 0.9695        | 0.0014      | 0.0243        | **0.9428**    | < 0.0001    |
| CNN-DDOG-CHOi       | 0.0270          | 0.7571          | 0.0813      | 0.0252       | 0.9703        | 0.0013      | 0.0261        | 0.9387        | < 0.0001    |

conventional model observers. Still, each CNN had better scores than the corresponding model observer used for data labeling. For example, CNN-NPWE4i achieved the better generalization performance than NPWE4i. This indicates that CNNs can learn more general non-linear templates from several linear observer templates optimized for each task.
In this study, we proposed new training strategies for CNN-based anthropomorphic model observers for SKE/BKS detection tasks in CBCT images with breast anatomical backgrounds. These new training strategies require the $P_c$ of human observers only on a portion of full datasets and utilize a conventional model observer for labeling the rest of the datasets. Therefore, it can effectively reduce the cost.
of data labeling. We showed that CNNs with our proposed training strategies generalize well from images with one specific noise structure to unseen images with the same noise structure (i.e., the first training–testing scheme), from several noise structures to the same noise structures (i.e., the second training–testing scheme), and from one image plane to the other image plane (i.e., the third training–testing scheme). We also showed that different conventional model observers (i.e., NPWEi, NPWEf, and DDOG-CHOi) can be used for data labeling to achieve the better generalization performance than they had. Note that the proposed methods were validated using the simulated data from our previously published work [6]. It will be an interesting future work to extend our work to real images with clinically relevant targets.

In this study, we focused on the training strategies rather than the design of the network architecture. We minimized the effort required to optimize the hyperparameters (e.g., the number of convolution layers and kernels) and used the same hyperparameters across all of the experiments. Although the simple three-layer CNN showed promising results for SKE/BKS detection tasks, we believe the more complex tasks (e.g., signal-known-statistically (SKS) detection tasks with unknown signal locations) will require a more complex and heavier neural network. Therefore, it will be an interesting future research topic to investigate appropriate network architectures and the effectiveness of the proposed training strategies for those detection tasks.

Although not presented in this paper, we also implemented a three-layer capsule network (CapsNet) with the routing algorithm [33] and compared its performance with our three-layer CNN. CapsNet was composed of convolution, two-dimensional primary capsule, and four-dimensional digit capsule layers. We observed no significant difference in performance between CNN and CapsNet. We conjecture that the possible information loss in strided convolution was not critical in our study because the shape of signals was relatively simple compared to the other objects composed of many entities. However, we believe that CapsNet can be more effective for the clinical tasks where the relationship between different parts of image is important.

For data labels, it is possible to use one-hot encoded answers or softmax outputs for 4-AFC tasks. However, one-hot encoding and softmax processes suffer from information loss because they focus on relative values and ignore each individual value [34]. Therefore, we trained CNNs directly on the decision variable to better learn the knowledge from conventional model observers.

To prevent overfitting of CNN, we used the early stopping method based on RMSE in $P_e$. We empirically found that this is effective compared to the conventional method based on loss. In our preliminary study, CNN constantly produced higher $P_e$ values at the minimum validation loss and lower $P_e$ values at the minimum training loss than those of human observers. The early stopping method is also compatible with other regularization methods such as weight decay, dropout, and reducing the number of neurons.

Therefore, generalization performance would be further improved by using those methods together with our method.

Our training approach is related to self-supervised learning in that training datasets are automatically labeled and thus networks can be trained without human-labeled data. Most of the self-supervised learning approaches aim to learn useful visual representations through so-called pretext tasks where labels can be automatically extracted from given images [35]–[39]. The learned representations are then used in transfer learning or knowledge transfer to perform a target task. However, these approaches are likely to lose a certain amount of learned knowledge due to the gap between pretext and target tasks. In contrast, our approach has no such problem because it applies self-supervised learning directly to the target task.

We would also like to discuss the computational cost of model observers in a testing phase. The average execution times per 2,000 images were 0.05, 0.05, 0.32, and 4.30 seconds for NPWEf, NPWEi, DDOG-CHOi, and CNN in a CPU mode, respectively. Although CNN took more time to process than conventional model observers, the execution time was still negligible compared to the image acquisition and reconstruction time in breast CT.

In this study, we predicted the human observer performance in single-slice images. However, in cone beam CT images, multiple slices are used for diagnosis. As described in previous work [40], detectability was improved when using multislice images. Therefore, we plan to extend this work for detection task in multi-slice images by training a 3-D CNN with transfer learning from a 2-D CNN.

ACKNOWLEDGMENT
(Byeongjoon Kim and Minah Han contributed equally to this work.)

REFERENCES

[1] H. H. Barrett, “Objective assessment of image quality: Effects of quantum noise and object variability,” J. Opt. Soc. Amer. A, Opt. Image Sci., vol. 7, no. 7, pp. 1266–1278, 1990.

[2] P. Timberg, M. Båth, I. Andersson, S. Mattsson, A. Tingberg, and M. Ruschin, “Visibility of microcalcification clusters and masses in breast tomosynthesis image volumes and digital mammography: A 4AFC human observer study,” Med. Phys., vol. 39, no. 5, pp. 2341–2437, Apr. 2012.

[3] L. Chen, J. M. Boone, C. Abbey, J. Hargreaves, C. Bateni, K. K. Lindfors, K. Yang, A. Nosratieh, A. Hernandez, and P. Gazi, “Simulated lesion, human observer performance comparison between thin-section dedicated breast CT images versus computed thick-section simulated projection images of the breast,” Phys. Med. Biol., vol. 60, no. 8, p. 3347, 2015.

[4] M. P. Eckstein, J. L. Bartroff, C. K. Abbey, J. S. Whiting, and F. O. Bochud, “Automated computer evaluation and optimization of image compression of X-ray coronary angiograms for signal known exactly detection tasks,” Opt. Express, vol. 11, no. 5, pp. 460–475, 2003.

[5] M. Han, C. Lee, S. Park, and J. Baek, “Investigation on slice direction dependent detectability of volumetric cone beam CT images,” Opt. Express, vol. 24, no. 4, pp. 3749–3764, 2016.

[6] M. Han, B. Kim, and J. Baek, “Human and model observer performance for lesion detection in breast cone beam CT images with the FDK reconstruction,” PLoS ONE, vol. 13, no. 3, Mar. 2018, Art. no. e0194408.

[7] A. E. Burgess, “Statistically defined backgrounds: Performance of a modified nonprewhitening observer model,” J. Opt. Soc. Amer. A, Opt. Image Sci., vol. 11, no. 4, pp. 1237–1242, 1994.
BYEONGJOON KIM received the B.S. degree in engineering from Yonsei University, South Korea, in 2017, where he is currently pursuing the Ph.D. degree in engineering. His research interests include deep learning, computed tomography image denoising, and image quality assessment.

MINAH HAN received the B.S. degree in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2013, and the Ph.D. degree in engineering from Yonsei University, Incheon, South Korea, in 2019, with a focus on the development of model observer for quality evaluation on computed tomography (CT) images. She is currently a Postdoctoral Researcher with the Yonsei Institute of Convergence Technology, Yonsei University. Her research interests include CT image quality assessment, model observer development with a deep-learning approach, and medical image processing.

JONGDUK BAEK received the B.S. degree in electrical engineering from Yonsei University, Seoul, South Korea, in 2004, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, CA, USA, in 2007 and 2009, respectively, with a focus on the development of inverse geometry computed tomography (CT). He is currently an Associate Professor with the School of Integrated Technology, Yonsei University. His research interests include image quality assessment using mathematical observer models, CT imaging systems, medical imaging processing, and low-dose CT with a deep-learning approach.