Generative Adversarial Networks for Noise Reduction in Low-Dose CT

Jelmer M. Wolterink, Tim Leiner, Max A. Viergever, and Ivana Išgum

***Abstract*— Noise is inherent to low-dose CT acquisition. We propose to train a convolutional neural network (CNN) jointly with an adversarial CNN to estimate routine-dose CT images from low-dose CT images and hence reduce noise. A generator CNN was trained to transform low-dose CT images into routine-dose CT images using voxelwise loss minimization. An adversarial discriminator CNN was simultaneously trained to distinguish the output of the generator from routine-dose CT images. The performance of this discriminator was used as an adversarial loss for the generator. Experiments were performed using CT images of an anthropomorphic phantom containing calcium inserts, as well as patient non-contrast-enhanced cardiac CT images. The phantom and patients were scanned at 20% and 100% routine clinical dose. Three training strategies were compared: the first used only voxelwise loss, the second combined voxelwise loss and adversarial loss, and the third used only adversarial loss. The results showed that training with only voxelwise loss resulted in the highest peak signal-to-noise ratio with respect to reference routine-dose images. However, CNNs trained with adversarial loss captured image statistics of routine-dose images better. Noise reduction improved quantification of low-density calcified inserts in phantom CT images and allowed coronary calcium scoring in low-dose patient CT images with high noise levels. Testing took less than 10 s per CT volume. CNN-based low-dose CT noise reduction in the image domain is feasible. Training with an adversarial network improves the CNNs ability to generate images with an appearance similar to that of reference routine-dose CT images.**

***Index Terms*— Coronary calcium scoring, deep learning, generative adversarial networks, low-dose cardiac CT, noise reduction.**

I. Introduction

COMPUTED tomography is a widely used imaging modality, allowing visualization of anatomical structures with high spatial and temporal resolution. However, ionizing radiation inherent to CT acquisition continues to raise concerns about potential health hazards [1], [2]. Therefore, the past decade has seen a trend towards dose reduction in CT examinations, and typical dose levels for e.g. coronary CT angiography have been reduced from around 12 mSv in 2009 [3] to 1.5 mSv in 2014 [4].

A drawback of dose reduction is the increase in noise in reconstructed CT images, which may lead to large local deviations in HU values [5]. A range of methods have been proposed to reduce noise and artifacts in low-dose CT, while preserving important details in the image. Iterative reconstruction (IR) techniques have recently been adopted by all major CT vendors [6], [7]. IR techniques iteratively estimate the denoised underlying image and have facilitated high levels of dose reduction [8]. However, these methods require long processing times, dedicated hardware and the availability of projection data.

Besides techniques that operate in the sinogram domain, there is a rich tradition of denoising methods that operate in the image domain, i.e. after image reconstruction from projection data. Non-local means filtering methods estimate the noise component based on multiple patches extracted at different locations in the image [9] and have been widely used for CT denoising [10]. Recently, Green *et al.* [11] proposed a method which replaces noisy image patches by their nearest neighbors in a database of high SNR patches. Alternatively, diffusion filters have been used to sharpen edges and other structures [12].

More recently, several supervised machine learning techniques have been proposed for noise reduction in low-dose CT. Such methods learn a relation between the voxel value in a low-dose image I_{LD} and the voxel value at the same location in a corresponding routine-dose image I_{RD}, based on training with pairs of images. For example, Chen *et al.* [13] proposed a convolutional neural network (CNN) that estimated routine-dose HU values based on local patches in low-dose CT. This regression method was used to transform low-dose chest and abdomen CT images into estimates of the corresponding routine-dose images. Kang *et al.* [14], [15] followed a similar approach, but applied the CNN to a directional wavelet transform of the CT image.

The methods proposed in [13]–[15] showed good quantitative noise reduction properties. However, the parameters of the CNNs were optimized to minimize the per-voxel squared error between the reference routine-dose image and the denoised low-dose image, or wavelet decompositions of these. When estimating a routine-dose CT image, a voxel in the target image may have different possible values, as noise is not only

Manuscript received January 31, 2017; revised May 9, 2017; accepted May 23, 2017. Date of publication May 26, 2017; date of current version November 29, 2017. This work was supported by the Netherlands Organization for Health Research and Development (ZonMw) in the framework of the Innovative Medical Devices Initiative research programme, through the project FSCAD under Project 104003009. *(Corresponding author: Jelmer M. Wolterink.)*

J. M. Wolterink, M. A. Viergever, and I. Išgum are with the Image Sciences Institute, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands (e-mail: j.m.wolterink@umcutrecht.nl).

T. Leiner is with the Department of Radiology, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2017.2708987

present in the low-dose acquisition, but also in the routine-dose acquisition. Minimizing the squared error between reference and predicted voxel values causes the CNN to predict the mean of these values, resulting in smoothed images that lack the texture of a typical routine-dose CT image. This smoothing may limit the quantification of small structures in denoised images.

In this work, to overcome the limitations of voxel-wise regression in noise reduction, we propose to train a noise reducing *generator* CNN together with an adversarial *discriminator* CNN, as a generative adversarial network (GAN) [16], [17]. The discriminator CNN aims to differentiate between real routine-dose CT images and transformed low-dose CT images. The performance of this discriminator CNN adds a loss term to optimization of the generator CNN, forcing the generator to provide more realistic estimates of the routine-dose CT image. The discriminator CNN is only used to provide feedback during training, and thus adds no complexity at test time.

In addition, we address the limitation that spatially aligned low-dose and routine-dose CT images are required for training of a noise reducing CNN. Well-aligned clinical scans acquired at different dose levels are often not available. Therefore, previous works have resorted to simulation of low-dose CT images based on routine-dose images [11], [13], which is a challenging problem [18]. Here, we show that a generator CNN trained with only adversarial feedback can learn the appearance of routine-dose CT images, without spatially aligned low-dose and routine-dose images.

The method is demonstrated on CT scans of an anthropomorphic phantom acquired at low dose and routine clinical dose and non-contrast-enhanced cardiac CT scans of patients who were scanned with a low dose and a routine clinical dose. We quantitatively analyze the noise in filtered back projection (FBP) reconstructed images at these two dose levels, as well as in low-dose FBP images transformed by our method, and show that the method has strong noise reducing properties while preserving the texture in the CT scan. The proposed solely image-based method is compared to a conventional iterative reconstruction method requiring projection data.

Cardiac CT images without contrast enhancement are clinically used to quantify coronary artery calcification (CAC), which is a strong and independent predictor of cardiovascular events [19]. To quantify CAC, connected components in the coronary arteries above 130 HU are identified. Recent guidelines recommend CAC quantification in patients at low-to-intermediate risk according to traditional cardiovascular risk factors [20]. Hence, there is a strong clinical need for low-dose calcium scoring CT scans [21]. However, excessive noise levels in low-dose CT scans make it difficult if not impossible to limit quantification to CAC lesions only. This may cause large overestimation of CAC or render images non-interpretable [22]. While noise reduction offers a solution to this problem, strong noise reduction may negatively affect the quantification or identification of small and low-density CAC lesions [23], [24]. We show how the proposed use of an adversarial network improves CAC quantification over standard FBP at low-dose and over a generator trained without an adversarial network. Furthermore, we show that the proposed method allows CAC quantification in patient low-dose CT images.

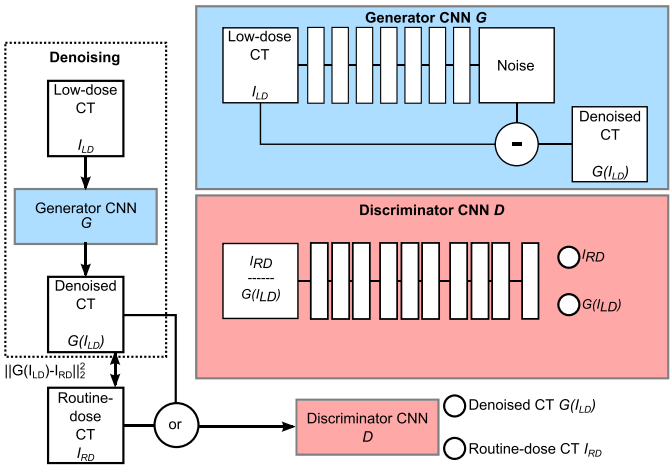


Fig. 1. Overview of the proposed pipeline for noise reduction in low-dose CT. The generative adversarial network consists of two components: a generator CNN and a discriminator CNN. The generator uses regression to determine the routine-dose HU value at every voxel in a low-dose CT. It does this through a skip connection which subtracts an estimated noise image from the input low-dose image. The discriminator tries to distinguish reduced noise CT images from real routine-dose images.

II. Methods

Fig. 1 illustrates the proposed system, which has two parts. The first part consists of a generator CNN G that analyzes a low-dose CT image I_{LD}. The generator returns $G(I_{LD})$, which is an approximation of a routine-dose CT image I_{RD}. The system has two ways to enforce similarity between $G(I_{LD})$ and I_{RD}. First, if voxels in I_{LD} and I_{RD} are spatially aligned, the voxel-wise error between $G(I_{LD})$ and I_{RD} may be minimized. Second, a discriminator CNN D may be simultaneously trained to differentiate between $G(I_{LD})$ and I_{RD}. If the discriminator can make this distinction easily, i.e. if the generated CT images do not resemble real routine-dose images, the generator needs to improve its estimations.

Hence, both networks have different tasks. While the generator performs regression of voxel values, the discriminator performs classification of images.

A. Generator CNN

The generator CNN G transforms the low-dose CT image I_{LD} into an image with a reduced noise level $G(I_{LD})$ approximating the reference routine-dose image I_{RD}. We assume that $I_{LD} = I_{RD} + N$, i.e. the noisy image is the reference routine-dose image with superimposed noise N. Hence, the task of the trainable layers in the CNN can be simplified to prediction of the noise N.

Cardiac CT images are typically anisotropic, with larger voxel spacing in the craniocaudal direction. Therefore, the input to the generator CNN was a 3D rectangular volume of $65 \times 65 \times 19$ voxels. The network contains seven consecutive convolution layers with small convolution kernels of

size $3 \times 3 \times 3$ voxels [25]. The number of kernels increases from 32 in the first layers, to 64 and 128 in the final layers. No padding is applied after convolutions. Hence, a receptive field of $15 \times 15 \times 15$ voxels in the input determines the result for a voxel in the output, which has size $51 \times 51 \times 5$ voxels. The final layer outputs the predicted noise through a linear activation function. The noise is then subtracted from the low-dose CT image to return a noise-reduced image $G(I_{LD})$.

All trainable layers except the final layer use leaky rectified linear activation functions (LReLUs) to increase training stability [17], [26]. Because the noise values are relatively small compared to the range of possible CT HU values, we initialize the weights in the convolution layers to a normal distribution ($\mu = 0.0, \sigma = 0.001$). Batch normalization [27] is used in the generator CNN, but not directly after the input layer or directly before the output layer.

B. Discriminator CNN

The discriminator takes either a routine-dose CT subimage I_{RD} or a processed low-dose CT subimage $G(I_{LD})$ as input, and determines whether the input is a real routine-dose image or not.

The input to the discriminator is a 3D rectangular volume of $51 \times 51 \times 5$ voxels, which is the size of the generator's output. The first two convolution layers use $3 \times 3 \times 3$ convolution kernels, which reduce the size of the volume to $47 \times 47 \times 1$ voxels. Hence, subsequent layers operate on 2D feature maps and consequently use 3×3 convolution kernels. Convolution layers are organized in three blocks. Each block contains two layers without strided convolution, and one layer with a stride of 2 [28]. The resulting feature maps are connected to an output node through a hidden layer with 256 nodes. As in the generator, we use LReLu activation functions and batch normalization. The final layer contains a sigmoid activation to determine whether the input is a real routine-dose CT image (label 1) or not (label 0). Weights in the discriminator network are initialized using the method proposed in [29].

C. Training

The system in Fig. 1 contains output $G(I_{LD})$ of the generator and outputs $D(G(I_{LD}))$ or $D(I_{RD})$ of the discriminator. Two loss components directly affect the generator. First, the squared error between $G(I_{LD})$ and I_{RD}. Second, the ability of the discriminator to identify images generated by the generator. Hence, the loss for the generator is defined as

$$\ell_G = \lambda_1 ||G(I_{LD}) - I_{RD}||_2^2 + \lambda_2 \ell_{bce}(D(G(I_{LD})), 1), \quad (1)$$

where $\ell_{bce}(D(G(I_{LD})), 1)$ is the binary cross-entropy between the discriminator's prediction and the label 1 that the generator wants the discriminator to predict. The parameters λ_1, λ_2 determine a weighting between the regression error of the generator and the classification error of the discriminator.

The discriminator has an adversarial goal to the generator and wants to correctly distinguish the processed low-dose CT images from the routine-dose CT images. Hence, the discriminator minimizes

$$\ell_D = \ell_{bce}(D(I_{RD}), 1) + \ell_{bce}(D(G(I_{LD})), 0), \quad (2)$$

where $\ell_{bce}(D(I_{RD}), 1)$ is the binary cross-entropy between the discriminator's decision on routine-dose CT samples and their target label 1, and $\ell_{bce}(D(G(I_{LD})), 0)$ is the binary cross-entropy between the discriminator's decision on processed low-dose CT samples and their target label 0.

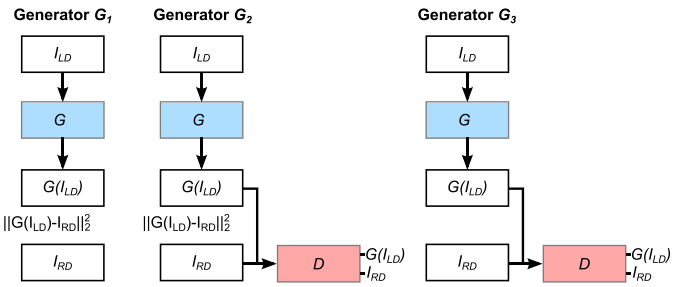


Fig. 2. Three different approaches to CNN-based low-dose CT noise reduction. Generator G_1 is trained only with the squared error loss between the reduced noise image and the routine-dose image. Generator G_2 is trained with the squared error and the discriminator feedback. Generator G_3 is trained with only the discriminator feedback.

The loss of the generator (Eq. 1) can be optimized in three ways. First, by only incorporating the voxel-wise loss between $G(I_{LD})$ and I_{RD}, i.e. $\lambda_1 > 0, \lambda_2 = 0$, as was previously proposed in [13] and [14]. Second, by combining the voxel-wise loss with the loss incurred through the discriminator ($\lambda_1 > 0, \lambda_2 > 0$). Third, by optimization based only on the discriminator feedback ($\lambda_1 = 0, \lambda_2 > 0$). In this case, an L2 regularization term on the noise map is added to the loss function. For the first two strategies, voxel-alignment between I_{LD} and I_{RD} is required, while no such alignment is required for the third strategy. We name the resulting trained generator CNNs G_1, G_2 and G_3 (Fig. 2).

All relevant parameters in the generator and discriminator are simultaneously optimized using the Adam optimizer [30], with a learning rate of 0.0002 and an exponential decay rate for the first moment estimates 0.5 [17]. While related methods on GAN-based image transformation proposed to alternate between optimization of the generator and the discriminator during training [31], [32], we found that simultaneously optimizing both networks led to more stable behavior of the system. The method was implemented in Theano and Lasagne. Experiments were run on a NVIDIA Titan X GPU with 12 GB memory.

III. Data

We include low-dose and routine-dose CT scans of an anthropomorphic thorax phantom, as well as low-dose and routine-dose patient cardiac CT scans.

A. Phantom CT Scans

An anthropomorphic thorax phantom (QRM anthropomorphic thorax phantom; QRM GmbH; Möhrendorf; Germany) with a central recess was used. This recess was filled with

water, in which one of two artificial coronary arteries was placed. The first coronary artery contained two inserts with densities of 196 and 380 mg hydroxyapatite (HA)/cm^3, the second coronary artery contained two inserts with densities of 408 and 800 mg HA/cm^3. All inserts had the same dimensions, and a volume of 196.3 mm^3. The phantom was embedded in an extension ring made of tissue equivalent material to simulate CT acquisition of average-sized patients.

The phantom was scanned on a Philips Brilliance iCT 256 scanner (Philips Healthcare, Best, The Netherlands), with a tube voltage of 120 kVp. Images were acquired in sequential mode at 20% routine dose (10 mAs) and 100% routine dose (50 mAs) [33]. The phantom was consecutively scanned at these two dose levels, so that low-dose and routine-dose images were spatially aligned. After the phantom was scanned once at both dose levels, a small translation and rotation were applied to the phantom. This process was repeated five times for both artificial arteries. Hence, the image set contained five acquisitions per dose level per artificial artery. Images were reconstructed using filtered backprojection (FBP) as well as an intermediate level of iterative reconstruction (iDose^4 level 3; Philips Healthcare, Best, The Netherlands) to a slice thickness and increment of 3.0 mm, matching the parameters for cardiac CT reconstruction. In-plane resolution was 0.49 mm.

B. Cardiac CT Scans

Non-contrast-enhanced cardiac CT scans of 28 patients were previously obtained in a study which was approved by the local institutional review board and for which written informed consent from all participants was obtained [34]. Patients were scanned on a Philips Brilliance iCT 256 scanner (Philips Healthcare, Best, The Netherlands), using a tube voltage of 120 kVp and a tube current of either 10/50 mAs or 12/60 mAs, depending on the patient's weight (threshold $\geq$ 80 kg). Hence, two scans were acquired: one at 20% and another at 100% routine cardiac dose, with an effective dose of 0.2 or 0.9 mSv. Image acquisition was ECG-triggered. The low-dose and routine-dose image of a patient were not spatially aligned. All images were reconstructed using FBP as well as an intermediate level of iterative reconstruction (iDose^4 level 4; Philips Healthcare, Best, The Netherlands) to slice thickness 3.0 mm and slice increment 1.5 mm. In-plane resolution ranged from 0.35 to 0.49 mm.

C. Evaluation

Noise levels were characterized using the mean and standard deviation of HU values in a homogeneous region of interest (ROI). The Friedman test was used to analyze noise differences among reconstructions and the Wilcoxon signed-rank test with Bonferroni correction was used to analyze pairwise differences.

The correspondence between voxel-aligned images was quantitatively evaluated using the peak signal-to-noise ratio (PSNR). In the case of CT, the PSNR is defined as

$$\mathrm{PSNR} = 20 \log_{10} \frac{4095}{\sqrt{MSE}}, \tag{3}$$

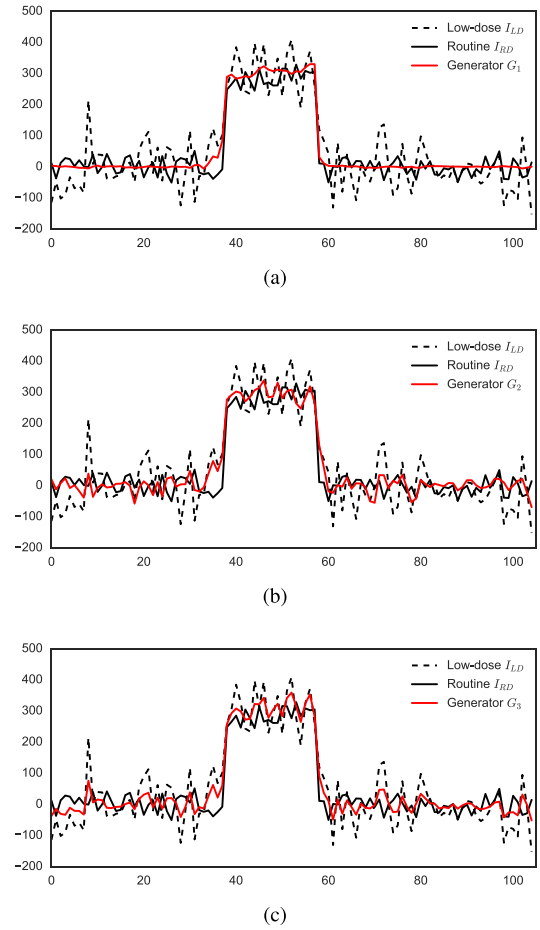


Fig. 3. Reducing noise in a synthesized 1D low-dose signal I_{LD} with generators G_1, G_2 and G_3. Generator G_1, which was trained without an adversarial discriminator, predicts smooth signals with low standard deviation, which do not resemble the synthetic routine-dose signal I_{RD}. The generators that were trained with an adversarial discriminator predict noisy values in the same distribution as the synthetic routine-dose signal I_{RD}.

where 4095 is the maximum range of HU values, and MSE is the mean squared error between two images. In all cases, the routine-dose FBP image was used as the reference standard.

The effect of noise reduction on the quantification of coronary artery calcification was analyzed. In the phantom CT images, we quantified the volume (in mm^3) and mass (in mg) of the four inserts. In the patient scans, we quantified total calcium burden using the Agatston score, which is a clinically used intensity-weighted measure of calcified area [35]. In all cases, a clinically used threshold of 130 HU was used for calcium quantification.

IV. Experiments and Results

A. Noise Reduction

To investigate the ability of the method to reduce noise, we performed experiments using synthesized 1D signals, as well as the 3D phantom and patient cardiac CT images described in the previous section.

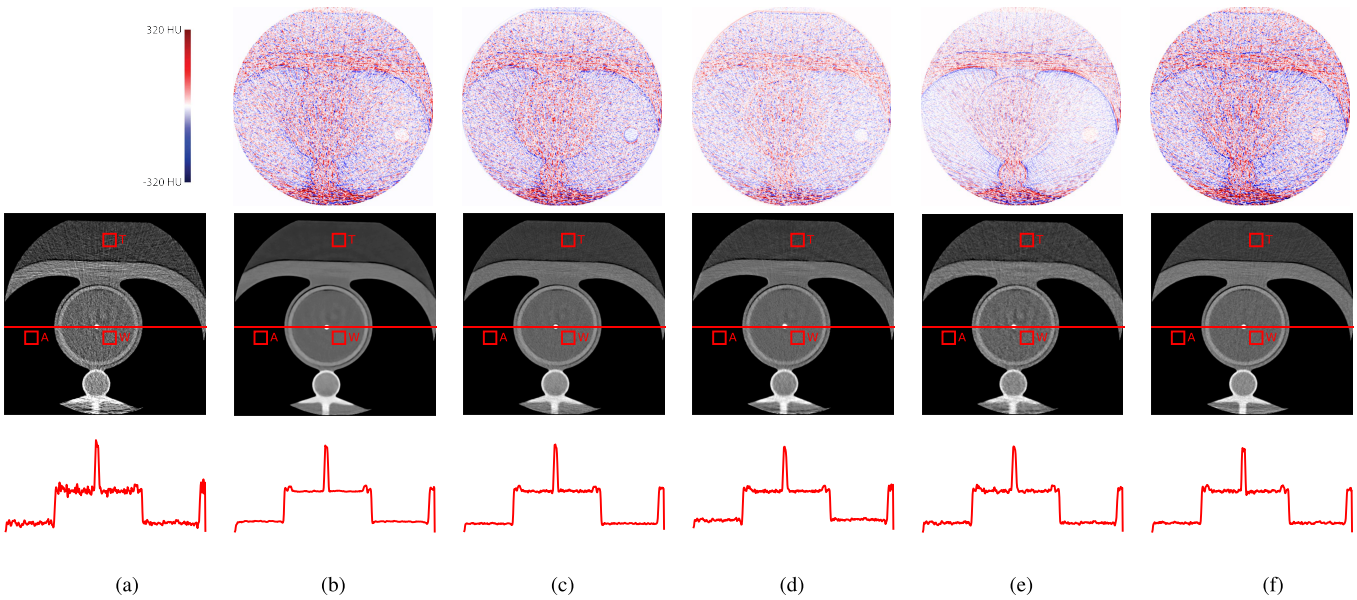


Fig. 4. Cardiac CT phantom with insert. (a) FBP-reconstructed low-dose CT image acquired at 10 mAs (I_{LD}), (b) low-dose CT image processed by generator G_1, (c) low-dose CT image processed by generator G_2, (d) low-dose CT image processed by generator G_3, (e) iteratively reconstructed low-dose CT image *IR* (I_{LD}), and (f) reference routine-dose CT acquired at 50 mAs (I_{RD}). Difference images with respect to I_{LD} are shown in the top row. Horizontal red lines in the images indicate intensity profiles, shown below the image. Red squares indicate air (A), tissue equivalent material (T) and water (W) ROIs used for noise measurement. All images have window level/width 90/750 HU.

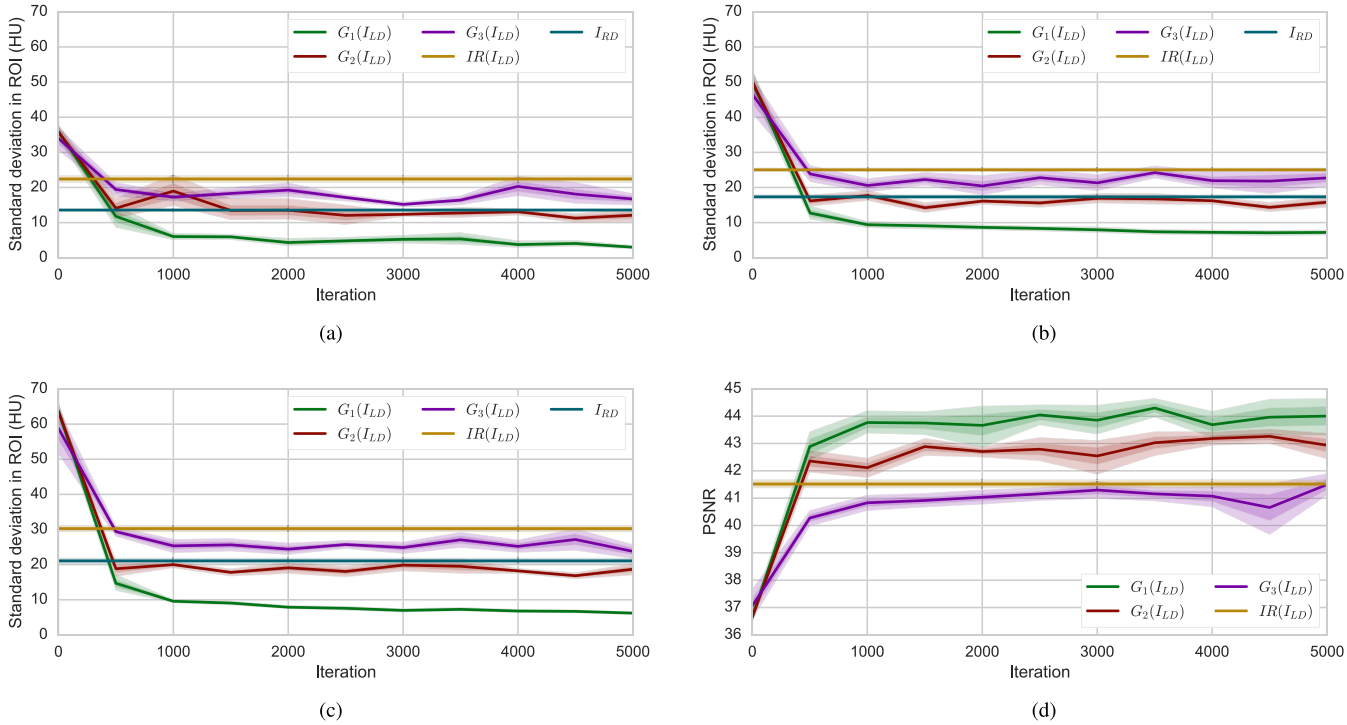


Fig. 5. (a),(b),(c) Standard deviation in homogeneous air, tissue equivalent material and water ROIs in the phantom CT scans, for generators G_1, G_2 and G_3 during generator training. The lines for *IR* (I_{LD}) and I_{RD} indicate the noise levels in low-dose IR reconstructions and routine-dose FBP reconstructions, respectively. (d) Peak-signal-to-noise ratio (PSNR) between generated images and routine-dose FBP during generator training. In all cases the average of five-fold cross-validation is shown.

1) Synthetic 1D Signals: To illustrate the characteristics and differences among generators G_1, G_2 and G_3, we performed experiments with synthesized 1D signals. Synthesized 1D training signals of length 119 were initialized to 0 HU. At a random point in the signal, a block activation with a random length and value between 0 and 500 HU was added. The resulting signal was considered the base signal. To synthesize low-dose and routine-dose signals for generators G_1 and G_2, Gaussian noise with $\mu = 0$ and $\sigma = 70$ or $\sigma = 20$ was added to the base signal. These values for σ were based on ROI

measurements in the phantom CT scans. To train generator G_3, separate base signals were acquired for the low-dose and routine-dose signal. The architecture of the generator and discriminator CNN was the same as described in the previous section, but all 2D and 3D operations were replaced by 1D operations. All generators were trained for 5,000 iterations with mini-batches of 48 samples. For generator G_2, the voxel-wise and adversarial loss were balanced with $\lambda_1 = 0.001$ and $\lambda_2 = 1.0$.

Fig. 3 shows the synthetic low-dose signal I_{LD}, the target synthetic routine-dose signal I_{RD} and the noise-reduced low-dose signals $G_1(I_{LD})$, $G_2(I_{LD})$ and $G_3(I_{LD})$. Signal $G_1(I_{LD})$ shows that the generator has learned to smooth the signal, and to predict values with low standard deviation. In contrast, generators G_2 and G_3 have both learned to reduce the noise level in the signal to that of the routine-dose signal I_{RD}.

2) Phantom CT Scans: Voxels in the phantom scans were aligned between images acquired at different dose levels. Hence, voxel-wise loss could be used during training. Generators G_1, G_2 and G_3 were each trained using five-fold cross-validation. Each fold contained one scan with 196 and 380 mg HA/cm^3 inserts, and one scan with 408 and 800 mg HA/cm^3 inserts. The generators were trained to transform 10 mAs low-dose CT images into 50 mAs routine-dose CT images. Each network was trained for 5,000 iterations, with mini-batches of 48 samples. For generator G_2, the voxel-wise and adversarial loss were balanced with $\lambda_1 = 0.001$ and $\lambda_2 = 1.0$.

Fig. 4 shows axial images of FBP, G_1, G_2, G_3 and IR reconstructions of the phantom scanned at low-dose, as well as an FBP reconstruction of the phantom scanned at routine-dose. It can be seen that the low-dose FBP reconstruction in Fig. 4a shows substantial noise, with local deviations up to 70 HU. Generator G_1 (Fig. 4b) generates a smooth image with low noise levels. Generators G_2 (Fig. 4c), G_3 (Fig. 4d) and iterative reconstruction (Fig. 4e) generate a slightly noisier image with a noise profile that better matches that of the reference image (Fig. 4f). Note that in the images generated by G_1, G_2 and G_3 a ring artifact remains visible after noise reduction. This artifact has a larger scale than the receptive field of the generator and discriminator and is thus not removed. Likewise, this artifact remains visible in the IR image. The difference images with respect to I_{LD} in Fig. 4 reveal that generator G_3 and IR apply more limited noise reduction than the reference, and that IR tends to enhance edges in the phantom image.

During training, reconstructions for G_1, G_2 and G_3 were obtained after every 500 iterations. In each reconstruction, the noise level was determined in three homogeneous $32 \times 32 \times 2$ voxel ROIs in the central recess (red squares in Fig. 4) containing air, tissue equivalent material or water. Fig. 5 shows the evolution of the standard deviations in these ROIs during training, where additional lines indicates the noise distribution in low-dose IR and reference routine-dose FBP images I_{RD}. For all generators, noise levels are initially the same as in the low-dose CT scan. While the standard deviation in scans reconstructed by G_1 continues to decrease throughout training, the standard deviation in scans reconstructed by G_2 plateaus around the level of the routine-dose scan. For G_3, the noise plateaus above the level of the routine-dose scan, indicating a less accurate estimation of I_{RD} at approximately the same level as iterative reconstruction. Fig. 5d shows the average PSNR during training. The PSNR is highest for scans generated by generator G_1, which minimizes the squared error between the low-dose and routine-dose image. Generator G_3 and iterative reconstruction show similar PSNR values.

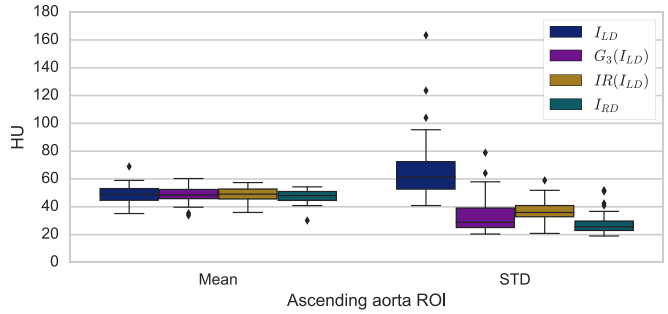
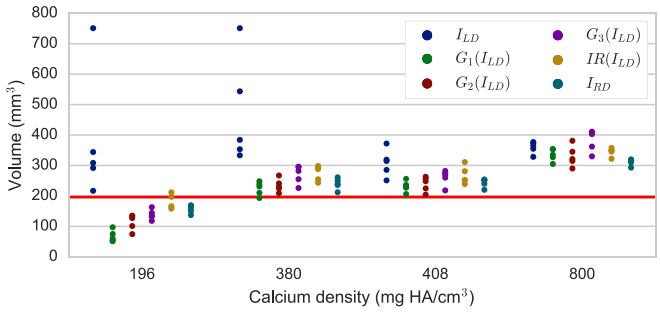


Fig. 6. Mean and standard deviation (STD) in an ROI in the ascending aorta at the level of the left coronary ostium in patient cardiac CT images. Results are shown for low-dose images reconstructed with FBP I_{LD}, generator G_3 ($G_3(I_{LD})$), iterative reconstruction ($IR(I_{LD})$), and routine-dose FBP reconstructions (I_{RD}).

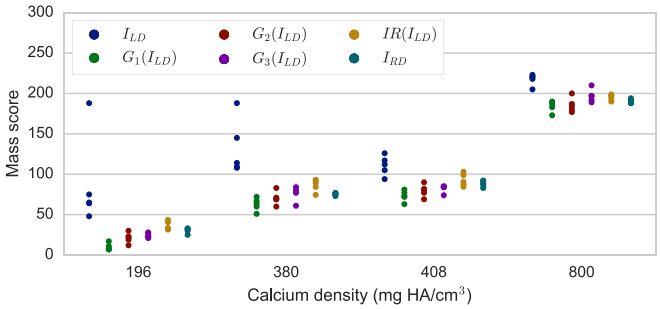
3) Cardiac CT Scans: Clinically acquired low-dose and routine-dose cardiac CT images share the same field-of-view and are in principle rigidly aligned. Nevertheless, we may not expect perfect voxel-wise correspondence between scans acquired at different dose levels, due to motion and breathing of the patient. Image registration may mitigate this problem to some extent, but also adds unwanted transformations to the image and noise patterns. Therefore, generator G_3 was trained using only adversarial loss.

The set of 28 patient scans was separated into two sets of 14 patients for a two-fold cross-validation, where patients were once in the training set and once in the test set. The generator was trained to transform low-dose images into routine-dose images. Training was performed as in the phantom scans, for 5,000 iterations with mini-batches of 48 3D samples.

To quantify image characteristics in each reconstruction, an ROI was placed in the ascending aorta at the level of the left coronary ostium and the HU mean and standard deviation in this ROI were determined (Fig. 6). There was a significant difference among mean HU values ($p = 0.03$) in different images, but post-hoc analysis with the Wilcoxon signed-rank test ($\alpha = 0.008$) did not reveal statistically significant differences between any two reconstructions. Likewise, there was a significant difference among the standard deviations in different images ($p<0.001$), with a median (IQR) value of 61.4 (52.5–73.9) HU for I_{LD}, 28.9 (24.7–41.3) HU for $G_3(I_{LD})$, 35.9 (32.7–41.2) for $IR(I_{LD})$, and 25.6 (22.9–30.4) for I_{RD}. Pairwise post-hoc analysis with the Wilcoxon signed-rank test revealed that noise was significantly higher in I_{LD} than in each of the other reconstructions ($p < 0.001$ in all comparisons). Conversely, noise levels in I_{RD} were significantly lower than in any of the other reconstructions ($p < 0.001$ in all comparisons). There was no significant difference between G_3 and IR ($p = 0.03$).



(a) Volume



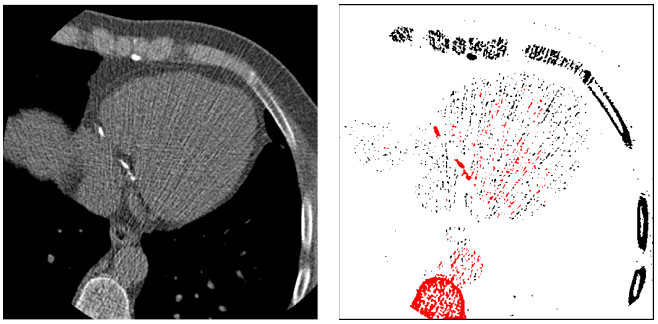
(b) Mass

Fig. 7. (a) Volume and (b) mass quantification of coronary calcification inserts in low-dose (I_{LD}) and routine-dose (I_{RD}) FBP images of the phantom, as well as low-dose IR reconstructions ($IR(I_{LD})$) and images reconstructed using generators G_1, G_2 and G_3.

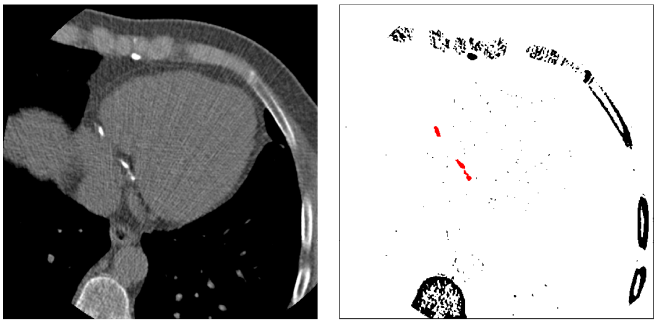
B. Coronary Calcium Quantification

1) Phantom CT Scans: In each phantom reconstruction, the volume and mass of two inserts were quantified. The inserts were identified and segmented using a clinically used threshold of 130 HU and 26-connectivity region growing [35]. Fig. 7 shows the volume and mass of each of the four inserts in different reconstructions. The reference volume of each insert was 196.3 mm^3, indicated by the red line. In the low-dose FBP reconstruction, the volume and mass of the inserts were systematically overestimated due to noise $\geq$ 130 HU in the image. In one acquisition, both inserts were connected to each other by voxels $\geq$ 130 HU. The three generators and IR reduced noise in a way that led to more accurate quantification of calcium volume. However, generator G_1 trained using only voxel-wise square error led to the largest underestimation of the volume of the insert with the lowest density. Similarly, the mass score was most underestimated when G_1 was used. Likely due to partial volume effects, the volume of the 800 mg HA/cm^3 inserts was overestimated in all reconstructions [36].

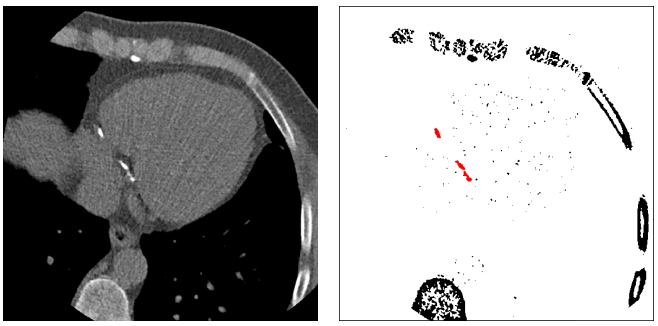
2) Cardiac CT Scans: Fig. 8 shows a low-dose cardiac CT FBP image, the same image transformed by generator G_3 and IR, and the routine-dose FBP image of the same patient. For each image, a CAC candidate mask is shown, indicating all voxels $\geq$ 130 HU. In the low-dose FBP image, calcium scoring is infeasible due to large clusters of connected voxels with density $\geq$ 130 HU (shown in red). However, after noise reduction with generator G_3, the distribution of noise voxels is similar to that in the routine-dose scan, and the CAC lesions remain above the density threshold. Noise reduction with iterative reconstruction shows a similar effect. Note the difference between the anatomy visualized in the low-dose scan and the routine-dose scan, making voxel-wise spatial alignment between the images difficult.



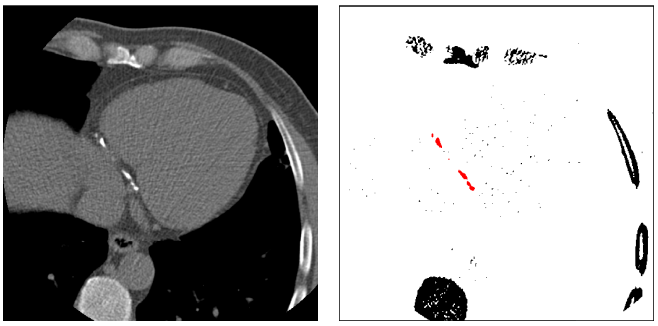
(a) I_{LD} (b) CAC mask I_{LD}



(c) $G_3(I_{LD})$ (d) CAC mask $G_3(I_{LD})$



(e) $IR(I_{LD})$ (f) CAC mask $IR(I_{LD})$



(g) I_{RD} (h) CAC mask I_{RD}

Fig. 8. Example CT slice for Patient 9 (Table I) of (a) 20% dose FBP reconstruction I_{LD} and (b) corresponding artery calcification (CAC) scoring mask, (c) 20% dose generator G_3 reconstruction $G_3(I_{LD})$ and (d) corresponding CAC scoring mask, (e) 20% dose generator IR reconstruction $IR(I_{LD})$ and (f) corresponding CAC scoring mask, and (g) routine-dose FBP reconstruction I_{RD} and (h) corresponding CAC scoring mask. All images have window level/width 90/750 HU. CAC scoring masks show all voxels $\geq$ 130 HU in black, and voxels selected by CAC scoring with connected component labeling in red.

TABLE I
AGATSTON SCORES IN LOW-DOSE IMAGES RECONSTRUCTED WITH FBP I_{LD}, GENERATOR G_3 ($G_3(I_{LD})$), ITERATIVE RECONSTRUCTION ($IR(I_{LD})$), AND ROUTINE-DOSE FBP (I_{RD}). DASHES INDICATE IMAGES IN WHICH CAC COULD NOT BE RELIABLY SCORED

Patient	I_{LD}	$G_3(I_{LD})$	$IR(I_{LD})$	I_{RD}
1	–	–	–	24.6
2	–	–	–	65.4
3	–	–	44.2	9.0
4	–	952.1	–	1065.8
5	–	20.6	36.5	92.7
6	–	76.8	71.3	94.9
7	–	367.1	405.6	379.1
8	–	294.8	302.6	394.7
9	–	1869.8	1991.9	2024.9
10	0	0	0	0
11	0	0	0	0
12	0	0	0	0
13	0	0	0	0
14	0	0	0	0
15	0	0	0	0
16	0	0	0	0
17	5.3	1.5	4.2	4.3
18	14.7	3.5	9.9	12.4
19	15.3	5.3	13.4	18.5
20	22.6	6.3	19.1	5.2
21	24.2	8.2	17.3	23.7
22	38.8	25.9	30.5	29.3
23	72.7	15.4	43.9	53.5
24	76.2	55.8	136.8	92.6
25	168.5	128.7	155.5	162.8
26	240.5	194.3	201.1	203.1
27	330.2	212	310.2	321.1
28	2097.9	2185.9	2166.3	2176.0

Coronary calcium was quantified in the low-dose FBP images, the images generated by generator G_3, iteratively reconstructed low-dose CT images and the routine-dose images (Table I). In 9 out of 28 low-dose FBP images, noise levels were too high to perform reliable calcium scoring, i.e. individual calcifications could not be identified in the CAC scoring mask. After noise reduction with generator G_3, noise levels were too high in 3 out of 28 images. These were the scans with the highest noise levels in low-dose CT FBP, with standard deviations in the aortic ROI of 163.4, 123.6 and 104.0 HU, respectively. Similarly, noise levels were too high in 3 out of 28 low-dose images reconstructed with IR. Calcium could reliably be scored in all routine-dose images. Out of 19 patients for whom calcium could be scored in I_{LD}, $G_3(I_{LD})$, $IR(I_{LD})$ and I_{RD}, 12 contained coronary calcium in all reconstructions. In these images, median (IQR) Agatston scores were 55.8 (20.8–186.5) in I_{LD}, 20.7 (6.1–145.1) in $G_3(I_{LD})$, 37.2 (16.3–166.9) in $IR(I_{LD})$, and 41.4 (17.0–172.9) in I_{RD}.

V. DISCUSSION

We have described a method to reduce noise in low-dose CT images using convolutional neural networks. The results show that training with adversarial feedback from a discriminator CNN can generate images with a more similar appearance to the routine-dose CT than training without a discriminator CNN. Feedback from the discriminator prevents smoothing in the image and allows more accurate quantification of low-density calcifications in phantom CT scans.

The results show that the proposed method is capable of substantial noise reduction in phantom CT images, and that combining a voxel-wise squared error loss with adversarial loss led to a noise distribution that was similar to that in the reference routine-dose image. Training with only squared error loss as proposed in [13] led to smooth images with low noise levels, while training with only adversarial loss led to images with slightly higher noise levels than in the routine dose. Hence, in practice it would be good to combine the two loss components when possible. A comparison to conventional iterative reconstruction at intermediate levels revealed that IR reduces noise, but not to the reference routine dose level. In future work, we will investigate if iteratively reconstructed images could be further processed using the proposed method.

In cardiac CT images, where voxel-wise spatially aligned training images are not available, we found that a generator trained with only adversarial loss was able to significantly reduce noise levels, while preserving mean tissue HU values. This shows that a convolutional neural network can learn to reduce noise in low-dose CT images even when no voxel-wise spatially aligned routine-dose scan is available. Voxel-wise accurate alignment of separate low-dose and routine-dose cardiac CT images is challenging due to poor image contrast, cardiac motion and large slice spacing. In previous studies, the problem of voxel alignment was mitigated by simulation of low-dose CT images based on routine-dose images [11], [13]. However, realistic low-dose CT simulation is a challenging problem, and simulated scans do not necessarily resemble real low-dose acquisitions [18]. In future work, we will investigate whether generator G_3 can also learn to reduce noise when the low-dose and routine-dose sets consist of different patients, i.e. if inter-patient learning is possible.

Calcium quantification in low-dose CT images of the phantom led to substantial overestimation of volume and mass. Calcium scores in the noise-reduced images were lower than those in the low-dose FBP reconstructions for all generators, as well as for the iteratively reconstructed images. The phantom calcium quantification results further indicate that the generator trained using only squared error applies too much smoothing to the image and affects quantification of low-density calcifications. Training with adversarial feedback resulted in calcium scores that were closer to those obtained in the routine-dose image.

In patient cardiac CT scans, noise reduction allowed calcium scoring in six scans that previously contained too much noise. This was similar to iterative reconstruction. There were 12 patients with coronary calcium whose images allowed scoring in I_{LD}, $G_3(I_{LD})$, $IR(I_{LD})$ and I_{RD}. For these patients, the obtained calcium scores in reduced noise images were lower than those in low-dose CT, which corresponds to our observation in the phantom images. Similarly, calcium scores obtained in $G_3(I_{LD})$ were generally lower than those in routine dose scans. This could be due to noise reduction, but may also be caused by interscan variability [37], and should be further investigated in a larger sample. The different reconstructions or dose levels only mildly affected mean HU values in an aortic ROI in the images. Hence, in all cases a clinical calcium threshold of 130 HU was used [35].

Smoothing effects in CNN-based image-to-image regression have been addressed in computer vision as well as medical image analysis. For colorization of grayscale natural images, Zhang *et al.* [49] proposed to pose the regression problem as a classification problem with one class for each potential pixel value. In our problem this would mean prediction of 4096 highly unbalanced classes, which is infeasible. Alternatively, GANs have been used for a wide range of image-to-image transformations, including image colorization, super-resolution [38], video frame prediction [39], segmentation [31] or general image-to-image conditioning [40]. In medical image analysis, Nie *et al.* [32] used a GAN to transform MRI images into estimates of CT images.

While methods for semantic image segmentation typically require CNNs with large receptive fields [41], [42], noise is localized. In the current approach, the generator has a relatively small receptive field of $15 \times 15 \times 15$ voxels. This is in line with patch-based methods such as presented in [11] and the finding of [5] that such receptive fields are sufficient to estimate the local extent of noise. In contrast, removal of streak artifacts caused by photon starvation may require CNNs with larger receptive fields [43]. The proposed generator CNN does not directly predict HU values in the routine-dose CT image, but predicts the amount of additive noise at each position in the image. An additional operation is performed to subtract this noise from the image. A similar approach was recently proposed by Han et al. *et al.* for subtraction of streak artifacts from CT images reconstructed with a low number of projections and by Kang *et al.* [15] for noise reduction in low-dose CT. Previously proposed GAN methods also used the input to G as input to D [31], [40]. We found that this led to a strong bias towards the discriminator's performance and was infeasible in the case of generator G_3.

In the proposed method, the discriminator network performs an auxiliary task during training and the network is not used during testing. However, after training, the discriminator has learned to extract certain features from low-dose and routine-dose non-contrast-enhanced cardiac CT images. In future work we will investigate if these features could be useful for other tasks in non-contrast-enhanced cardiac CT, such as automatic coronary calcium scoring [44]–[46].

A potential limitation of the current method is that pathologies might be introduced that are actually not present in the image, based on their presence in the training set. Therefore, in future work it would be interesting to estimate the certainty of the method at each location in the image. Furthermore, in the current study only a limited number of phantom training images were available to compare generators G_1, G_2 and G_3. An increase in training data may lead to slightly improved performance of the individual generators. However, the differences between the generators will likely stay the same, as they are trained to minimize different loss functions.

A major advantage of the proposed method is its processing speed. The discriminator CNN is only used during training, which restricts computational requirements during testing. The method has a runtime of less than 10 s on a $512 \times 512 \times 90$ CT volume, and may thus be efficiently applied to an already reconstructed low-dose scan, without the need for sinogram data. The current study focused on the quantification of coronary artery calcification in cardiac CT. However, there is high interest in development of low-dose CT scanning protocols for all body areas and diseases, especially in pediatric patients and young adults, since they are most susceptible to the deleterious effects of X-ray radiation. Potential applications include imaging of the brain, head and neck, chest, abdomen and pelvis [47], [48]. In future work, we will validate the proposed method on a variety of clinically relevant tasks. Finally, using the proposed method we aimed to translate low-dose CT images into routine-dose CT images. However, the method could likely be applied to translate between any other pairs of CT images.

VI. Conclusion

Low-dose CT noise reduction in the image domain using a convolutional neural network is feasible. Training with an adversarial network allows the generator to better learn the noise distribution in routine-dose CT and produce more realistic images for more accurate coronary calcium quantification.

References

[1] D. J. Brenner and E. J. Hall, "Computed tomography—An increasing source of radiation exposure," *New England J. Med.*, vol. 357, no. 22, pp. 2277–2284, Nov. 2007.

[2] M. S. Pearce *et al.*, "Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: A retrospective cohort study," *Lancet*, vol. 380, no. 9840, pp. 499–505, Aug. 2012.

[3] J. Hausleiter *et al.*, "Estimated radiation dose associated with cardiac CT angiography," *JAMA*, vol. 301, no. 5, pp. 500–507, 2009.

[4] M. H. Al-Mallah, A. Aljizeeri, M. Alharthi, and A. Alsaileek, "Routine low-radiation-dose coronary computed tomography angiography," *Eur. Heart J. Suppl.*, vol. 16, pp. B12–B16, Nov. 2014.

[5] J. Padgett, A. M. Biancardi, C. I. Henschke, D. Yankelevitz, and A. P. Reeves, "Local noise estimation in low-dose chest CT images," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 9, no. 2, pp. 221–229, Mar. 2014.

[6] L. L. Geyer *et al.*, "State of the art: Iterative CT reconstruction techniques," *Radiology*, vol. 276, no. 2, pp. 339–357, 2015.

[7] M. J. Willemink *et al.*, "Iterative reconstruction techniques for computed tomography part 1: Technical principles," *Eur. Radiol.*, vol. 23, no. 6, pp. 1623–1631, Jun. 2013.

[8] M. J. Willemink *et al.*, "Iterative reconstruction techniques for computed tomography part 2: Initial results in dose reduction and image quality," *Eur. Radiol.*, vol. 23, no. 6, pp. 1632–1642, Jun. 2013.

[9] A. Buades, B. Coll, and J.-M. Morel, "A review of image denoising algorithms, with a new one," *Multiscale Model. Simul.*, vol. 4, no. 2, pp. 490–530, 2005.

[10] Z. Li *et al.*, "Adaptive nonlocal means filtering based on local noise level for CT denoising," *Med. Phys.*, vol. 41, no. 1, p. 011908, Jan. 2014.

[11] M. Green, E. M. Marom, N. Kiryati, E. Konen, and A. Mayer, *Efficient Low-Dose CT Denoising by Locally-Consistent Non-Local Means (LC-NLM)*. Cham, Switzerland: Springer, 2016, pp. 423–431. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-46726-9_49

[12] A. M. Mendrik, E.-J. Vonken, A. Rutten, M. A. Viergever, and B. van Ginneken, "Noise reduction in computed tomography scans using 3-D anisotropic hybrid diffusion with continuous switch," *IEEE Trans. Med. Imag.*, vol. 28, no. 10, pp. 1585–1594, Oct. 2009.

[13] H. Chen *et al.*, "Low-dose CT via convolutional neural network," *Biomed. Opt. Exp.*, vol. 8, no. 2, pp. 679–694, 2017.

[14] E. Kang, J. Min, and J. C. Ye. (2016). "A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction." [Online]. Available: https://arxiv.org/abs/1610.09736

[15] E. Kang, J. Min, and J. C. Ye. (2017). "Wavelet domain residual network (WavResNet) for low-dose X-ray CT reconstruction." [Online]. Available: https://arxiv.org/abs/1703.01383

[16] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[17] A. Radford, L. Metz, and S. Chintala. (2015). "Unsupervised representation learning with deep convolutional generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1511.06434

[18] C. W. Kim and J. H. Kim, "Realistic simulation of reduced-dose CT with noise modeling and sinogram synthesis using DICOM CT images," *Med. Phys.*, vol. 41, no. 1, p. 011901, 2014.

[19] J. Yeboah *et al.*, "Comparison of novel risk markers for improvement in cardiovascular risk assessment in intermediate-risk individuals," *JAMA*, vol. 308, no. 8, pp. 788–795, 2012.

[20] D. C. Goff *et al.*, "2013 ACC/AHA guideline on the assessment of cardiovascular risk," *Circulation*, vol. 129, pp. S49–S73, 2014. [Online]. Available: https://doi.org/10.1161/01.cir.0000437741.48606.98

[21] K. B. Baron, A. D. Choi, and M. Y. Chen, "Low radiation dose calcium scoring: Evidence and techniques," *Current Cardiovascular Imag. Rep.*, vol. 9, no. 4, pp. 1–8, 2016.

[22] M. J. Willemink *et al.*, "Finding the optimal dose reduction and iterative reconstruction level for coronary calcium scoring," *J. Cardiovascular Comput. Tomogr.*, vol. 10, no. 1, pp. 69–75, 2016.

[23] J. A. van Osch *et al.*, "Influence of iterative image reconstruction on CT-based calcium score measurements," *Int. J. Cardiovascular Imag.*, vol. 30, no. 5, pp. 961–967, Jun. 2014.

[24] C. Gebhard *et al.*, "uller, E. Kazakauskaite, O. Gaemperli, "Coronary artery calcium scoring: Influence of adaptive statistical iterative reconstruction using 64-MDCT," *Int. J. Cardiol.*, vol. 167, no. 6, pp. 2932–2937, Sep. 2013.

[25] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[26] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. Int. Conf. Mach. Learn.*, vol. 30. 2013, pp. 1–6.

[27] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[28] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. (2014). "Striving for simplicity: The all convolutional net." [Online]. Available: https://arxiv.org/abs/1412.6806

[29] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Aistats*, vol. 9. 2010, pp. 249–256.

[30] D. P. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980

[31] P. Luc, C. Couprie, S. Chintala, and J. Verbeek. (2016). "Semantic segmentation using adversarial networks." [Online]. Available: https://arxiv.org/abs/1611.08408

[32] D. Nie, R. Trullo, C. Petitjean, S. Ruan, and D. Shen. (2016). "Medical image synthesis with context-aware generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1612.05362

[33] N. van der Werf, M. Willemink, T. Willems, M. Greuter, and T. Leiner, "Influence of dose reduction and iterative reconstruction on CT calcium scores: A multi-manufacturer dynamic phantom study," *Int. J. Cardiovascular Imag.*, vol. 33, no. 6, pp. 899–914, 2017.

[34] A. M. den Harder *et al.*, "Submillisievert coronary calcium quantification using model-based iterative reconstruction: A within-patient analysis," *Eur. J. Radiol.*, vol. 85, no. 11, pp. 2152–2159, 2016.

[35] A. S. Agatston, W. R. Janowitz, F. J. Hildner, N. R. Zusmer, M. Viamonte, and R. Detrano, "Quantification of coronary artery calcium using ultrafast computed tomography," *J. Amer. College Cardiol.*, vol. 15, no. 4, pp. 827–832, Mar. 1990.

[36] J. Dehmeshki, X. Ye, H. Amin, M. Abaei, X. Lin, and S. D. Qanadli, "Volumetric quantification of atherosclerotic plaque in CT considering partial volume effect," *IEEE Trans. Med. Imag.*, vol. 26, no. 3, pp. 273–282, Mar. 2007.

[37] A. Rutten, I. Išgum, and M. Prokop, "Coronary calcification: Effect of small variation of scan starting position on Agatston, volume, and mass scores," *Radiology*, vol. 246, no. 1, pp. 90–98, 2008.

[38] C. Ledig *et al.* (2016). "Photo-realistic single image super-resolution using a generative adversarial network." [Online]. Available: https://arxiv.org/abs/1609.04802

[39] M. Mathieu, C. Couprie, and Y. LeCun. (2015). "Deep multi-scale video prediction beyond mean square error." [Online]. Available: https://arxiv.org/abs/1511.05440

[40] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. (2016). "Image-to-image translation with conditional adversarial networks." [Online]. Available: https://arxiv.org/abs/1611.07004

[41] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. N. L. Benders, and I. Išgum, "Automatic segmentation of MR brain images with a convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1252–1261, May 2016.

[42] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, *Dilated Convolutional Neural Networks for Cardiovascular MR Segmentation in Congenital Heart Disease*. Cham, Switzerland: Springer, 2017, pp. 95–102.

[43] Y. S. Han, J. Yoo, and J. C. Ye. (2016). "Deep residual learning for compressed sensing CT reconstruction via persistent homology analysis." [Online]. Available: https://arxiv.org/abs/1611.06391

[44] J. M. Wolterink, T. Leiner, R. A. P. Takx, M. A. Viergever, and I. Išgum, "Automatic coronary calcium scoring in non-contrast-enhanced ECG-triggered cardiac CT with ambiguity detection," *IEEE Trans. Med. Imag.*, vol. 34, no. 9, pp. 1867–1878, Sep. 2015.

[45] N. Lessmann *et al.*, "Deep convolutional neural networks for automatic coronary calcium scoring in a screening study with low-dose chest CT," *Proc. SPIE*, vol. 9785, p. 978511, Mar. 2016.

[46] I. Išgum, A. Rutten, M. Prokop, and B. van Ginneken, "Detection of coronary calcifications from computed tomography scans for automated risk assessment of coronary artery disease," *Med. Phys.*, vol. 34, no. 4, pp. 1450–1461, Apr. 2007.

[47] R. D. A. Khawaja *et al.*, "Dose reduction in pediatric abdominal CT: Use of iterative reconstruction techniques across different CT platforms," *Pediatric Radiol.*, vol. 45, no. 7, pp. 1046–1055, Jul. 2015.

[48] L. N. Morimoto *et al.*, "Reduced dose CT with model-based iterative reconstruction compared to standard dose CT of the chest, abdomen, and pelvis in oncology patients: Intra-individual comparison study on image quality and lesion conspicuity," *Abdominal Radiol.*, pp. 1–10, 2017. [Online]. Available: https://link.springer.com/article/10.1007/s00261-017-1140-5, doi: 10.1007/s00261-017-1140-5.

[49] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Computer Vision*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 649–666.