Assessing the Impact of Deep Neural Network-Based Image Denoising on Binary Signal Detection Tasks

Kaiyan Li, *Graduate Student Member, IEEE*, Weimin Zhou, Hua Li, and Mark A. Anastasio, *Senior Member, IEEE*

***Abstract*— A variety of deep neural network (DNN)-based image denoising methods have been proposed for use with medical images. Traditional measures of image quality (IQ) have been employed to optimize and evaluate these methods. However, the objective evaluation of IQ for the DNN-based denoising methods remains largely lacking. In this work, we evaluate the performance of DNN-based denoising methods by use of task-based IQ measures. Specifically, binary signal detection tasks under signal-known-exactly (SKE) with background-known-statistically (BKS) conditions are considered. The performance of the ideal observer (IO) and common linear numerical observers are quantified and detection efficiencies are computed to assess the impact of the denoising operation on task performance. The numerical results indicate that, in the cases considered, the application of a denoising network can result in a loss of task-relevant information in the image. The impact of the depth of the denoising networks on task performance is also assessed. The presented results highlight the need for the objective evaluation of IQ for DNN-based denoising technologies and may suggest future avenues for improving their effectiveness in medical imaging applications.**

***Index Terms*— Image denoising, task-based image quality assessment, numerical observers, ideal observer, deep learning.**

I. Introduction

IMAGE denoising is a classical image processing operation that is commonly employed in medical imaging applications [1]–[7]. Recently, denoising methods based on deep neural networks (DNNs) have been proposed and widely investigated [5], [6], [8]–[15]. These methods are typically trained by minimizing loss functions that quantify a distance between the denoised image and the defined target image (e.g., a noise-free or low noise image) and have demonstrated high performance in terms of traditional image quality metrics such as root mean square error (RMSE), structural similarity index metric (SSIM) [16] or peak signal-to-noise ratio (PSNR).

In medical imaging, images are often acquired for specific purposes and the use of objective measures of image quality (IQ) has been widely advocated for assessing imaging systems and image processing algorithms [17]–[22]. Despite this, the objective evaluation of modern DNN-based medical image denoising methods remains largely lacking [23]. Although DNN-based denoising methods, by conventional design, can improve traditional IQ measures, it is well-known that such measures may not always correlate with objective task-based IQ measures [24]–[28]. For example, Yu *et al.* [23] conducted a study in which a DNN-based denoising method was observed to reduce RMSE compared to an alternative method, but signal detectability was unimproved [23].

Even more concerning is the fact that image denoising methods can compromise the visibility of important structural details in the denoised images even though traditional measurement metrics (such as RMSE or SSIM) are improved [2], [8], [29]. While DNN-based denoising operations may succeed at lowering noise levels, the extent to which they perturb the second- and higher-order statistical properties of an image that are relevant to signal detection is not understood. Finally, according to data processing inequality [30], the performance of an ideal observer cannot be increased via image processing operations such as denoising. However, conditions under which DNN-based denoising methods can improve the performance of sub-optimal observers on detection tasks remains relatively unexplored.

The purpose of this study is to assess modern DNN-based denoising methods by use of objective IQ measures, in a preliminary attempt to address the issues described above. Three canonical DNN-based denoising methods are identified for analysis. The convolutional neural network (CNN)-based observer, the Hotelling observer, the Regularized Hotelling observer, an anthromorphic channelized Hotelling observer,

Manuscript received March 28, 2021; accepted April 20, 2021. Date of publication April 30, 2021; date of current version August 31, 2021. This work was supported in part by NIH under Award R01EB020604, Award R01EB023045, Award R01NS102213, Award R01CA233873, and Award R21CA223799. *(Corresponding authors: Mark A. Anastasio; Hua Li.)*

Kaiyan Li and Mark A. Anastasio are with the Department of Bioengineering, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA (e-mail: kaiyanl2@illinois.edu; maa@illinois.edu).

Weimin Zhou was with the Department of Electrical and Systems Engineering, Washington University in St. Louis, St. Louis, MO 63130 USA. He is now with the Department of Psychological and Brain Sciences, University of California at Santa Barbara, Santa Barbara, CA 93106 USA (e-mail: weiminzhou@ucsb.edu).

Hua Li is with the Department of Bioengineering, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA, and also with the Carle Cancer Center, Carle Foundation Hospital, Urbana, IL 61801 USA (e-mail: huali19@illinois.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TMI.2021.3076810, provided by the authors.

Digital Object Identifier 10.1109/TMI.2021.3076810

and a non-prewhitening matched filter are implemented as NOs. The performances of these NOs acting on the original noisy images and the corresponding denoised images are quantified via receiver operating characteristic (ROC) analysis, and signal detection efficiencies are computed to assess the impact of the denoising operations on NO performance. The impact of the network depth of a DNN-based denoising method on NO performance is assessed to understand if the deep learning mantra "deeper is better" necessarily holds true for signal detection performance. A covariance matrix propagation analysis is also performed, to gain insights into how DNNs modify the covariance structure of image data as they are propagated through layers of a linear convolutional network. Finally, the depth of the CNN-based observer is varied to demonstrate how the benefit of the denoising operation is dependent on the specification of the NO. The presented analysis highlights the importance of objective IQ evaluation for DNN-based denoising technologies and may suggest future avenues for improving their effectiveness in medical imaging applications.

The remaining of the paper is organized as follows. Section II describes the necessary background on binary signal detection task, numerical observers, and image denoising. The numerical studies and the results of the proposed evaluations of different denoising networks are provided in Sections III and IV. Finally, the article provides a discussion of the key findings in Sec. V.

II. Background

A. Formulation of Binary Signal Detection Task

A linear digital imaging system can be described as a continuous-to-discrete (C-D) mapping process [17]:

$$\mathbf{g} = \mathcal{H} f(\mathbf{r}) + \mathbf{n}, \tag{1}$$

where $\mathbf{g} \in \mathbb{R}^{N\times 1}$ is the measured image vector, $f(\mathbf{r})$ denotes the object function that is dependent on the coordinate $\mathbf{r} \in \mathbb{R}^{k\times 1}$, $k \geq 2$, $\mathcal{H}$ denotes a linear imaging operator that maps $\mathbb{L}_2(\mathbb{R}^k)$ to $\mathbb{R}^{N\times 1}$, and $\mathbf{n} \in \mathbb{R}^{N\times 1}$ denotes the measurement noise. When its spatial dependence is not important to highlight, $f(\mathbf{r})$ will be denoted as $\mathbf{f}$.

A binary signal detection task requires an observer to classify the measured image data $\mathbf{g}$ as satisfying either a signal-present hypothesis H_1 or a signal-absent hypothesis H_0. These two hypotheses can be described as:

$$H_0 : \mathbf{g} = \mathcal{H}\mathbf{f_b} + \mathbf{n} = \mathbf{b} + \mathbf{n}, \tag{2a}$$

$$H_1 : \mathbf{g} = \mathcal{H}(\mathbf{f_b} + \mathbf{f_s}) + \mathbf{n} = \mathbf{b} + \mathbf{s} + \mathbf{n}, \tag{2b}$$

where $\mathbf{f_s}$ and $\mathbf{f_b}$ denote the signal and background, respectively, and $\mathbf{s} = \mathcal{H}\mathbf{f_s}$ and $\mathbf{b} = \mathcal{H}\mathbf{f_b}$ denote the signal and background images. For the case of a signal-known-exactly (SKE) and background-known-statistically (BKS) task, $\mathbf{s}$ is known while $\mathbf{b}$ is a random vector.

To perform this task, a deterministic observer computes a test statistic that maps the measured image $\mathbf{g}$ to a real-valued scalar variable that is compared to a predetermined threshold τ to determine if $\mathbf{g}$ satisfies H_0 or H_1. By varying the threshold τ, a ROC curve can be formed to quantify the trade-off between the false-positive fraction (FPF) and the true-positive fraction (TPF) [17]. The area under the ROC curve (AUC) can be subsequently calculated as a figure-of-merit (FOM) for signal detection performance.

B. Numerical Observers for IQ Assessment

In preliminary assessments of medical imaging technologies, NOs have been employed to quantify task-based measures of IQ for various image-based inferences [24]. The NOs that are employed in this study to perform binary SKE/BKS signal detection tasks are described briefly below.

1) Ideal Observer (IO) and CNN-Based Observer: The Bayesian Ideal Observer (IO) sets an upper limit of observer performance for signal detection tasks and has been advocated for use in optimizing medical imaging systems and data-acquisition designs [17]–[21]. The IO test statistic $t_{\text{IO}}(\mathbf{g})$ is any monotonic transformation of the likelihood ratio $\Lambda_{\text{LR}}(\mathbf{g})$:

$$\Lambda_{\text{LR}}(\mathbf{g}) = \frac{p(\mathbf{g}|H_1)}{p(\mathbf{g}|H_0)}, \tag{3}$$

where $p(\mathbf{g}|H_1)$ and $p(\mathbf{g}|H_0)$ are the conditional probability density functions that describe the measured data $\mathbf{g}$ under the hypotheses H_1 and H_0, respectively. Equation (3) is analytically intractable, in general, and Markov-chain Monte Carlo (MCMC) techniques have been proposed to approximate the IO test statistic [31]. In this study, an alternative method based on supervised learning is employed to approximate $\Lambda_{\text{LR}}(\mathbf{g})$. Specifically, this will be accomplished by use of an appropriately designed CNN-based classifier as described elsewhere [32]. The resulting NO will be referred to as the *CNN-IO observer*.

Please note that when a CNN-based classifier is employed as a NO but it does not possess sufficient model capacity to accurately approximate $\Lambda_{\text{LR}}(\mathbf{g})$, it will simply be referred to as a *CNN-based observer*. Therefore, the CNN-based observer is, by definition, a sub-optimal observer.

2) Hotelling Observer and Regularized Hotelling Observer: The Hotelling Observer (HO) is the IO that is restricted to employ test statistics that are linear functions of the data [17]. The HO employs the Hotelling discriminant, which is the population equivalent of the Fisher linear discriminant, and is optimal among all linear observers in the sense that it maximizes the signal-to-noise ratio of the test statistic [17]. The HO test statistic $t_{\text{HO}}(\mathbf{g})$ is defined as:

$$t_{\text{HO}}(\mathbf{g}) = \mathbf{w}_{\text{HO}}^T\mathbf{g} = (\mathbf{K_g}^{-1}\Delta\bar{\mathbf{g}})^T\mathbf{g}, \tag{4}$$

where $\mathbf{w}_{\text{HO}}^T \in \mathbb{R}^N$ denotes the Hotelling template, $\Delta\bar{\mathbf{g}} \in \mathbb{R}^N$ denotes the difference between the ensemble mean of the measurements $\mathbf{g}$ under the two hypotheses H_0 and H_1, and $\mathbf{K_g} \equiv \frac{1}{2}(\mathbf{K}_0(\mathbf{g}) + \mathbf{K}_1(\mathbf{g}))$. Here $\mathbf{K}_0(\mathbf{g}) \in \mathbb{R}^{N\times N}$ and $\mathbf{K}_1(\mathbf{g}) \in \mathbb{R}^{N\times N}$ denote the covariance matrices of $\mathbf{g}$ under the two hypotheses H_0 and H_1. If a linear imaging system and a SKE signal detection task are considered, $\Delta\bar{\mathbf{g}} = \mathbf{s}$. Note that the HO only employs first and second order statistical information about $\mathbf{g}$, whereas the IO requires full knowledge of the image data statistics.

In some cases, the covariance matrices $\mathbf{K}_0(\mathbf{g})$ and $\mathbf{K}_1(\mathbf{g})$ can be ill-conditioned and therefore the Hotelling template cannot

be stably computed. To address this, a regularized HO (RHO) can be employed that implements the test statistic $t_{\mathrm{RHO}}(\mathbf{g})$:

$$t_{\mathrm{RHO}}(\mathbf{g}) = \mathbf{w}_{\mathrm{RHO}}^T\mathbf{g} = (\mathbf{K}_\lambda^+\Delta\bar{\mathbf{g}})^T\mathbf{g}, \tag{5}$$

where $\mathbf{K}_\lambda$ represents a low-rank approximation of $\mathbf{K_g}$ that is formed by keeping only the singular values greater than $\lambda\sigma_{max}$. Here, $\mathbf{K}_\lambda^+$ is the Moore–Penrose inverse of $\mathbf{K}_\lambda$, λ is a threshold for the singular values and σ_{max} represents the largest singular value of $\mathbf{K_g}$, The value of λ can be tuned on an independent set of data and the value that leads to the best RHO performance can be selected.

3) Channelized Hotelling Observer: When the HO is employed with a channeling mechanism to reduce the dimensionality of the image data, a channelized HO (CHO) is formed. When implemented with difference-of-Gaussian (DOG) channels and an internal noise mechansim, the CHO can be interpreted as an anthropomorphic observer [33]–[35]. Let $\mathbf{T}$ denote a channel matrix and $\mathbf{v} \equiv \mathbf{Tg}$ the corresponding channelized image data. The CHO test statistic $t_{\mathrm{CHO}}(\mathbf{g})$ is given by:

$$t_{\mathrm{CHO}}(\mathbf{g}) = \left[(\mathbf{K_v} + \mathbf{K_{int}})^{-1}\Delta\bar{\mathbf{v}}\right]^T(\mathbf{v} + \mathbf{v_{int}}), \tag{6}$$

where $\mathbf{K_v}$ denotes the covariance matrix of the channelized data $\mathbf{v}$, $\mathbf{K_{int}}$ denotes the covariance matrix of the channel internal noise, and $\mathbf{v_{int}}$ is a noise vector sampled from the Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{K_{int}})$. Based on previous studies [35], in this work $\mathbf{K_{int}}$ will be defined as:

$$\mathbf{K_{int}} = \epsilon \cdot diag(\mathbf{K_v}), \tag{7}$$

where $diag(\mathbf{K_v})$ represents a diagonal matrix with diagonal elements from $\mathbf{K_v}$ and ϵ is the internal noise level. The parameters of the DOG channels and the internal noise level employed in this study are described below in Sec. III-C4.

4) Non-Prewhitening Matched Filter (NPWMF): The non-prewhitening matched filter (NPWMF) is a simple NO that utilizes only first-order statistical information [36], [37]. The NPWMF test statistic $t_{\mathrm{NPWMF}}(\mathbf{g})$ is given by:

$$t_{\mathrm{NPWMF}}(\mathbf{g}) = \Delta\bar{\mathbf{g}}^T\mathbf{g}, \tag{8}$$

where $\Delta\bar{\mathbf{g}} \in \mathbb{R}^N$ represents the difference of the means of the ensemble of measured images $\mathbf{g}$ under the two hypotheses H_0 and H_1, respectively. By design, the NPWMF will not be affected by changes to the second- and higher-order statistics of the image data.

C. DNN-Based Image Denoising

Denoising methods based on DNNs hold significant potential for medical imaging applications [1]–[8], [38], [39]. Due to their flexibility and ability to exploit image features, many such denoising methods have been proposed based on CNNs. Given a noisy image $\mathbf{g}$, the action of a DNN-based denoising method can be described generically as:

$$\hat{\mathbf{g}} = \mathcal{F}(\mathbf{g}, \Theta), \tag{9}$$

where the mapping $\mathcal{F}$ denotes the DNN that is parameterized by the weight vector Θ and $\hat{\mathbf{g}}$ denotes the estimated denoised image. Depending on how the target data are defined when training the DNN, $\hat{\mathbf{g}}$ can be interpreted as an estimate of the noiseless $\mathbf{g}$ or an estimate of $\mathbf{g}$ that contains a reduced noise level. When pre-training networks by use of simulated data, the former approach has been commonly employed [1]–[7], [29], [38], [39].

In addition to CNN-based methods, a variety of other approaches, including residual learning [40], have been employed for medical image denoising [5], [39]. The performance of denoising networks has commonly been evaluated by use of traditional metrics such as structural similarity index metric (SSIM) [16] and peak signal-to-noise ratio (PSNR).

III. Numerical Studies

Computer-simulation studies were conducted to objectively evaluate DNN-based denoising methods for SKE/BKS binary signal detection tasks. Three different DNNs were investigated, which were trained on simulated image data. The performances of the five different NOs reviewed in Sec. II-B on the noisy and denoised image data were analyzed under different conditions to gain insights into the potential impact of DNN-based denoising on signal detection.

A. Simulated Nuclear Medicine Images From a Parallel-Hole Collimator Imaging System

Planar scintigraphy images were simulated via an idealized linear parallel-hole collimator imaging system. The system was described by a linear C-D mapping $[\mathbf{g}]_m \equiv \int_V f(\mathbf{r})\, h_m(\mathbf{r})\, d\mathbf{r}$ that was specified by Gaussian point response functions [31]:

$$h_m(\mathbf{r}) = A_m \exp\left[-\frac{(\mathbf{r}-\mathbf{r}_m)^T(\mathbf{r}-\mathbf{r}_m)}{2w_m^2}\right], \tag{10}$$

where $[\mathbf{g}]_m$ denotes the m^{th} component of $\mathbf{g}$, V denotes the support of $f(\mathbf{r})$, and the amplitude $A_m = \frac{h}{2\pi w_m^2}$ with the height h and width w_m. The to-be-imaged objects $f(\mathbf{r}) = f_b(\mathbf{r}) + f_s(\mathbf{r})$ contained a random background and a superimposed deterministic signal in the signal present case. The random background $f_b(\mathbf{r})$ was specified by lumpy object model [31] as:

$$f_b(\mathbf{r}) = \sum_{n=1}^{N_b} l(\mathbf{r} - \mathbf{r}_n|a, w_b), \tag{11}$$

where $N_b \sim P(\bar{N})$ denotes the number of the lumps with $P(\bar{N})$ denoting a Poisson distribution with the mean $\bar{N}$. The lump function $l(\mathbf{r} - \mathbf{r}_n|a, w_b)$ was modeled by a 2D Gaussian function with lump amplitude a and lump width w_b:

$$l(\mathbf{r} - \mathbf{r}_n|a, w_b) = a\exp(-\frac{(\mathbf{r}-\mathbf{r}_n)^T(\mathbf{r}-\mathbf{r}_n)}{2w_b^2}), \tag{12}$$

where $\mathbf{r}_n$ denotes the center location of the n^{th} lump that was sampled from a uniform distribution over the spatial support of the image. For the signal present cases, the signal corresponded to a Gaussian signal:

$$f_s(\mathbf{r}) = A_s \exp\left[-\frac{(\mathbf{r}-\mathbf{r}_s)^T(\mathbf{r}-\mathbf{r}_s)}{2w_s^2}\right], \tag{13}$$

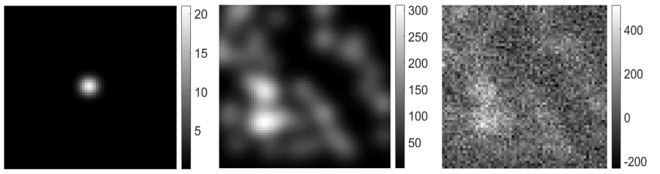

Fig. 1. These images are examples that depict (a) a possible signal $\mathbf{s}$, (b) a noise-free signal-present image $\mathbf{s}+\mathbf{b}$, and (c) the corresponding noisy measurement $\mathbf{g}$. The dimensions of the images are 64×64. As described in the text, the signal amplitude was relatively small to emulate a situation where the detection task is challenging.

where A_s is the signal amplitude, w_s is the signal width and $\mathbf{r_s}$ is the center of signal. The images $\mathbf{s} = \mathcal{H}\mathbf{f_s}$ and $\mathbf{b} = \mathcal{H}\mathbf{f_b}$ are given by:

$$[\mathbf{s}]_m = \frac{A_s h w_s^2}{w_m^2 + w_s^2} \exp\left[-\frac{(\mathbf{r}_m - \mathbf{r}_s)^T(\mathbf{r}_m - \mathbf{r}_s)}{2(w_m^2 + w_s^2)}\right], \quad (14)$$

and

$$[\mathbf{b}]_m = \frac{a h w_b^2}{w_m^2 + w_b^2} \sum_{n=1}^{N_b} \exp\left[-\frac{(\mathbf{r}_n - \mathbf{r}_m)^T(\mathbf{r}_n - \mathbf{r}_m)}{2(w_m^2 + w_b^2)}\right]. \quad (15)$$

The measurement noise $\mathbf{n}$ was described by an uncorrelated mixed Possion-Gaussian noise model. Details regarding the signal, background and noise are provided in Sec. III-C below. Figure 1 shows an example of the signal and a noise free signal-present image along with the corresponding noisy image data $\mathbf{g}$.

The relatively simple image models employed in our study provided a means by which simulated image data could be computed and degraded in a clear and controlled way, without being influenced by unknown noise sources that could potential be present in clinically acquired images.

B. DNN-Based Denoising Methods, Training, and Validation

A simple linear denoising network and two nonlinear denoising networks with CNN-based or ResNet-based architectures were considered as three representative examples to be evaluated in this study. Figure 2 shows the architectures of these three networks, which are described next.

1) Linear DNN-Based Denoising Method: As depicted in Fig. 2(a), the linear DNNs include only a collection of D linear convolutional layers. Although such networks will not achieve state-of-the art performance, they are considered here because they permit the analytic propagation of covariance matrices, and hence Hotelling templates, through the different layers of the network. Therefore, preliminary insights into how DNNs perturb information relevant to binary signal detection tasks can be gained. The network input was a noisy image $\mathbf{g}$ of dimension 32×32 and the output was the estimated $\hat{\mathbf{g}}$ with the same dimensions. In the first layer of the network, 32 filters of dimension $3 \times 3 \times 1$ were employed to generate 32 feature maps. In each of the 2^{nd} to the $(D\text{-}1)^{th}$ layers, 32 filters of dimension $3 \times 3 \times 32$ were employed. In the penultimate layer, a single filter of dimension $3 \times 3 \times 32$ was applied to map the tensor-valued feature map to the scalar-valued output image.

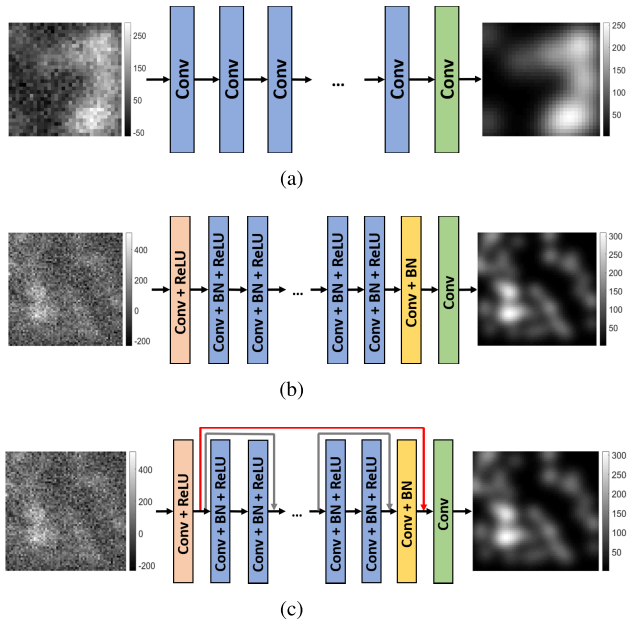

Fig. 2. The three denoising networks evaluated in this study were based upon a (a) linear CNN, (b) non-linear CNN, and (c) non-linear ResNet denoising network, respectively. The dimensions of the input and output images are (a) 32×32, (b) 64×64, (c) 64×64, respectively.

As described in Eqn. (1), let $\mathcal{H}\mathbf{f}_j$ denote a given ground truth (noiseless) image corresponding to either a signal absent or signal present case and let $\mathbf{g}_j$ denote the corresponding measured noisy image. Here, the subscript j has been added to index the objects and images. Given the collection of paired training data $\{(\mathbf{g}_j, \mathcal{H}\mathbf{f}_j)\}_{j=1}^{J}$, the linear network was trained by minimizing the mean-square-error (MSE) loss function:

$$\mathcal{L}_{\text{MSE}}(\Theta) = \frac{1}{J}\sum_{j=1}^{J} \|\mathcal{F}(\mathbf{g}_j; \Theta) - \mathcal{H}\mathbf{f}_j\|_2^2. \quad (16)$$

2) Nonlinear CNN-Based Denoising Network: As depicted in Fig. 2(b), a traditional non-linear CNN architecture of depth D was considered. The network input was a noisy image $\mathbf{g}$ of dimension 64×64 and the output was the estimated $\hat{\mathbf{g}}$ with the same dimensions. The CNN contained four types of layers. The first layer was a Conv+ReLU layer, in which 64 convolution filters of dimension $3 \times 3 \times 1$ were applied to generate 64 feature maps. In each of the 2^{nd} to $(D\text{-}2)^{th}$ Conv+BN+ReLU layers, 64 convolution filters of dimension $3 \times 3 \times 64$ were employed and batch normalization was included between the convolution and ReLU operations. In the $(D\text{-}1)^{th}$ Conv+BN layer, 64 convolution filters of dimension $3 \times 3 \times 64$ were employed and batch normalization was performed. In the last Conv layer, one single convolution filter of dimnension $3 \times 3 \times 64$ was employed to form the final denoised image of dimension 64×64. The network was trained by use of the MSE-based loss function.

3) Nonlinear ResNet-Based Denoising Network: An alternative nonlinear denoising network based on a ResNet architecture [40] was also investigated. As shown in Fig. 2(c),

the ResNet architecture employs shortcut connections (the so-called skip connections) between non-adjacent convolutional layers. This network design can better address the vanishing gradient issue [40], and allows for a deeper network with more convolutional layers. In this study, skip connections were added every other layer, as depicted by the gray line in Fig. 2(c). An additional skip connection, depicted as the brown line in Fig. 2(c), was added to connect the output of the 1^{st} layer and the input of the D^{th} (i.e., last) layer. Except for the skip connections, the network architecture was identical to that described above for the non-linear CNN.

Instead of using MSE-based loss, the perceptual loss was employed to train this network:

$$\mathcal{L}_{\text{Perceptual}}(\Theta) = \frac{1}{J}\sum_{j=1}^{J} \|\phi(\mathcal{F}(\mathbf{g}_j;\Theta)) - \phi(\mathcal{H}\mathbf{f}_j)\|_2^2, \quad (17)$$

where $\phi(\cdot)$ represents a feature extraction operator. It has been observed that denoising networks trained by use of a perceptual loss function can be effective in reducing noise while retaining image details [5].

4) Datasets and Denoising Network Training Details: The standard convention of utilizing separate training/validation/testing datasets was adopted. The training dataset included 10,000 noisy signal-present and 10,000 noisy signal-absent measurement images along with the corresponding noise-free target images. The validation datatset included 200 signal-present images and 200 signal absent images and the corresponding noise-free target images. Finally, the testing dataset comprised 10,000 signal-present images and 10,000 signal-absent noisy images.

These datasets were computed as follows. First, lumpy background images, which were generated according to Eqn. (15), were employed as the noise-free signal-absent images. Then, a Gaussian signal was inserted to the background images to create noise-free images under the signal-present hypothesis. The signal was defined in Eqn. (14). Finally, mixed Poisson and Gaussian noise was added to the noise-free images under both hypotheses. The training, validation, or testing datasets were generated separately according to the steps described above. The statistical properties of these images varied between studies and are described below.

All the denoising networks were trained on mini-batches at each iteration by use of the Adam optimizer [41] with a learning rate of 0.0001. Each mini-batch contained 200 signal-present images and 200 signal absent images that were randomly selected from the training dataset. The network model that possessed to the best performance on the validation dataset was selected for use. Keras [42] was employed for implementing and training all networks on a single NVIDIA TITAN X GPU.

When training the nonlinear ResNet-based denoising network, the output before the first pooling layer from a pre-trained VGG19 [43] network was employed as a feature extraction operator to compute the perceptual loss in Eqn. (17). A similar feature extraction operator was utilized by Gong *et al.* [5]. The VGG19 network contained 16 convolutional layers, 5 max pooling layers, and 3 fully connected layers, and was trained by use of images from ImageNet [44]. A total of 64 feature maps were extracted with spatial size 64×64 to compute the perceptual loss.

C. Objective Evaluation of Denoising Networks

1) Studies Involving Linear Denoising Networks: A study was implemented to assess the performance of the RHO when acting on data corresponding to the outputs of different intermediate layers in the linear denoising network. In this way, the RHO performance could be observed as it propagates through the network. The RHO was utilized because the resulting covariance matrices were generally ill-conditioned.

In the detection task, the signal defined in Eqn. (13) was employed with $A_s = 2.5$, $w_s = 1$, and $\mathbf{r}_s = [16; 16]^T$. The parameters of the lumpy background model defined in Eqn. (11) were $\bar{N} = 15$, $a = 5$, and $w_b = 3$. The dimensions of $\mathbf{s}$, $\mathbf{b}$, $\mathbf{n}$, and $\mathbf{g}$ in Eqn. (2) were 32×32. The assumed parameters of the imaging system defined in Eqn. (10) were $A_m = \frac{h}{2\pi w_m^2}$, $h = 20$ and $w_m = 2$. For the mixed Possion-Gaussian noise, the Gaussian noise was sampled from a Gaussian distribution with the mean 0 and the standard deviation 25. Based on these settings, the training/validation/testing datasets were established and the linear denoising networks with depths varying from $D = 2$ to $D = 15$ were trained as described above in Sec. III-B1. Each network with different D was trained separately to achieve the optimal performance based on the defined loss function.

In order to compute the RHO acting on the tensor-valued feature data produced by each network layer, the covariance matrix $\mathbf{K}_d$ of the output data tensor of each layer needed to be estimated. Here, d denotes a layer index. To accomplish this, the tensor-valued data were vectorized and the associated covariance matrices corresponding to each layer were computed by propagating the covariance matrix $\mathbf{K_0}$ of the noisy input image through the network. Details regarding this procedure are provided in Sec. 1 of the Supplementary file.

2) Studies Involving the Non-Linear Denoising Networks: A study was designed to investigate the performance of NOs when acting on the original noisy measurement images and the corresponding denoised images produced by the non-linear CNN and ResNet-based networks. Several parameters of the simulated images and denoising networks were varied to gain insights into the potential impact of denoising on NO performance.

For the considered detection tasks, the signals, the lumpy object model, and the parallel-hole collimator imaging system were defined as in Sec. III-C1 but with different parameter settings. The signal possessed an amplitude $A_s = 3$, width $w_s = \sqrt{2}$, and center location $\mathbf{r}_s = [32; 32]^T$. The parameters of the lumpy background model defined in Eqn. (11) were $\bar{N} = 50$, $a = 5$, and $w_b = 3$. The dimensions of $\mathbf{s}$, $\mathbf{b}$, and $\mathbf{n}$ in Eqn. (2) were 64×64. The parallel-hole collimator imaging system was specified as $A_m = \frac{h}{2\pi w_m^2}$, $h = 20$ and $w_m = 2$. The standard deviation of Gaussian noise was set to 75. Based on these settings, the training/validation/testing datasets were established and nonlinear denoising networks of depth $D = \{3, 5, 7, 9, 11, 13\}$ were trained as described above in

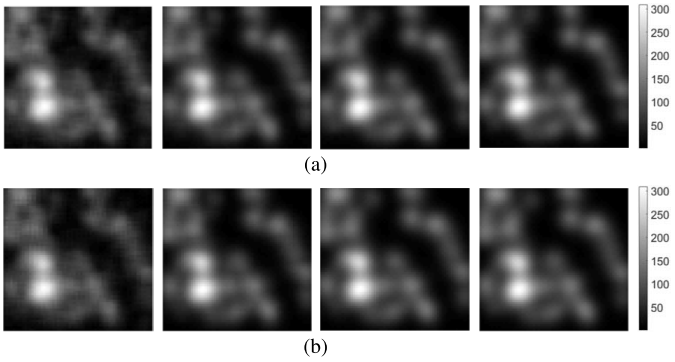

Fig. 3. The images, from left to right, in each row represent the denoised estimates $\hat{\mathbf{g}}$ obtained by use of a) the CNN-based and b) the ResNet-based non-linear networks with varied {3, 7, 11, 13} layers, respectively. The related noise-free signal-present target image $\mathcal{H}(\mathbf{f_s} + \mathbf{f_b})$ and the original noisy image $\mathbf{g}$ were the second and third images shown in Fig. 1. The dimensions of the images are 64×64.

Sec. III-B2. Examples of denoised images $\hat{\mathbf{g}}$ produced by use of the CNN-based and ResNet-based denoising networks of different depths D are shown in Fig. 3. This study was also repeated for the case where low-noise, instead of noiseless, target images, were employed for training. Those studies are presented in Sec. 2 of the Supplementary file.

Finally, the impact of the signal size on the performance of the RHO was investigated. Signals of width $w_s = \{1, \sqrt{2}, 2, 2.5, 3\}$ were considered. All other parameters were kept the same as that described above.

3) Observer Performance Evaluation Metrics: To evaluate the performance of the NOs, ROC analysis was conducted and AUC values were computed and employed as a figure-of-merit. The ROC curves were fit by use of the Metz-ROC software [45] that employs the proper binormal model [46]. The error bars of the AUC values were estimated as well. Detection efficiencies for a given NO and denoising method were defined as

$$e \equiv \frac{\text{AUC}_{\text{denoised}}}{\text{AUC}_{\text{noisy}}}, \tag{18}$$

where $\text{AUC}_{\text{denoised}}$ and $\text{AUC}_{\text{noisy}}$ denote the AUC values corresponding to a NO acting on the denoised and original noisy image data, respectively. The detection efficiency quantifies the impact of the denoising operation on the performance of the NO. It should be noted that this definition is different from that employed elsewhere in the literature, where detection efficiency is typically referenced to an IO [47]. As such, it is possible that $e > 1$ when the IO is not employed. The denoised images were also assessed by use of RMSE and SSIM.

4) Numerical Observer Computation: The CNN-IO was employed to approximate the IO test statistic [32]. Details regarding the implementation of the CNN-IO and CNN-based observers are provided in Sec. 5 of the Supplementary file.

For computing the HO and RHO test statistics, the covariance matrix $\mathbf{K_g}$ need to be estimated. For use in evaluating the linear denoising networks, the covariance matrix decomposition method [17], [32] was initially employed to estimate the covariance matrix of the original noisy images. To estimate the covariance matrix of the background images, 100,000 signal-present and 100,000 signal-absent noiseless images were utilized. Subsequently, to examine how task-performance propagates through the networks, the covariance matrices corresponding to the vectorized feature tensors at each network layer were computed by use of the propagation strategy described in Sec. 1 of the Supplemental file. For evaluating the nonlinear denoising networks, the covariance matrices corresponding to both the noisy and denoised images were empirically estimated by use of 100,000 signal-present and 100,000 signal-absent images.

When computing the RHO test statistic, the threshold parameter λ in Eqn. (5) was swept from $1e-3$ to $1e-7$ and the corresponding detection performance was estimated based on a separate validation dataset including 2,000 signal-present images and 2,000 signal-absent images. The value which led to the best RHO detection performance was selected. The RHO with selected parameter was then applied to the testing dataset described below and the corresponding observer performance was estimated. The NPWMF template was established by use of the same training data as employed to establish the RHO.

For computing the CHO test statistic, 2,000 signal-present and 2,000 signal-absent images were utilized to estimate the channelized covariance matrix. A set of 10 DOG channels [35] was employed with channel parameters $\sigma_0 = 0.005$, $\alpha = 1.4$, and $Q = 1.67$. The internal noise level ϵ was 2.5, which was the same value employed by Abbey *et al.* [35].

The performance of the NOs on the original noisy images was evaluated by use of a testing dataset with 10,000 signal-present noisy images and 10,000 signal-absent noisy images that was described above in Sec. III-B4. Subseqently, the performance of the NOs was assessed by use of the denoised testing images.

IV. Results

A. Propagation of Task-Based Information Through a Linear Denoising Network

The performance of the RHO acting on the noisy test data and on data corresponding to the outputs of different intermediate layers in the linear denoising network is summarized in Table I. The covariance matrices needed to compute the RHO test statistic corresponding to the output of each network layer were calculated by use of the propagation strategy described in Sec. 1 of the supplementary file. With the exception of the network with three layers, the RHO performance on the denoised images was lower than on the original noisy images, and the performance decreases more on the image denoised by deeper networks.

To gain insights into this behavior, the singular value spectra of the covariance matrices estimated from the original noisy images and from the images denoised by networks with varied depths were examined. The results, shown in Fig. 4, reveal that the spectra corresponding to the denoised images decay more rapidly than that corresponding to the original noisy image. Additionally, the spectra corresponding to the denoised images decayed more rapidly as the denoising network became deeper. Accordingly, the number of singular values that exceeded the value of the threshold $\lambda\sigma_{max}$ that specified the RHO

TABLE I

RHO SIGNAL DETECTION PERFORMANCE PROPAGATION THROUGH LINEAR CNN-BASED DENOISING NETWORK WITH {3, 5, 7, 9, 11, 13, 15} LAYERS WERE DEMONSTRATED BY USE OF AUC VALUES. THE STANDARD ERROR OF EACH AUC VALUE WAS THE SAME OF 0.003

RHO detection performance (AUC value) at the output of different layers of the linear denoising network								
Noisy measurements	**Layer Index**	**The denoising network with different layers**						
		3 layers	**5 layers**	**7 layers**	**9 layers**	**11 layers**	**13 layers**	**15 layers**
0.6376	3	0.6376	0.6376	0.6376	0.6376	0.6376	0.6376	0.6376
	5		0.6372	0.6376	0.6376	0.6376	0.6376	0.6376
	7			0.6316	0.6376	0.6376	0.6376	0.6376
	9				0.6283	0.6376	0.6376	0.6376
	11					0.6213	0.6376	0.6376
	13						0.6188	0.6376
	15							0.6158

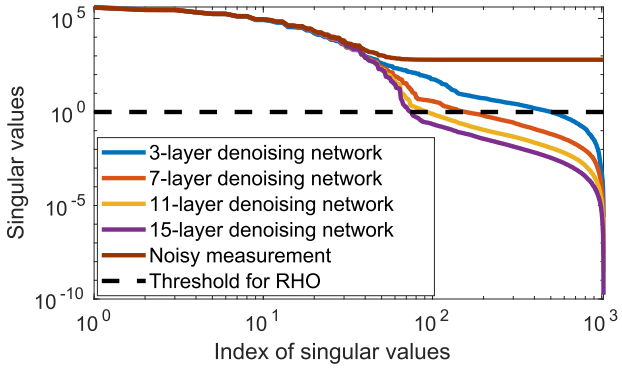

Fig. 4. The singular value spectra of the covariance matrices corresponding to the original noisy images and the images denoised by use of the linear denoising networks with depths of {3, 7, 11, 15} were demonstrated.

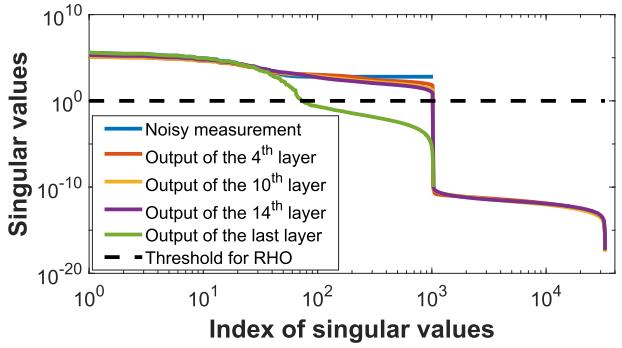

Fig. 5. The singular value spectra of covariance matrices corresponding to the original noisy images and the outputs of different layers in a linear CNN denoising network with the depth $D = 15$ were illustrated.

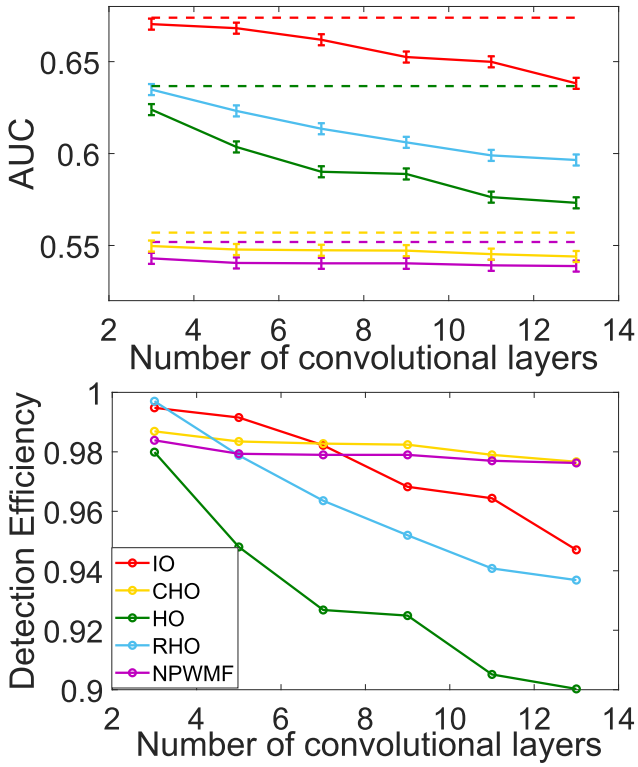

Fig. 6. The relationships between AUC (top figure) and detection efficiency (bottom figure) and the depth (the number of convolutional layers) of a CNN-based non-linear denoising method when different NOs are employed were quantified. The two figures share the same legend that is displayed in the bottom figure. The dashed lines in the upper figure depict the performances of the NOs on the original noisy images.

via Eqn. (5) decreased as the network depth increased. This resulted in the RHO performance to degrade as the network depth increased.

The propagation of RHO performance through the networks is summarized in Table I. It was observed that the RHO performance on data produced by the intermediate denoising network layers remained approximately constant until the last layer, at which point it decreased. It should be noted that the last layer of the denoising network transforms a high-dimensional feature tensor to the denoised output image. This operation possesses a null space and is therefore non-invertible. The drop in RHO performance at the last layer suggests that some of the features that were important to task-performance resided in the null space of the learned transformation.

To understand why NO performance remained constant until the last layer, the singular value spectra of the covariance matrices estimated from the original noisy images and the feature tensors corresponding to intermediate layers of the denoising network were further analyzed for the case of the network of depth $D = 15$. The results, shown in Fig. 5, reveal

TABLE II
THE RMSE AND SSIM VALUES ASSOCIATED WITH NOISY IMAGES AND THE OUTPUT IMAGES OF TWO DIFFERENT NONLINEAR DENOISING NETWORKS WERE COMPARED

Denoising networks	Measurement Metrics			
	CNN + MSE		ResNet + Perceptual	
	RMSE	SSIM	RMSE	SSIM
Noise image	75.4161	0.3663	75.4161	0.3663
3 layers	13.1478	0.9370	13.3280	0.9337
5 layers	12.1819	0.9469	12.3120	0.9463
7 layers	11.5499	0.9526	11.6433	0.9519
9 layers	11.4584	0.9535	11.4340	0.9540
11 layers	11.4563	0.9536	11.3016	0.9549
13 layers	11.4548	0.9537	11.2556	0.9555

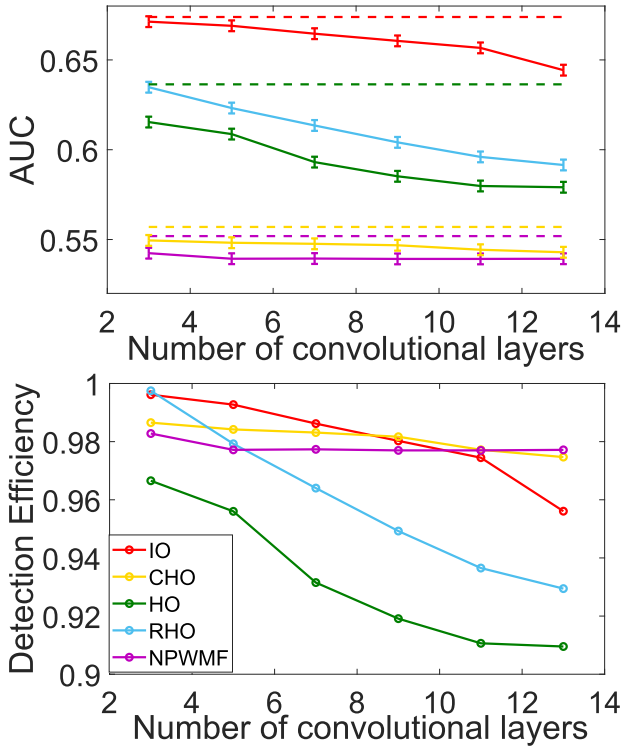

Fig. 7. The relationship between NO performance and the depth (the number of convolutional layers) of the ResNet-based non-linear denoising networks was quantified. The two figures on each panel share the same legend. The dashed lines in the upper figure represent the performance of the NOs on the noisy images.

that the spectra corresponding to the intermediate layers were similar to that corresponding to the original noisy images. Accordingly, the number of singular values that exceeded the value of the threshold $\lambda\sigma_{max}$ that specified the RHO via Eqn. (5) at different intermediate network layers remained constant as the network depth increased. This resulted in the RHO performance remaining fixed as the network depth increased, until the last layer was reached as discussed above.

B. Impact of Denoising Network Depth

1) Performance Changes: The impact of depth of the non-linear CNN and ResNet networks on the NO performance

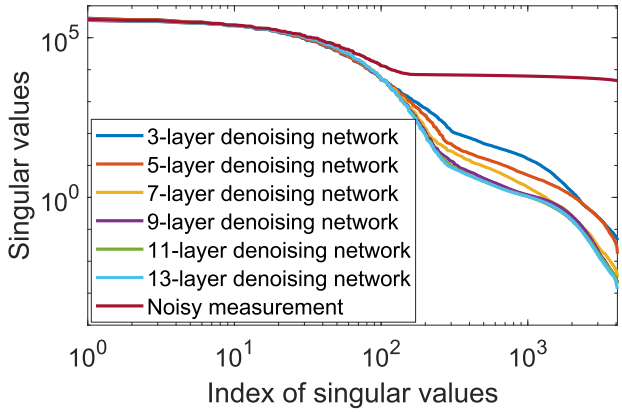

Fig. 8. The singular values of the covariance matrices from noisy images and images denoised by CNN-based non-linear denoising networks with $\{3, 5, 7, 9, 11, 13\}$ convolutional layers were compared, respectively. The denoising operation changes the structure of data covariance matrix. The changes are more obvious for deeper networks.

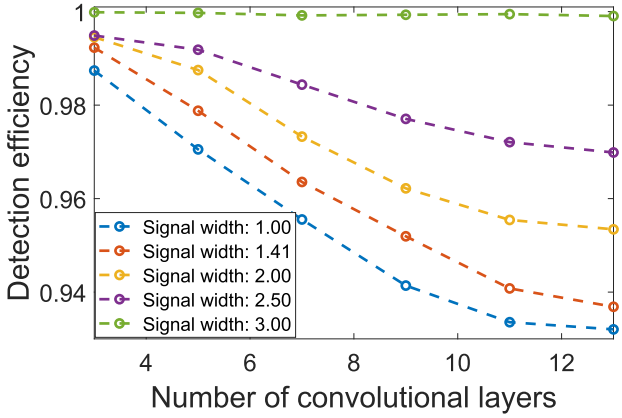

Fig. 9. The relationships between signal size and RHO detection efficiency were quantified. Here, the CNN-based denoising method was employed with an MSE loss and the network depth was varied: $D = \{3, 5, 7, 9, 11, 13\}$. Detection efficiency reduced more rapidly as a function of network depth when the signal size was reduced.

as measured by AUC and detection efficiency is shown in Fig. 6 and Fig. 7. For all cases, it was observed that the performance of the NOs on the original noisy images was higher than on the denoised images. The performance of the CNN-IO, HO and RHO on the denoised images decreased as the depth of the denoising networks increased. Contrarily, the performance of CHO and NPWMF on the denoised images was relatively insensitive to the depth of the denoising networks. These observations suggest that the second- and potentially higher-order statistical properties of the images were degraded by the denoising networks; this is confirmed below in Fig. 8. The quality of the denoised images as measured by RMSE and SSIM values for the networks of varying depth are shown in Table II. As expected, these metrics improved as the depth of the denoising networks increased. These results confirm that objective measures of IQ based on signal detection performance can show conflicting trends as compared to traditional metrics when comparing different denoising networks.

2) Changes in Covariance Matrix Induced by Denoising: The degradation of HO performance was further analyzed by computing the SVD of the covariance matrices corresponding to the images denoised by use of the CNN-based method. The results, shown in Fig. 8, reveal that the covariance matrix corresponding to the denoised images was ill-conditioned, while that corresponding to the original noisy images was well-conditions. Moreover, the singular value spectra tended to decrease more rapidly as the depth of the denoising network was increased. Although not shown, similar observations were made in the case of the ResNet-based denoising method. These results confirm that the denoising networks changed the second-order statistical properties of the denoised images. As mentioned above, the performance of the NPWMF observer, which uses only first-order statistical information, was not strongly degraded by denoising. Together, these observations support the assertion that the reduction in performance of the NOs that were sensitive to second- and higher-order image statistics was caused by the the changes in these properties induced by the denoising operation.

C. Detection Efficiency vs. Signal Size

The impact of signal size on RHO detection efficiency is shown in Fig. 9. Here, the width w_s of the Gaussian signal in Eqn. (13) took on the values: $\{1, \sqrt{2}, 2, 2.5, 3\}$. It was observed that, for each signal size, the detection efficiency was reduced as the denoising network depth increased. Additionally, the detection efficiency reduced more rapidly as a function of network depth for smaller signal sizes as compared to larger ones. Specifically, for $w_s = 3$, there was no statistically significant decrease in detection efficiency as the denoising network depth increased. Moreover, the detection efficiency was close to one. This is due to the relatively large size of the signal and use of an MSE loss function to train the denoising network. An MSE loss function treats every pixel in an image equally and therefore a large signal contributes more than a small one during the network training (i.e., more task-specific information is potentially preserved).

D. Situations Where Denoising Improved Detection Performance

CNN-based observers of varying depths were employed to demonstrate conditions under which the CNN- and ResNet-based denoising methods could improve signal detection performance. Detection performance was assessed on the original noisy images and the outputs of the two denoising networks with the depth of $\{3, 9, 11\}$, respectively. The evaluated CNN-based observers for this study were set with {1, 2, 4, 6, 8, 10} convolutional layers, respectively. It should be noted that the CNN-based observer with 10 layers coincided with the CNN-IO, and therefore approximated the IO for this task.

The results shown in Fig. 10 reveal, as expected, that the performance of the CNN-based observer increases with observer network depth. More interestingly, the detection performance of the shallow CNN-based observer with 3 layers on the original noisy images was worse than that on the images

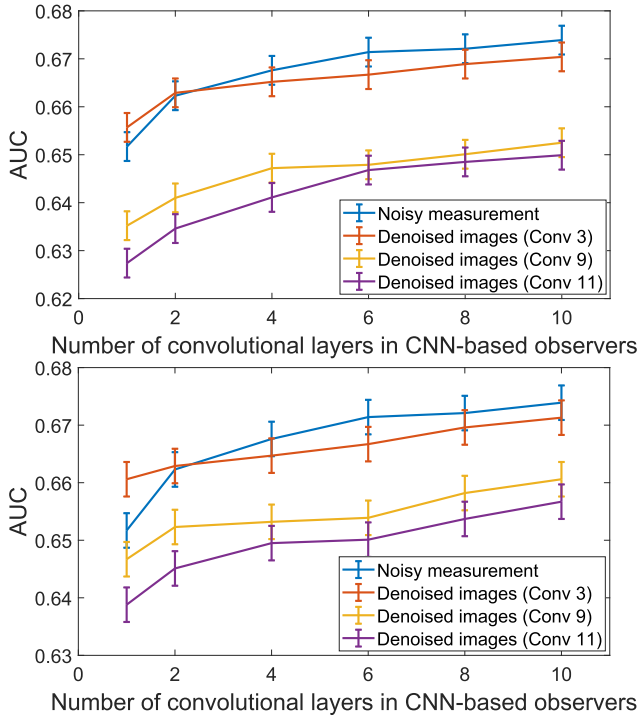

Fig. 10. The performance of the CNN-based observers with different number of convolutional layers acting on the original noisy image and the outputs of two non-linear denoising networks were compared. The upper panel shows the results on the CNN-based nonlinear denoising networks; The lower panel shows the results on the ResNet-based nonlinear denoising networks. Note that the y-axis range is clipped for display purposes.

denoised by a non-linear denoising network that also had 3 convolutional layers. This represented a situation in which the denoising operation resulted in improved signal detection performance.

As observed and discussed above in Section IV-A, the use of deeper denoising networks resulted in a stronger degradation in signal detection performance as compared to use of shallower networks for the NOs considered. Additionally, according to data processing inequality [30], it is known that the performance of an IO cannot be increased via image processing operations such as denoising. As such, it is to be expected that the performance of the CNN-IO on the original noisy image data will not be improved by use of any denoising operation. These factors suggest that the extent to which a denoising operation will improve signal detection performance depends, in a complicated way on (at least) the following: 1) the extent to which the denoising operation degrades the image statistics that are employed by a given NO for a specified inference; and 2) the extent to which the NO approximates the IO.

V. Summary and Discussion

In this work, the performance of DNN-based denoising methods was evaluated by use of task-based IQ measures. Specifically, binary signal detection tasks under SKE/BKS conditions were considered. The performance of the IO and common linear NOs were quantified to assess the impact of the denoising operation on task performance. This study was

motivated by the scarcity of works that have evaluated such modern denoising methods by use of objective methods.

The numerical results showed that, in the cases considered, the denoising operation can result in a loss of task-relevant information. Moreover, it was observed that while increasing the depth of the denoising network improved RMSE and SSIM, it resulted in a decrease in NO performance. This is consistent with the well-known fact that physical IQ measures may not always correlate with task-based ones [26]. This result also suggests that the mantra "deep is better" should be qualified and may not always hold true for objective IQ measures. The considered networks were analyzed to gain insights into the observed behavior and it was found that the denoising operation resulted in ill-conditioned covariance matrices. As such, denoising networks, while seeking to minimize a traditional (non-task-based) loss function, have the potential to degrade the image statistics that are important for signal detection.

Conditions under which the considered denoising operations could improve NO performance were also investigated. In the presented studies, it was observed that a shallow denoising network could improve the performance of a shallow CNN-based observer. When the depth of either the denoising or observer networks increased, the benefit of denoising was lost and NO performance was degraded. This suggests that the impact of denoising on signal detection performance depends, in a complicated way, on the specification of the denoising network, task, and the NO. As such, there is an urgent need to objectively evaluate new DNN-based denoising methods.

There remain numerous important topics for future investigation. The binary SKE detection task considered in this study is simplistic relative to many real-word clinical tasks. It will be important to consider more complicated tasks that involve signal variability and hybrid tasks that involve detection and estimation [48]. The study design presented can also readily be applied to assess alternative DNN-based denoising methods that use varying network architectures and loss functions. Ultimately, it will be critical to conduct human reader studies to assess the utility of new DNN-based denoising methods for specific clinical tasks.

Finally, the presented results will motivate the development of new approaches to establishing DNN-based denoising methods that mitigate the loss of task-relevant information by incorporating task-relevant information in the training strategy.

REFERENCES

[1] A. Manduca *et al.*, "Projection space denoising with bilateral filtering and CT noise modeling for dose reduction in CT," *Med. Phys.*, vol. 36, no. 11, pp. 4911–4919, Oct. 2009.

[2] Z. Li *et al.*, "Adaptive nonlocal means filtering based on local noise level for CT denoising," *Med. Phys.*, vol. 41, no. 1, Dec. 2013, Art. no. 011908.

[3] J.-W. Lin, A. F. Laine, and S. R. Bergmann, "Improving PET-based physiological quantification through methods of wavelet denoising," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 2, pp. 202–212, Jun. 2001.

[4] A. Le Pogam, H. Hanzouli, M. Hatt, C. C. Le Rest, and D. Visvikis, "Denoising of PET images by combining wavelets and curvelets for improved preservation of resolution and quantitation," *Med. Image Anal.*, vol. 17, no. 8, pp. 877–891, Dec. 2013.

[5] K. Gong, J. Guan, C.-C. Liu, and J. Qi, "PET image denoising using a deep neural network through fine tuning," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 3, no. 2, pp. 153–161, Mar. 2019.

[6] H. Chen *et al.*, "Low-dose CT denoising with convolutional neural network," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 143–146.

[7] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[8] Q. Yang *et al.*, "Low-dose CT image denoising using a generative adversarial network with wasserstein distance and perceptual loss," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1348–1357, Jun. 2018.

[9] L. Gondara, "Medical image denoising using convolutional denoising autoencoders," in *Proc. IEEE 16th Int. Conf. Data Mining Workshops (ICDMW)*, Dec. 2016, pp. 241–246.

[10] E. Kang, J. Min, and J. C. Ye, "A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction," *Med. Phys.*, vol. 44, no. 10, pp. e360–e375, Oct. 2017.

[11] H. Chen *et al.*, "Low-dose CT with a residual encoder-decoder convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 36, no. 12, pp. 2524–2535, Dec. 2017.

[12] P. Liu, M. D. El Basha, Y. Li, Y. Xiao, P. C. Sanelli, and R. Fang, "Deep evolutionary networks with expedited genetic algorithms for medical image denoising," *Med. Image Anal.*, vol. 54, pp. 306–315, May 2019.

[13] X. You, N. Cao, H. Lu, M. Mao, and W. Wanga, "Denoising of MR images with rician noise using a wider neural network and noise range division," *Magn. Reson. Imag.*, vol. 64, pp. 154–159, Dec. 2019.

[14] D. Liu, B. Wen, J. Jiao, X. Liu, Z. Wang, and T. S. Huang, "Connecting image denoising and high-level vision tasks via deep learning," *IEEE Trans. Image Process.*, vol. 29, pp. 3695–3706, 2020.

[15] C. Tian, Y. Xu, L. Fei, and K. Yan, "Deep learning for image denoising: A survey," in *Proc. Int. Conf. Genet. Evol. Comput.* Singapore: Springer, 2018, pp. 563–572.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[17] H. H. Barrett and K. J. Myers, *Foundations Image Science*. Hoboken, NJ, USA: Wiley, 2013.

[18] C. E. Metz, R. F. Wagner, K. Doi, D. G. Brown, R. M. Nishikawa, and K. J. Myers, "Toward consensus on quantitative assessment of medical imaging systems," *Med. Phys.*, vol. 22, no. 7, pp. 1057–1061, Jul. 1995.

[19] W. Vennart, "ICRU Report 54: Medical imaging—The assessment of image quality: ISBN 0-913394-53-X. April 1996, Maryland, USA," *Radiography*, vol. 3, no. 3, pp. 243–244, 1997.

[20] R. F. Wagner and D. G. Brown, "Unified SNR analysis of medical imaging systems," *Phys. Med. Biol.*, vol. 30, no. 6, p. 489, 1985.

[21] X. He and S. Park, "Model observers in medical imaging research," *Theranostics*, vol. 3, no. 10, p. 774, 2013.

[22] K. Li, W. Zhou, H. Li, and M. A. Anastasio, "Supervised learning-based ideal observer approximation for joint detection and estimation tasks," *Proc. SPIE Med. Imag., Image Perception, Observer Perform., Technol. Assessment*, vol. 11599, Oct. 2021, Art. no. 115990F.

[23] Z. Yu *et al.*, "Ai-based methods for nuclear-medicine imaging: Need for objective task-specific evaluation," *J. Nucl. Med.*, vol. 61, no. 1, p. 575, 2020.

[24] H. H. Barrett, J. Yao, J. P. Rolland, and K. J. Myers, "Model observers for assessment of image quality," *Proc. Nat. Acad. Sci. USA*, vol. 90, no. 21, pp. 9758–9765, 1993.

[25] O. Christianson *et al.*, "An improved index of image quality for task-based performance of CT iterative reconstruction across three commercial implementations," *Radiology*, vol. 275, no. 3, pp. 725–734, Jun. 2015.

[26] K. J. Myers, H. H. Barrett, M. C. Borgstrom, D. D. Patton, and G. W. Seeley, "Effect of noise correlation on detectability of disk signals in medical imaging," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 2, no. 10, pp. 1752–1759, Oct. 1985.

[27] A. Badal, K. H. Cha, S. E. Divel, C. G. Graff, R. Zeng, and A. Badano, "Virtual clinical trial for task-based evaluation of a deep learning synthetic mammography algorithm," *Proc. SPIE*, vol. 10948, Oct. 2019, Art. no. 109480O.

[28] K. Li, W. Zhou, H. Li, and M. A. Anastasio, "Task-based performance evaluation of deep neural network-based image denoising," *Proc. SPIE Med. Imag. Image Perception, Observer Perform., Technol. Assessment*, vol. 11599, May 2021, Art. no. 115990L.

[29] S. Li, J. Zhou, D. Liang, and Q. Liu, "MRI denoising using progressively distribution-based neural network," *Magn. Reson. Imag.*, vol. 71, pp. 55–68, Sep. 2020.
[30] N. J. Beaudry and R. Renner, "An intuitive proof of the data processing inequality," *Quantum Inf. Comput.*, vol. 12, nos. 5–6, pp. 432–441, May 2012.
[31] M. A. Kupinski, J. W. Hoppin, E. Clarkson, and H. H. Barrett, "Ideal-observer computation in medical imaging with use of Markov-chain Monte Carlo techniques," *JOSA A, Opt. Image Sci. Vis.*, vol. 20, no. 3, pp. 430–438, 2003.
[32] W. Zhou, H. Li, and M. A. Anastasio, "Approximating the ideal observer and hotelling observer for binary signal detection tasks by use of supervised learning methods," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2456–2468, Oct. 2019.
[33] K. J. Myers and H. H. Barrett, "Addition of a channel mechanism to the ideal-observer model," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 4, no. 12, pp. 2447–2457, 1987.
[34] C. Abbey and F. Bochud, "Modeling visual detection tasks in correlated image noise with linear model observers," in *Handbook of Medical Image Physics and Psychophysics*, vol. 1, R. L. Van Metter, J. Beutel, and H. L. Kundel, Eds. Bellingham, WA, USA: SPIE, 2000, pp. 629–654.
[35] C. K. Abbey and H. H. Barrett, "Human-and model-observer performance in ramp-spectrum noise: Effects of regularization and object variability," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 18, no. 3, pp. 473–488, 2001.
[36] R. F. Wagner, D. G. Brown, and M. S. Pastel, "Application of information theory to the assessment of computed tomography," *Med. Phys.*, vol. 6, no. 2, pp. 83–94, Mar. 1979.
[37] 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.
[38] J. V. Manjón, P. Coupé, A. Buades, D. Louis Collins, and M. Robles, "New methods for MRI denoising based on sparseness and self-similarity," *Med. Image Anal.*, vol. 16, no. 1, pp. 18–27, Jan. 2012.
[39] W. Jifara, F. Jiang, S. Rho, M. Cheng, and S. Liu, "Medical image denoising using convolutional neural network: A residual learning approach," *J. Supercomput.*, vol. 75, no. 2, pp. 704–718, Feb. 2019.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[42] F. Chollet. (2015). *Keras*. [Online]. Available: https://keras.io
[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[45] C. E. Metz, B. A. Herman, and C. A. Roe, "Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets," *Med. Decis. Making*, vol. 18, no. 1, pp. 110–121, 1998.
[46] L. L. Pesce and C. E. Metz, "Reliable and computationally efficient maximum-likelihood estimation of 'proper' binormal ROC curves," *Academic Radiol.*, vol. 14, no. 7, pp. 814–829, Jul. 2007.
[47] S. Park, E. Clarkson, M. A. Kupinski, and H. H. Barrett, "Efficiency of the human observer detecting random signals in random backgrounds," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 22, no. 1, pp. 3–16, Jan. 2005.
[48] E. Clarkson, "Estimation receiver operating characteristic curve and ideal observers for combined detection/estimation tasks," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 24, no. 12, p. B91, 2007.