

Learning Bone Suppression from Dual Energy Chest X-rays using Adversarial Networks

Dong Yul Oh¹ and II Dong Yun^{2,*}

¹Interdisciplinary Program in Bioengineering, Seoul National University, Korea

²Division of Computer and Electronic System Engineering, Hankuk University of Foreign Studies, Korea

*Correspondence: yun@hufs.ac.kr

ABSTRACT

Suppressing bones on chest X-rays such as ribs and clavicle is often expected to improve pathologies classification. These bones can interfere with a broad range of diagnostic tasks on pulmonary disease except for musculoskeletal system. Current conventional method for acquisition of bone suppressed X-rays is dual energy imaging, which captures two radiographs at a very short interval with different energy levels; however, the patient is exposed to radiation twice and the artifacts arise due to heartbeats between two shots. In this paper, we introduce a deep generative model trained to predict bone suppressed images on single energy chest X-rays, analyzing a finite set of previously acquired dual energy chest X-rays. Since the relatively small amount of data is available, such approach relies on the methodology maximizing the data utilization. Here we integrate the following two approaches. First, we use a conditional generative adversarial network that complements the traditional regression method minimizing the pairwise image difference. Second, we use Haar 2D wavelet decomposition to offer a perceptual guideline in frequency details to allow the model to converge quickly and efficiently. As a result, we achieve state-of-the-art performance on bone suppression as compared to the existing approaches with dual energy chest X-rays.

1 Introduction

Over twenty-thousand people die every year due to diseases related to the lung and its surroundings, such as chronic obstructive pulmonary disease (COPD), emphysema, and pneumonia¹. Radiologists first obtain chest X-rays in order to diagnose these pulmonary diseases, however, the ribs interfere with careful observation of the lesions, which frequently occurs near parenchyma, heart, peritoneum, etc. except for musculoskeletal system. Previous studies by^{2,3} have proved that lung cancer lesions located behind ribs potentially have key features associated with abnormalities. In addition, most patients, particularly those who need regular observation, are able to cope with more precise pathologic outcomes through the difference between the current image and the one previously recorded. The process for matching two images is required but the ribs could also disturb the diagnosis.

Currently, the commercialized method for acquisition of bone suppressed X-rays is dual energy imaging⁴, which captures two radiographs at a very short interval with different energy levels. It performs bone cancellation by exploiting subtraction between the attenuation of soft tissue and bone at different intensities. However, this method has a significant clinical defect in which the patient is exposed to the radiation twice and artifacts arise due to heart beat between two shots. Although low-dose imaging techniques have been developed, it is rarely true that X-ray exposure does not increase the probability of causing other diseases such as skin cancer. Since heart beat is not a function that a human can temporarily stop, additional techniques are required to solve the artifacts caused by the heart movement. Furthermore, a specialized equipment, which is expensive to purchase and maintain, is required to obtain dual energy X-rays (DXRs). Other conventional techniques are limited in their performance because X-rays, technically radiographs, have a wide range of clinical settings in medical imaging, and inter-class variation is very high.

We therefore tackle this problem with a novel approach using deep learning based model to learn bone suppression on single energy chest X-rays from previously acquired dual energy chest X-rays. Similar problems have already been addressed by^{5–8}. As big data become readily available, most solutions adopt the architectures of such approach as existing family of convolutional auto-encoders⁹. They have optimized the network parameters to minimize the *average* pixel-wise difference (with some other designed pixel-related functions) between the prediction and its ground truth. This is very straightforward and easy for the model to converge, however the bone suppressed images are quite blurry due to the nature of minimizing average pixel values, which we will discuss by comparing with our approach in Section 4.1 and 4.3.

Inspired by the recent success of the deep generative models^{10–13}, we fundamentally focus not only on de-noising approach that considers bone as a noise but also learning conditional probability distribution of bone suppressed image respect to its original one. The approach of¹² is the closest to ours in using Generative Adversarial Networks (GANs)¹⁴. The objective function to optimize the model parameters is the amount of noise, Euclidean distance between pairwise outputs and labels,

which is equivalent to other previous approaches. Here we add an adversarial training framework to maintain the sharpness of specific lesions on single energy X-rays and avoid undesirably suppressing them. The key difference from¹² is the choice of improved techniques to leverage a finite set of data based on the original GAN framework.

1.1 Main Contributions

This work first of all addresses the problem of minimizing average pixel-wise differences to learn bone suppression on single energy chest X-rays. Existing conditional adversarial networks of¹² is purposely modified to accomplish such a goal. Our contributions are summarized as:

- This work experimentally verifies that adversarial training framework for modeling de-noising approach with conditional image-to-image translation on bone suppression is able to outperform existing state-of-the-art methods.
- We propose to explicitly exploit frequency details using Haar 2D wavelet decomposition to offer a perceptual guideline for minimizing pairwise image differences.
- To the best of our knowledge, the model discussed in this paper is the first approach using deep generative models for bone suppression with DXRs, which has been rigorously evaluated.

1.2 Related Work

The present work is a partial solution of bone suppression on chest X-rays improving pathologic outcomes of both computer-assisted diagnosis (CAD) and radiologists. Many recent efforts to address this problem have been proposed. All of them utilize their method to extract specific information of bones from given chest X-rays and recognize where to suppress.

Bone suppression was first introduced by¹⁵, removing the dominant effects of the bony structure within the X-ray projection and reconstructing residual soft tissues components. Most of general studies in relation to bone suppression received relatively less attention and have been conducted for very specific purpose until the actual clinical effect from bone suppression has been verified. However,² proved that currently learned diagnosis suffers from lung cancer lesions obscured by anatomical structures such as ribs, and³ showed that the superposition of ribs highly affects the performance of automatic lung cancer detection. Both studies re-examined the invisibility of abnormalities caused by the superposition of bones and the improvement of automatic or human-level pathologic classification by the detection of these abnormalities.

Since then, great progress has been done in bone suppression. We categorize them into deep learning and non-deep learning approaches. For non-deep learning approaches, one of the most sensational method that received much attention in medical fields is dual energy imaging⁴. It also refers to dual energy subtraction (DES) since it acquires information about specific intensities through a series of subtractions between two X-rays at different energies. Both images at different energies have different attenuation values, hence they can be subtracted to perform bone or tissue cancellation that is able to detect the lesion such as a calcified nodule that did not appear in either of them.¹⁶ employed Active Shape Model, which is a parametric model of a curve for bones where the parameters are determined from the statistics of many sets of points in similar images, then the segmentation data is used to remove bones by subtraction.¹⁷ followed a similar curve fitting model to get rib segments obtained through Gabor filtering, and used several pre-processing from CAD, local contrast enhancement and lung segmentation.¹⁸ refined the final ribs with the dynamic programming-based active contour algorithm. The key aspects of these previous methods are detecting the position of lung and ribs border first and finally refining the final rib shadows based on vertical intensity profiles.

As deep learning algorithms are further developed, current related studies focus more on deep learning based model on bone suppression.⁵ used a massive artificial neural network, which the sub regions of input passes linear dense layers with single output, to obtain the bone image from a single energy chest X-ray. Then they subtract the bone image from the original image to yield virtual dual energy image, similar to a soft-tissue image.⁶. the extension model of⁵, additionally employed a total variation-minimization smoothing method and multiple anatomically specific networks to improve previously achieved performance. A new approach combined with deep learning and dual energy X-rays data has been commonly used recently;⁷ trained with 404 dual-energy chest X-rays with a multi-scale approach, and also subtracted the bone image from the original image to obtain a virtual soft tissue image using its vertical gradient as previously introduced.⁸ proposed two end-to-end architecture, convolutional auto-encoder network and non-down-sampling convolutional network that directly output the bone suppressed images based on DXR training set. They combined mean squared error (MSE) with the structural similarity index (SSIM) that addresses sensitivity of the human visual system to changes in local structure¹⁹.

Such a naive adoption of convolutional auto-encoder families often fails to capture the sharpness since the network misses high frequency details, which are the main reason of blurry images, in its encoding and decoding system.⁹ have overcome this limitation and achieved high performance on segmentation task with skip connection in the auto-encoding process. The segmentation task can be addressed by creating mask with its pixel-wise probability, however, with an intensity profile in the bone suppression task can potentially act as a bias.¹² employed very heuristic loss function using conditional GAN framework

for image translation similarly to neural style transfer. The success of such approach motivates us to do research on more effective and easier method not only to converge on learning bone suppression from a finite set of DXRs, also eliminate bias in suppressing region. We combine the suppressing noisy bones approach with image-to-image translation and purposely re-designed existing conditional adversarial network; the input system and improved techniques in the training process.

2 Background

2.1 Generative Adversarial Networks

This study aims to learn bone suppression on single energy X-rays from previously acquired DXRs through de-noising approach with conditional image-to-image translation. We use adversarial training within GAN framework¹⁴ to learn the conditional probability distribution of the output (bone suppressed X-ray images) according to the input (original X-ray images).

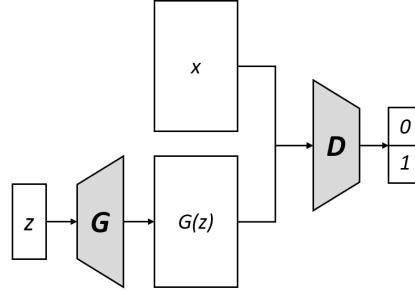


Figure 1. The overall schematic of Generative Adversarial Networks.

GAN is a generative model that consists of two networks called *generator* and *discriminator* in an adversarial relationship. The generator creates an image similar to the training set, and the discriminator distinguish whether the input is a fake image, which comes from the generator, or a real one coming from the training set. As depicted in Figure 1, the GAN is a structured probabilistic model. The generator is a differentiable function G , which basically takes latent variable z for the prior information of the model, then outputs the samples $G(z)$ that are intended to be drawn from the same distribution as observed variables x . Here z is regarded as random noise of which sampling method is generally taken in commonly known distribution such as Gaussian or exponential family. The discriminator is a differentiable function D which is a binary classifier taking both x and $G(z)$ and outputs a single probability for either case, $D(x)$ or $D(G(z))$. The discriminator thereby is trained with two mini-batch datasets for real and fake samples unlike the usual case in traditional supervised learning. In this scenario, two networks compete; the discriminator strives to make $D(x)$ to be near 1 while $D(G(z))$ to 0, which can be derived from binary cross-entropy using sigmoid function. Thus, the cost function of the discriminator is as follows:

$$J^{(D)}(\theta^{(D)}, \theta^{(G)}) = -\frac{1}{2}\mathbb{E}_{x \sim p_{data}} \log D(x) - \frac{1}{2}\mathbb{E}_{z \sim p_z} \log(1 - D(G(z))) \quad (1)$$

where $\theta^{(D)}$ and $\theta^{(G)}$ are the parameter of generator and discriminator, respectively. (1) offers extremely huge penalty if the discriminator does not properly distinguish both cases. This algorithm often refers to the *game theory* competing the participants (players), where the player's cost is dependent each other and each player cannot control the other player's parameters, hence GAN framework is called adversarial training. The simplest solution is a *Nash equilibrium* corresponding to the $G(z)$ being drawn from the same distribution as the training data x , and $D(x) = 0.5$ for all x in this scenario. This is also regarded as a zero-sum game or minimax game that the goal is for the sum of the players' cost is to be zero. Therefore, the cost function for the generator is:

$$J^{(G)} = -J^{(D)} \quad (2)$$

However, this minimax game algorithm is very inefficient in an actual training process. Minimizing cross-entropy has been proven for its efficiency because the loss never saturates when the network fails to predict given problem. (2) intuitively shows that when the discriminator minimizes its cross-entropy, the generator maximizes the same cross-entropy. In other words, the gradient vanishing problem where the gradient saturates to 0, occurs in the generator and vice-versa. To end this, we maintain

the concept of minimizing the generator's cross-entropy instead of flipping the sign and re-design the cost function for the generator as the cross-entropy of the generated image.

$$J^{(G)} = -\frac{1}{2} \mathbb{E}_{z \sim p_z} \log D(G(z)) \quad (3)$$

Now the generator maximizes the discriminator being mistaken unlike previously introduced minimax game where the generator strives to minimize the discriminator being correct. This is a very heuristic method to maintain a strategy of minimizing the existing cross-entropy without a disadvantage to the generator in the actual training process. This game is no longer zero-sum game; all players have a strong gradient when the opponent is losing the game however can be considered in a cooperative relationship since each player grows further to lead growing opponent being mistaken. This is equivalent to the maximum likelihood estimation under the assumption that the discriminator is optimal. The expected gradient of this function is equal to the expected gradient of $D_{KL}(p_{data}||p_g)$ since the problem is approximate the true data distribution by G . Note that minimizing KL-divergence between the training data and the model is equivalent to maximum likelihood.

To theoretically yield the global optimum of GAN, we first take the value function, $V(D, G)$ that specifies the discriminator's payoff in zero-sum game framework. Note that (3) is a heuristic mechanism to improve the actual training process. Therefore, the value function in this scenario is represented as minimization and maximization in an inner loop and outer loop, respectively.

$$\min_G \max_D V(D, G) = \min_G \max_D -J^{(D)}(\theta^{(D)}, \theta^{(G)}) \quad (4)$$

Next we take the derivative of (4) respect to a single entry $D(x)$ to obtain the optimal discriminator. In this process, the constants are ignored in advance and the expected values are formalized as integral. Let the probability distribution of real data and fake data created from the generator be denoted by p_{data} and p_g respectively. Since $G(z)$ is derived from latent variable z and desired to resemble true data x , the cross-entropy for G which is denoted by $D(G(z))$ can be re-written as $D(x)$ where x is belong to $p_g(x)$. The optimal case for the discriminator can then be computed as:

$$\max_D V(D, G) = \int_x p_{data}(x) \log D(x) + p_g(x) \log(1 - D(x)) dx \quad (5)$$

$$D^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)} \quad (6)$$

It is intuitively obvious that an optimal case for this scenario is $p_g(x) = p_{data}(x)$ because the generator creates the samples that are intended to be drawn from the same distribution as training data x , which would mean that the generator maximizes the discriminator being mistaken for distinction between true data $x \sim p_{data}$ and generated data $x \sim p_g$. Thus, the probability that the discriminator distinguishes either case is equal to 0.5 ($D(x) = 0.5$) if the generator correctly learns the distribution of true data. Note that the assumption that the discriminator is optimal is required to obtain the lower bound of this optimal case for the generator. All these can be derived by taking (6) into (5) and considering the JS-divergence (7).

$$D_{JS}(p_{data}||p_g) = \frac{1}{2} D_{KL}\left(p_{data} \parallel \frac{p_{data} + p_g}{2}\right) + \frac{1}{2} D_{KL}\left(p_g \parallel \frac{p_{data} + p_g}{2}\right) \quad (7)$$

$$\min_G V(D^*, G) = \int_x p_{data}(x) \log \frac{p_{data}(x)}{p_{data}(x) + p_g(x)} + p_g(x) \log \frac{p_g(x)}{p_{data}(x) + p_g(x)} dx \quad (8)$$

By solving the equivalence between (7) and (8),

$$\min_G V(D^*, G) = -\log(4) + 2 \cdot D_{JS}(p_{data}(x)||p_g(x)) \quad (9)$$

Finally, the optimal point for (4) is $p_g(x) = p_{data}(x)$ which refers to $D_{JS}(p_{data}||p_g) = 0$, hence $p_g(x)$ minimizing (8) has a distribution similar to $p_{data}(x)$.

Maximum likelihood estimation is the way we want to achieve high probability in all ranges where true data appears. Note that this is equivalent to minimizing cross-entropy such as (1), as described in (4). GANs are still in such estimation, however, behave in a way to get low probability in areas where true data does not appear. It shows the main difference from minimizing KL-divergence and that JS-divergence (9) is rather similar to reverse KL-divergence. The choice of divergence has not clearly explained why GANs make sharper samples, but they have received more attention as they outperform the existing generative models minimizing pixel-wise differences.

2.2 Image-to-Image Translation

As previously introduced in Section 2.1, the GAN approximates the maximum likelihood using a metric of JS-divergence through sampling without explicitly defining the probability model.¹⁴ introduced GAN frameworks with the aims to obtain the generator mapping z which is the latent variable, to the high dimensional space of observation x . Inspired by this strong ability that simply learns the distribution of x by competing the generator and discriminator, compared to previous generative models, many approaches using other sources instead of z that was recently proposed.

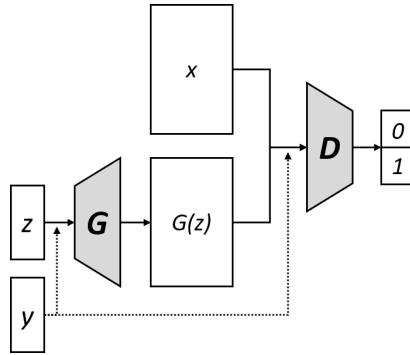


Figure 2. The overall schematic of Conditional GANs. The key difference from the original one is conditioning the networks, in which random noise z with the source data y as condition is transferred to the target data domain through the generator.

They are specifically called domain-to-domain translation including text, images, audio signals and etc. with conditional probability model that generates a target when given a source. As depicted in Figure 2, it is optional to use the random noise, z , but the generator and discriminator's job does not change; The generator is trained to give out the output that cannot be distinguished from target images by the discriminator, which is trained to do so. Note that most of the time, it is desirable to observe the source image y for the discriminator to complete conditional probability model in adversarial training framework. Therefore, the value function in this scenario is as follows:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x,y \sim p_{data}} \log D(x, y) + \mathbb{E}_{y \sim P_{data}, z \sim p_z} \log(1 - D(y, G(y, z))) \quad (10)$$

where x is target data, and y is source data according to x . To further improve the performance of the generator, the most common way is to use a traditional loss minimizing the distance between the source image mapped to the target domain, and its reference image, hence the model finds the properties to which they are linked between given domains providing data in pairs.

$$L_1 = \mathbb{E}_{x,y \sim p_{data}, z \sim p_z} \|x - G(y, z)\|_1 \quad (11)$$

$$G^* = \arg \min_G \max_D V(D, G) + \lambda L_1 \quad (12)$$

The generator not only fool the discriminator but also minimize L1 or L2 distance from the ground truth within pairwise data. The choice of using random noise z does not significantly contribute to learning conditional probability, however the model would loss stochasticity and only produce deterministic output if z is not used. This is previously employed and attempted by^{12,20,21}, but the effectiveness of random noise clearly depends on given problem type. Thus, the final objective generator of the generator is described in (12).

If pairwise data is not available, manually the feature is often determined for re-mapping to the target domain after the source is mapped to the low dimensional latent space, which suffers over-fitting. However¹³ proposed unpaired image-to-image translation using cycle consistency where the source image transferred to the target domain is able to be returned its original domain. This approach uses very heuristic mechanism particularly in a situation where the acquisition of pairwise data is labor-intensive, but the performance for the image quality is lower than the one that uses the pairwise data.

3 Method

In this chapter, we introduce our method for bone suppression using specifically designed GAN. As mentioned in previous section, the GAN approximates the intractable maximum likelihood using a metric of JS-divergence through sampling the latent variable from commonly known distribution, without explicitly defining the probability model. However, the definition of the sampling space does not fundamentally contribute to our problem since obtaining the output according to the input can be regarded as conditional image translation. A pair of the X-ray images with ribs and those with no ribs are available due to previously acquired data via DES. Therefore, L1-distance between the predicted value and the actual value for the bone suppressed image can practically guide the distribution learning with GAN. This guidance has theoretically global-convergence as the GAN approach, however, is unlikely to work a main objective function in training process. It is typically used in a weighted manner to assist the other criteria because it is one of the pixel-related functions that reduces the average difference of input and output. Here we use additional support mechanism to outperform existing state-of-the-art methods.

3.1 Haar 2D Wavelet Decomposition

Wavelet is a signal of the form firstly introduced by²² where a short localized oscillation repeats near zero and slowly vanishes. The wavelet is designed to have specific properties that are useful for signal processing; the convolution between wavelets and the target signal extracts certain information in a frequency or time domain. The principle can be described as the wavelet resonates if the target signal and the wavelet have the same frequency. The convolution of the signal to be analyzed with such wavelets is very similar to the Fourier Transform for examining the frequency band of a certain part of the signal. This is called wavelet transform, which is the process of separating the signal into a set of specific wavelets that are obtained from shifting or scaling one basic wavelet basis function. Its application is not only for the signal processing, but also for time series analysis or digital control system. The key features of time-frequency analysis with the wavelet transform from Short Time Fourier Transform (STFT) is that it adaptively selects frequency band based on the characteristics of the signal. The time resolution of the wavelet transform differs depending on frequency bands, whereas the STFT has same resolution at all frequency bands. Therefore, since the sudden change of the signal such as noise is very visible in frequency changing and important for perceptual quality, wavelet transform is more effective. All these performances have been verified by²³⁻²⁵.

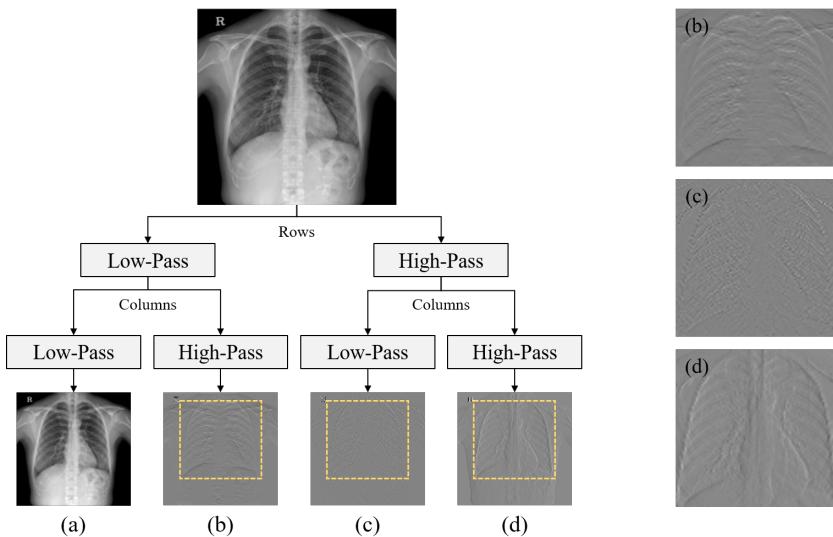


Figure 3. Haar 2D wavelet decomposition. The row direction in image is split into high-pass and low-pass sub-bands, then the column direction repeats this step. The decomposition results are put in four components; (a) sub-sampled original image, the directional feature images in (b) vertical, (c) horizontal, and (d) diagonal details.

We adopted Haar wavelet transform, which is one of the most popular wavelet transforms. Note that Haar wavelet is the basis wavelet in Haar wavelet transform and appears in square-shaped functions thereby is not continuous and differentiable. Haar transform using such wavelets can be used to analyze the localized feature of signals due to the orthogonal property. Our problem addresses two-dimensional signals, thus when the image is two-dimensionally wavelet-transformed, the high-frequency components are collected at the upper right and the low ones at the bottom left as shown in Figure 3. This is also regarded as 2D wavelet decomposition.

Frequency information obtained from wavelet decomposition have a very critical role in training deep neural network. In terms of successfully applied deep learning based applications, the main strength is to approximate complex source-to-target function with non-linearity when a large scale of training data is provided. The network learns the feature of interest without manually defining the features by human that often suffer from the lack of strong prior information of source and target domain. However, directly using normal X-ray images in our case can be more challenging for the neural network. Most of the time, it is desirable to provide conceptual hints instead of entirely relying on its neural system. It also pre-defines the features that the network should learn, which allows the model to converge more quickly and efficiently. This behavior has already been proven by²⁶ and its extension²⁷.

3.2 Network Architecture

The network architecture is based on Pix2Pix proposed by¹². The overall concept is equivalent to¹², which is that the generator minimizes pairwise difference and simultaneously attempts to fool the discriminator. In this process, GAN framework helps the network overcome the limitation by reducing the average error between input and output. In this study, we have added two purposely modified techniques to improve our specific task, bone suppression. First, as previous section introduced, we changed the input system from normal gray-scale X-ray images to wavelet decomposed X-ray images. This can efficiently decompose the directional components of X-ray, vertical, horizontal, diagonal frequency details to facilitate easier training of a deep network. Second, we partially modified training system in GAN framework, which will be further introduced in next section. The proposed model consists of the basic network in GAN; generator and discriminator. The architecture of the generator that receives the original image and produces bone suppressed images is depicted in Figure 4.

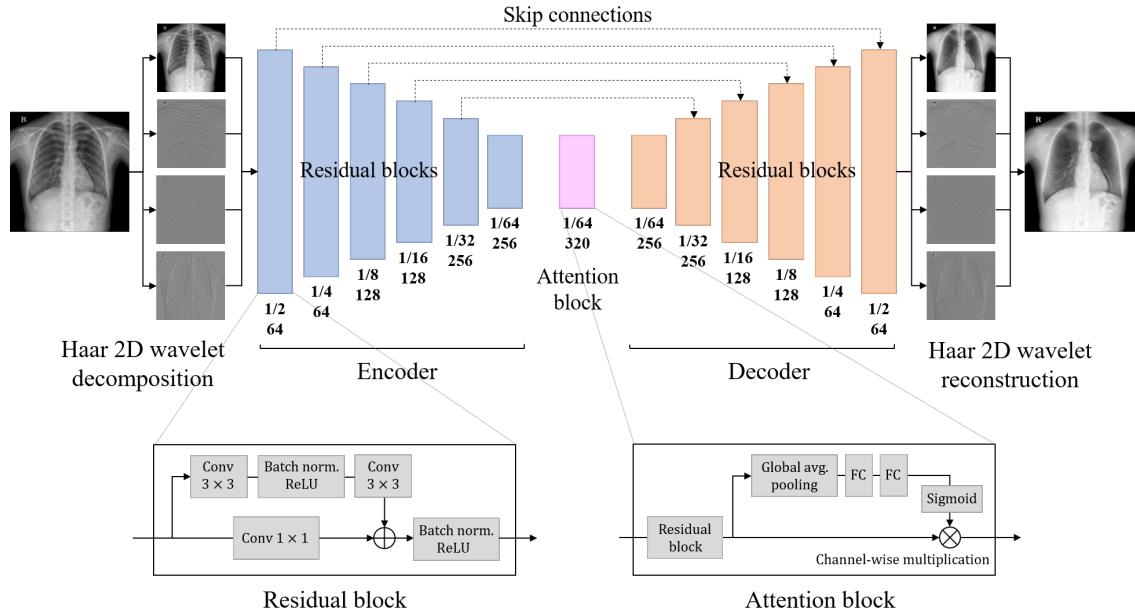


Figure 4. The architecture of the generator. The two values below each colored block represent the sub-sampling ratio respect to the original input size, and the output channels. The residual block enhances the gradient flow of the generator by shuttling the information to the next layer, and the last encoded feature finally receives self-attention through an attention block.

The generator takes the input size as 1024×1024 with gray-scale (1 channel) then converts the input to $512 \times 512 \times 4$ by Haar 2D wavelet decomposition and concatenating its results. As depicted in Figure 4, the overall architecture is based on convolutional auto-encoder with skip connections, which is regarded as U-Net⁹. The network consists of 12 residual blocks from²⁸ and an attention block (a squeeze and excitation block) firstly proposed by²⁹. The robustness of residual network, which

overcomes the limitation that deep networks are hard to train, have been proven in many computer vision tasks such as image recognition. Each residual block has two 3×3 convolution layers, and an additional 1×1 convolution layer that translates the input when changing the output channel. Translating the feature maps from shallower layer to following deeper layer has a critical role in training deep networks; it is rarely desirable for the deeper layer to directly fit the highly abstracted features, and such flow of the feature maps also improves gradient flow in back-propagation. In terms of the skip connections, the residual block in the encoder shuttles the high frequency information to its corresponding block in the decoder, thus the model can maintain the spatial frequency resolution and result in the sharp images. At the center of the network, a squeeze and excitation block is used for the attention mechanism facilitating the convergence of the model. This block summarizes all the feature maps through global average pooling, which is very important in the deep neural network where the local receptive field is small. The global spatial information is compressed into a channel descriptor and re-calibrated to calculate channel-wise dependencies.

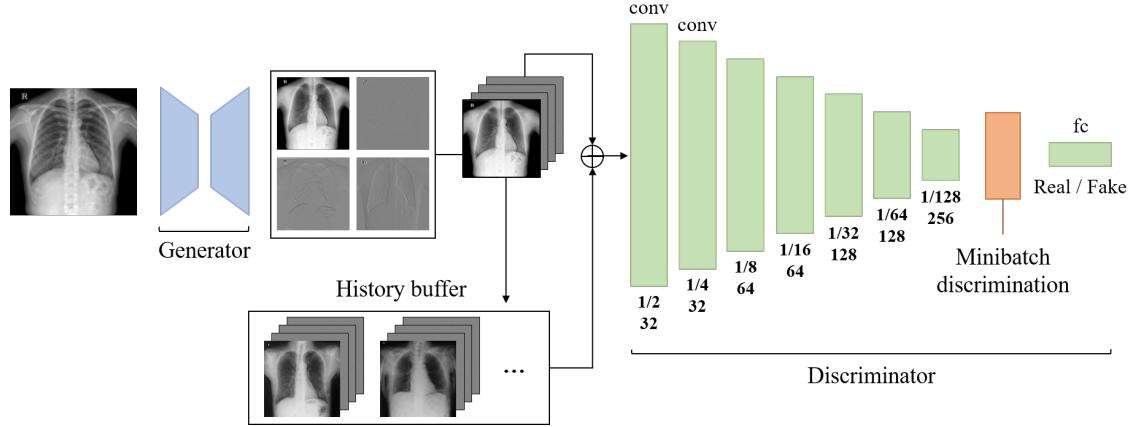


Figure 5. The architecture of the discriminator. The numbers below each convolution block is equivalent to those in Figure 4. The discriminator also takes the history of the generator's samples and considers the distribution of batch of images instead of the single image.

The discriminator contains 7 convolution layers and a fully connected layer to output a single probability whether given image is a fake image, which comes from the generator or not. Note that a stride in convolution operation is doubled instead of using a pooling layer. Maintaining the sharpness of other tissues by removing only the ribs in X-ray corresponding to the horizontal noise is still challenging while the bone suppressed image is blurry in general convolutional auto-encoder families. In this problem, the discriminator has the most important role; the degree to which the generator gets stronger (to trick the discriminator) depends on how we design the input that the discriminator looks. Therefore, we also took four components obtained by Haar 2D wavelet decomposition as the input hence the generator not only tries to make the four components shown in Figure 3 equal to those of the output, but also simultaneously avoid the blur to fool the discriminator. To make this more useful, we added history buffer and minibatch discrimination between the last convolution layer and the fully connected layer as depicted in Figure 5, improving both discriminator and generator.

3.3 Training

The discriminator and generator in the proposed method models are independently parameterized, and update the parameters by stochastic gradient descent based one their objective function (to minimize the cost function). The generator optimizes the Maximum Log-Likelihood Estimation (MLE) criteria previously described in (3) and the guidance term (11) with Haar 2D wavelet decomposed details. Note that maximizing the log likelihood in the logistic regression on both discriminator and generator is equivalent to minimizing their cross entropy. The discriminator also optimizes its MLE criteria in (1). Here we use Adam optimizer³⁰ with initial learning rate = 0.0008 and batch size = 8.

However, the GAN still fails to fully address mode collapse although it has grown dramatically in recent years. Mode collapse is when the generator creates similar samples only where the discriminator does not distinguish well. These samples are so-called ‘strange’ that the discriminator decided them as real and that the generator succeeds in tricking the discriminator, because such success does not consider the shape or texture that they have. This is primarily due to the loss function of the generator, which is a cross-entropy with its generated image focusing on images that are not well distinguished. In terms of adversarial frameworks, the discriminator network neither improves the generator by distinguishing all the given samples nor failing to distinguish them all, and often fail to converge. Thus, we need an equilibrium in their strength as long as using

adversarial framework. In order to solve these problems and improve learning convergence speed, recurrent optimization method that involves history buffer and minibatch discrimination are used.

3.4 History Buffer

The history buffer is a buffer that reflects the previous training results in the next training steps by the generator saving some images it has created. The wide range occurrence of the mode collapse in training process has a critical drawback; most of deep learning frameworks that do not use recurrent network such as Long Short-Term Memory (LSTM), apply the loss and the gradient calculation only respect to the currently given batch data. For this reason, the GAN frameworks also exhibits unstable learning because the discriminator forgets the past generation.

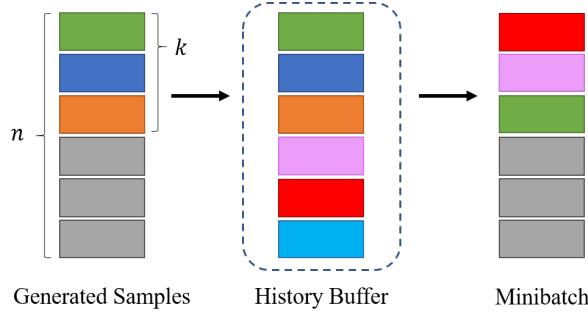


Figure 6. The illustration of history buffer that temporarily takes the half of generated samples in minibatch, and re-fills it with the samples randomly picked after shuffling the data.

This problem is not first addressed in this paper, and in particular the mechanism of using the history buffer has already been proposed by³¹. They noticed significant performance improvement depending on the presence of using a history of generated images. The authors of³¹ addressed that this lack of memory of the discriminator can cause divergence of the adversarial training, and lead the generator to re-introduce the artifacts that the discriminator had forgotten.

The history buffer simply takes k generated samples from $(x_{i1}, x_{i2}, \dots, x_{ik}, x_{ik+1}, \dots, x_{in})$, which is the output mini-batch in i -th step from the generator. Then randomly shuffling the data in the buffer, and the k -size of batch data in the buffer are popped and concatenated with the remaining $(x_{ik+1}, \dots, x_{in})$ thereby the batch size for training the networks is constant as depicted in Figure 6. Note that the size of the history buffer is $2k$, equivalent to batch size n , and such concatenation is available only when the buffer is full; i.e. the initialization starts with (x_{11}, \dots, x_{1k}) , then the mini-batch in i -th step finally looks like $(x_{r_11}, x_{r_22}, \dots, x_{r_nn})$ where $r = \{r_1, r_2, \dots, r_k\}$ is randomly picked from 1 to i -step. Now the Discriminator learns to distinguish all the samples from the corresponding buffer, which leads to more stable convergence of both networks and alternatively takes the same effect as recurrent optimization.

3.5 Minibatch Discrimination

Minibatch discrimination has been proposed by³², which simply transposes the feature maps to measure the distance between each feature map, thereby the discriminator network sees the distribution of images in given batch instead of a single image. Mode collapse often indicates that all outputs from the generator concentrates a single data point that the discriminator currently believes is highly realistic. Setting the discriminator to identify multiple samples is a straightforward solution to address this problem. It is also regarded as exploiting the dependency among generated images in mini-batch so that the discriminator can tell the outputs of the generator to become more dissimilar to each other.

The actual training process in an original architecture including general classification models or generative models, is to optimize the model based on the value of the objective function in mini-batch unit. Note that ‘mini-batch’ that we typically use for gradient descent indicates the average or the sum of individually calculated for each single data. Although most of time it is preferable to observe each data independently, our main purpose of using the adversarial training framework is to emphasize the sharpness of the image. In addition,³² shows that this minibatch discrimination mechanism does not work better in the task where the goal is to obtain a strong classifier in both supervised and semi-supervised learning.

Minibatch discrimination layer generally measures L1-distance between the batch of outputs that passed the last intermediate layer of the discriminator. Let the feature maps in i -th image in batch size of n be denoted by $f(x_i) \in \mathbb{R}^A, i \in \{1, 2, \dots, n\}$, where A is the number of output channel. In order to get the dependency between images represented as distance, it obtains the matrix $M_i \in \mathbb{R}^{B \times C}$ through multiplying $f(x_i)$ by any tensor vector (kernel) $T \in \mathbb{R}^{A \times B \times C}$ that will be optimized where B and C is the number of kernels and kernel size. Then it calculates the L1-distance between the rows of $M_{i,b}$ across the samples,

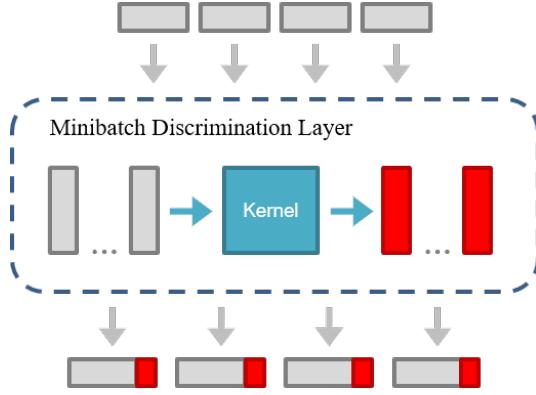


Figure 7. The illustration of minibatch discrimination layer multiplying a specific tensor vector, measuring the distance between samples, and concatenating the results to the input.

$b \in \{1, 2, \dots, B\}$ and finally applies a negative exponential $o(x_i) = \sum_{j=1}^n -e^{(\|M_{i,b} - M_{j,b}\|_1)} \in \mathbb{R}^B$. As a results, this layer yields as many inter-dependencies among batch images as the number of kernels. The authors of³² suggest to use the other samples as ‘side information’, thereby the output of minibatch discrimination layer is concatenated to the original feature maps on channel axis as depicted in Figure 7. The discriminator now distinguishes whether the input is a fake ‘batch’, or a real ‘batch’ from the training set, which allows much more visually realistic images than the one looking at a single image.

4 Experiment

4.1 Dataset

To verify the performance of the proposed model, we conducted experiments on the paired dataset of normal X-ray images and bone suppressed X-ray images via DES, which are regarded as DXRs (see Figure 8). It contained 348 patients for paired frontal-view chest X-rays and DXRs in total, and we randomly split the dataset into 80% for training, 10% for validation and 10% for test set. The dataset was originally released in DICOM format with 2017×2017 as each image size, and we rescaled them to 1024×1024 due to memory issue on GPU.

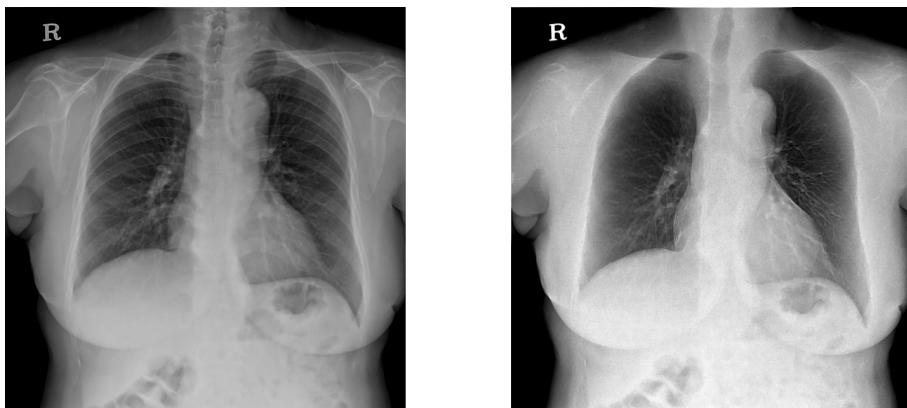


Figure 8. Sample data of bone suppressed X-ray image via DES (right) and its original image (left).

Since DICOM images exceed the commonly supported pixel dynamic range (from 0 to 255), it is preferable to select the specific dynamic range where the user tries to observe and linearly stretches the pixel intensities that lie within given range, to the original range. It is called linear windowing, and enables us to highlight bony structure rather than soft issue, or to highlight the abnormalities including lesions or at the expense of other structures present within the field-of-view. Thus, we use linear windowed images instead of a full dynamic range of images using windowing parameters provided in DICOM tags. We also normalize each image in the dataset that is subtracted by individually calculating the average of its pixels and dividing by the standard deviation.

As previously introduced in Section 1, dual energy imaging captures two radiographs at a very short interval with different energy levels to eliminate bone by subtraction between the attenuation of soft tissue and bone at different intensities. Therefore, the artifacts may arise due to heart beat between two radiographs. We manually examined the dataset since there was no post processing to handle this problem in acquisition of original images. 11 X-ray images were excluded from the training set and used for additional test which will be discussed in Section 4.3. In addition, this paper proposes to learn bone suppression on single energy X-ray by analyzing the pair of DXRs, and we only used the X-ray images at commonly known level of energy and discard those at lower energy.

4.2 Performance Metrics

We consider the following three objective image quality metrics to quantitatively evaluate the proposed method. Their advantage and drawbacks outlined below:

Peak Signal-to-Noise Ratio (PSNR): This metric measures the ratio between the maximum possible power of signal (pixel value) and the power of noise that corrupts the image and affects the fidelity of the image. It is an improved metric of Mean Squared Error (MSE) that does not reflect the image scale. i.e. the difference between 9 and 10 is that the pixel interval is ranging from 0 to 255 (8-bit) is more noticeable than the one ranging from 0 to 4096 (12-bit). In addition, it is often expressed in logarithmic scale due to various pixel dynamic range. Given a reference $m \times n$ image a and its approximation image b , we can obtain MSE and PSNR from the following definitions:

$$MSE = \frac{1}{mn} \sum_i^m \sum_j^n ||a(i,j) - b(i,j)||_2 \quad (13)$$

$$PSNR = 20 \log_{10} \left(\frac{\sqrt{MSE}}{MAX_a} \right) \quad (14)$$

where MAX_a is the maximum possible pixel value of the reference image.

Noise Power Spectrum (NPS) This metric gives a complete description of the noise with its amplitude over frequency resolution. It can be regarded as an improved metric of standard deviation within a specified region of interest (ROI), because the standard deviation does not consider the distribution of its noise according to frequency level. For NPS calculation, it is required to select ROI to characterize the noise correlations with 2D Fourier Transform:

$$NPS = \frac{1}{N_{ROI}} \sum_{i=1}^{N_{ROI}} \frac{1}{L_x L_y} ||FT_{2D}\{ROI_i(x,y) - \overline{ROI}_i\}||_2 \quad (15)$$

where L_x, L_y are the lengths of x and y dimension of ROIs, N_{ROI} is the number of ROIs used for NPS calculation, and \overline{ROI}_i is the mean pixel value of i -th ROI. Note that NPS represents the noise amplitude on Fourier space in the x and y dimension, not a single value. Since the result of (15) is a spectrogram, which is a 3D figure visualized in 2D by describing the amplitude over x and y dimensional frequency with color, it is common to average this NPS along 1D radial frequency to represent spatial resolution.

Structural Similarity Index (SSIM): This metric is proposed by¹⁹, also a full reference metric such as PSNR, in which the assessment of image quality relies on an initial noise-free image. However, it improves PSNR that measures absolute pixel-by-pixel errors, considering perceptual image degradation, luminance and contrast as human-perceived change in structural information; the pixels that are spatially close are likely to have strong inter-dependencies. Given a reference image a and its approximation image b , SSIM is defined as a product of luminance, contrast and structure functions:

$$SSIM = \frac{(2\mu_a\mu_b + c_1)(2\sigma_{ab} + c_2)}{(\mu_a^2 + \mu_b^2 + c_1)(\sigma_a^2 + \sigma_b^2 + c_2)} \quad (16)$$

where μ and σ^2 are the average and variance of corresponding image denoted by subscript, respectively. Note that σ_{ab} is the covariance of image a and b , and the constants c_1 and c_2 are set as $c_1 = (0.01L)^2$, $c_2 = (0.03L)^2$ by default where L is the dynamic range of pixel.

4.3 Quantitative results

In our overall bone suppression work-flow, we noticed the perceptual difference in the luminance due to the pixel value slightly exceeded the expected its dynamic range since there was no post-processing to adjust the pixel dynamic range of the output corresponding to its the normalized image. We could use histogram stretching, a process of simply increasing or decreasing the histogram when the images have the same contents. However, our problem takes the input as a general X-ray image and the output as a bone suppressed image. To handle this problem, we adopted histogram matching, which transforms the gray values corresponding to i -th cumulative histogram of the source image to have same one of the target image. The source image (bone suppressed image) and the target image (original image) in histogram matching are depicted in Figure 9. Since the difference between two images was the presence of the ribs, and the pixels with the closest difference in cumulative histogram was converted first, the bone suppressed image became more visually natural; the soft tissue that appeared relatively dark due to the intensities of the bones was brightened and vice versa. Note that our initial assumption of bone suppression was not designed for musculoskeletal diagnosis and most abnormalities are more likely to be found in soft tissues with lower intensity than bones. Therefore, we concluded that histogram matching as post-processing did not severely affect the image fidelity, however in future work, we would like to further verify this issue in clinical view.

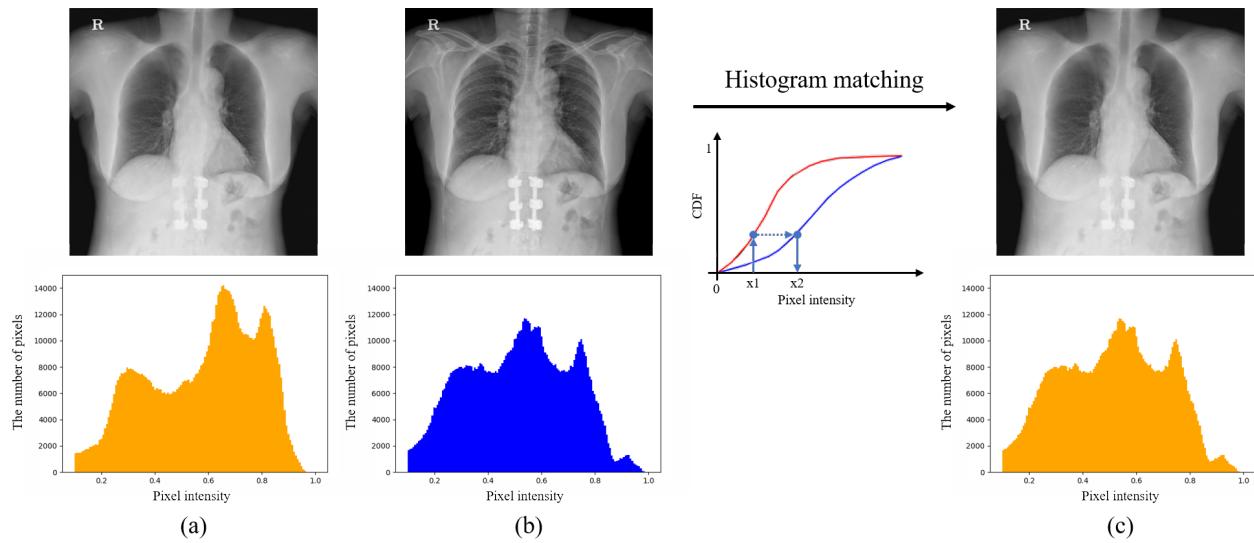


Figure 9. How histogram matching works and the perceptual difference changes (top row) as the pixel intensities changes (bottom row): (a) target image, (b) source image and (c) histogram matched source image. Note that the DC term is omitted in each histogram.

Finally, we conducted in total three trials of training the model, and selected one model with the best performance evaluated by 34 images in validation set. Then we measured the three metrics described in previous section using the test set. The sample experiments result with the proposed method can be found in Appendix. Since the region of interest on bone suppression is lung area, the evaluation of the entire image area and the lung area is carried out. Noise Power Spectrum (NPS) is calculated by manually extracting the 120×120 ROIs for the lung area in the error (noise) matrix between the prediction and its ground truth. In addition, we proceeded simple ablation studies about how much our purposely modified technique improves the performance on bone suppression; adoption of the main network architecture as GAN and the input system as Haar 2D wavelet decomposed frequency details. The method that we propose in section 3 outperformed the rest of the differently designed models as shown in Table 1.

The baseline of our study, convolutional auto-encoder (CNN), has the second highest performance on both PSNR and SSIM in the lung area where as the original PSNR is low due to the overall blurry image. The CNN+Haar Wavelets shows the worst SSIM, and its bone suppressed images are very blurry and even blood vessels in the lungs are not recognizable, which will be discussed in section 4.3.2. The CNN+GAN model shows that the PSNR results are not inferior to the baseline model, however very poor SSIM results because the adversarial training sharpens the image including the bones. This may increase human-perceived changes on the ribs, which have sudden difference in the pixel intensities. Therefore, not only better removal of the bones but also high visibility due to its sharpness affects the noise power in the high frequency bands, as depicted in Figure 10.

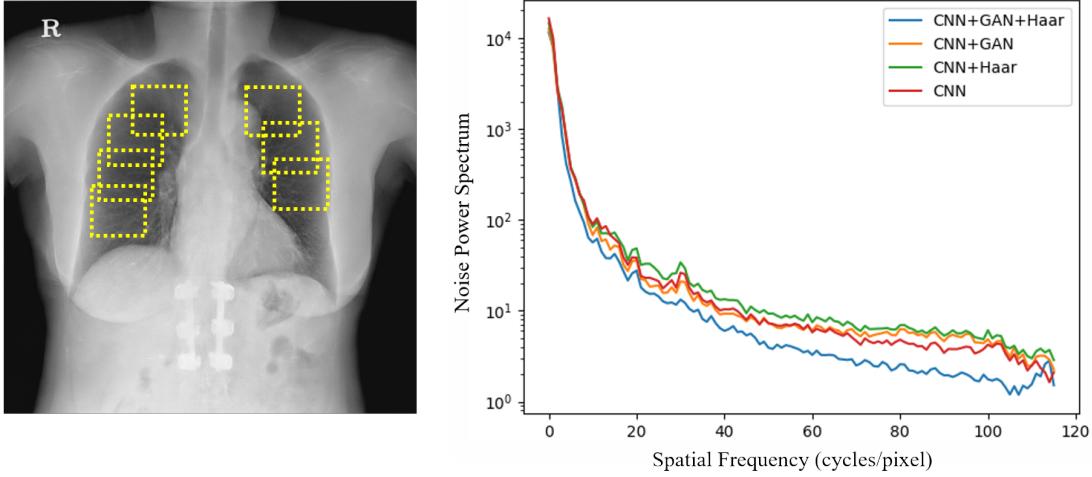


Figure 10. Sample ROI locations (left). Only 7 ROIs are shown for clarity, but 5 ~ 10 ROIs for each image are used and taken from the difference between the prediction and its ground truth. Average NPS is calculated across all patients in test set (right).

Table 1. The comparison of the performance with different conditions on the presence of purposely designed techniques in our problem.

Model	PSNR	PSNR (Lung)	SSIM (Lung)
CNN	19.229	26.350	0.9031
CNN + Haar Wavelets	22.289	25.840	0.7906
CNN + GAN	21.477	26.343	0.8496
CNN + GAN + Haar Wavelets (Ours)	24.080	28.582	0.9304

We also conducted bone suppression on the images that we manually excluded from the training set due to the conspicuous artifact. In this case, the ground truth obtained via DES can not be used as a reference image to evaluate the results. As shown in Figure 11, we observed, in a qualitative manner, that the motion artifacts due to heart beat did not appear and almost all information was maintained without blurry results. However, it still suffered from the lack of training data, which leads the model to often fail to capture the outline of the small blood vessels in the lungs and chest and remains further required extension of our study.

4.4 Analysis of Adversarial Training

The objective function where the discriminator distinguishes whether a given image is fake or real and the generator fools the discriminator not to do so, is very abstract. It works well even if we do not exactly define the features that we want the networks to learn in numerical form. In other words, we can only acknowledge that such features are one of style or patterns that the discriminator identifies as real. This can be solved by providing a reasonable guidance such as L1-distance to control a specific feature of interest, instead of visualizing the feature map or attention. In addition, many of GAN variants have shown sensational results beyond the pixel-related functions. When either cyclic consistency, the ability to return oneself with various domain, or the data pairs is available, it forces the training direction to make GAN converge quickly. In practice, this work verifies the quality of bone suppression using the adversarial training framework is able to outperform those with existing state-of-the-art methods.

4.5 Analysis of Haar 2D Wavelet Decomposition

Since our problem is de-noising the problem of considering the bone as a specific noise and removing only the bone, the bone suppression performance can be improved by providing a frequency details of the noise. Interestingly, we observed that the proposed input system, Haar 2D wavelet decomposition, works better only when used with adversarial training. As depicted in Figure 12, general convolutional auto-encoder with Haar wavelet decomposed information is blurrier and has less contrast. We

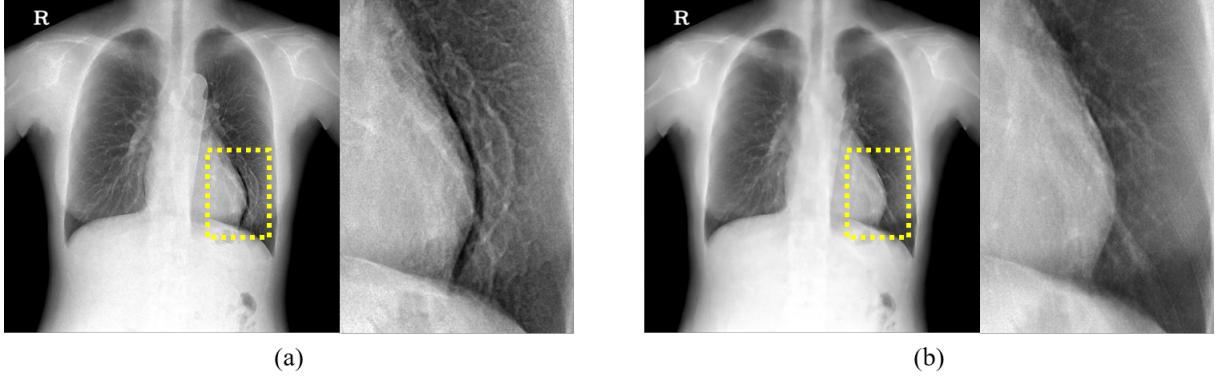


Figure 11. The example of artifacts due to temporal interval between two radiographs in DES (a) and the results of the proposed method to first radiograph (b).

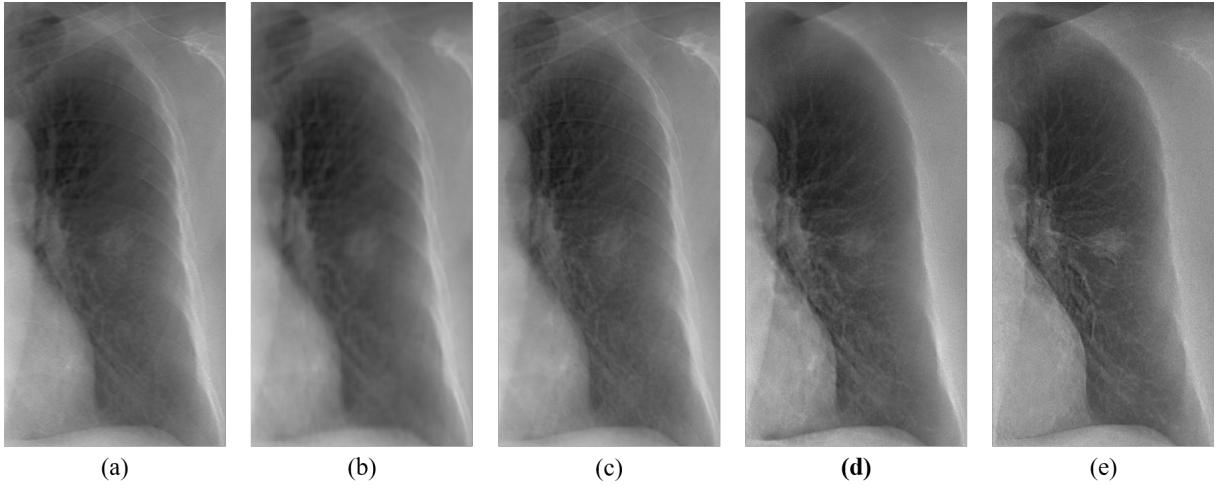


Figure 12. The side-by-side comparison of the quality of bone suppression results with difference conditions based on the ablation studies, which is described in Table 1: (a) CNN, (b) CNN + Haar, (c) CNN + GAN, (d) CNN + GAN + Haar (ours), and (e) DES.

firstly aimed to provide wavelet decomposed frequency details to help train unsupervised conditional GAN and to accelerate model convergence. However, this may act as the burden to the networks because the difference between the prediction and its ground truth becomes four times greater than the original system. When the overall data size is fixed, sharing weights for convolution for a single image is considered to be less complex compared to taking four sharing weights on each of the four images. Our proposed method specifically leverages the wavelet decomposition system and shows better results on bone suppression.

5 Conclusion

Bone suppression has received more attention to reduce the mis-diagnosis of radiologists due to the hidden lesion behind the bony structures. However, there are major drawbacks to currently commercialized method, dual energy subtraction (DES) within acquiring bone suppressed images. As many studies had contributed to this purpose, we successfully predicted the bone suppression results on single energy chest X-rays by analyzing previous acquired dual energy chest X-rays. We also built a model that outperforms existing approaches with a very intuitive approach; using adversarial training with frequency information as a guideline, and this method is not limited to bone suppression, but potentially contributes to other related scopes as well. Once suppressing bones on chest X-rays, the model understands the attenuation coefficient and spatial distribution of bones. In other words, it enables us to obtain that images highlighting the bony structures and bone landmarks through linear

system, improving diagnosis performance on skeletal system and the registration of two chest X-rays. In future work, additional experimentation will be required to further explore the clinical meaning of this study with subjective image quality assessment.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Science, Technology (No. 2017R1A2B4004503), Hankuk University of Foreign Studies Research Fund of 2018.

Appendix

We show the sample experiment results of the proposed method on single energy chest X-rays in Figure 13. Note that, the original image and its ground truth in Figure 13 are linearly windowed using windowing parameters (default) in DICOM tags, and the bone suppressed image is histogram matched to the original one.

References

1. Murphy, S. L., Xu, J., Kochanek, K. D., Curtin, S. C. & Arias, E. Deaths: Final data for 2015. (2017).
2. Shah, P. K. *et al.* Missed non–small cell lung cancer: radiographic findings of potentially resectable lesions evident only in retrospect. *Radiology* **226**, 235–241 (2003).
3. Loog, M., van Ginneken, B. & Schilham, A. M. Filter learning: application to suppression of bony structures from chest radiographs. *Med. image analysis* **10**, 826–840 (2006).
4. Vock, P. & Szucs-Farkas, Z. Dual energy subtraction: principles and clinical applications. *Eur. journal radiology* **72**, 231–237 (2009).
5. Suzuki, K., Abe, H., MacMahon, H. & Doi, K. Image-processing technique for suppressing ribs in chest radiographs by means of massive training artificial neural network (mtann). *IEEE Transactions on medical imaging* **25**, 406–416 (2006).
6. Chen, S. & Suzuki, K. Bone suppression in chest radiographs by means of anatomically specific multiple massive-training anns combined with total variation minimization smoothing and consistency processing. In *Computational Intelligence in Biomedical Imaging*, 211–235 (Springer, 2014).
7. Yang, W. *et al.* Cascade of multi-scale convolutional neural networks for bone suppression of chest radiographs in gradient domain. *Med. image analysis* **35**, 421–433 (2017).
8. Gusarev, M., Kuleev, R., Khan, A., Rivera, A. R. & Khattak, A. M. Deep learning models for bone suppression in chest radiographs. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2017 IEEE Conference on*, 1–7 (IEEE, 2017).
9. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241 (Springer, 2015).
10. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *stat* **1050**, 10 (2014).
11. Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R. & Xing, E. P. Toward controlled generation of text. In *International Conference on Machine Learning*, 1587–1596 (2017).
12. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1125–1134 (2017).
13. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2223–2232 (2017).
14. Goodfellow, I. *et al.* Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680 (2014).
15. Reed, I. S., Glenn, W. V., Truong, T., Kwok, Y. S. & Chang, C. M. X-ray reconstruction of the spinal cord, using bone suppression. *IEEE Transactions on Biomed. Eng.* **293**–298 (1980).
16. Juhász, S., Horváth, Á., Nikházy, L. & Horváth, G. Segmentation of anatomical structures on chest radiographs. In *XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010*, 359–362 (Springer, 2010).
17. Oğul, H., Oğul, B. B., Ağıldıre, A. M., Bayrak, T. & Sümer, E. Eliminating rib shadows in chest radiographic images providing diagnostic assistance. *Comput. methods programs biomedicine* **127**, 174–184 (2016).
18. Horváth, Á., Orbán, G. G., Horváth, Á. & Horváth, G. An x-ray cad system with ribcage suppression for improved detection of lung lesions. *Period. Polytech. Electr. Comput. Sci.* **57**, 19 (2013).

19. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**, 600–612 (2004).
20. Wang, X. & Gupta, A. Generative image modeling using style and structure adversarial networks. In *European Conference on Computer Vision*, 318–335 (Springer, 2016).
21. Mathieu, M., Couprie, C. & LeCun, Y. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440* (2015).
22. Stollnitz, E. J., DeRose, A. D. & Salesin, D. H. Wavelets for computer graphics: a primer. 1. *IEEE Comput. Graph. Appl.* **15**, 76–84 (1995).
23. Xizhi, Z. The application of wavelet transform in digital image processing. In *2008 International Conference on MultiMedia and Information Technology*, 326–329 (IEEE, 2008).
24. Cohen, R. Signal denoising using wavelets. *Proj. Report, Dep. Electr. Eng. Tech. Isr. Inst. Technol. Haifa* (2012).
25. Talukder, K. H. & Harada, K. Haar wavelet based approach for image compression and quality assessment of compressed image. *arXiv preprint arXiv:1010.4084* (2010).
26. Kang, E., Min, J. & Ye, J. C. A deep convolutional neural network using directional wavelets for low-dose x-ray ct reconstruction. *Med. physics* **44** (2017).
27. Kang, E., Chang, W., Yoo, J. & Ye, J. C. Deep convolutional framelet denosing for low-dose ct via wavelet residual network. *IEEE transactions on medical imaging* **37**, 1358–1369 (2018).
28. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
29. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507* **7** (2017).
30. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
31. Shrivastava, A. *et al.* Learning from simulated and unsupervised images through adversarial training. In *CVPR*, vol. 2, 5 (2017).
32. Salimans, T. *et al.* Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, 2234–2242 (2016).

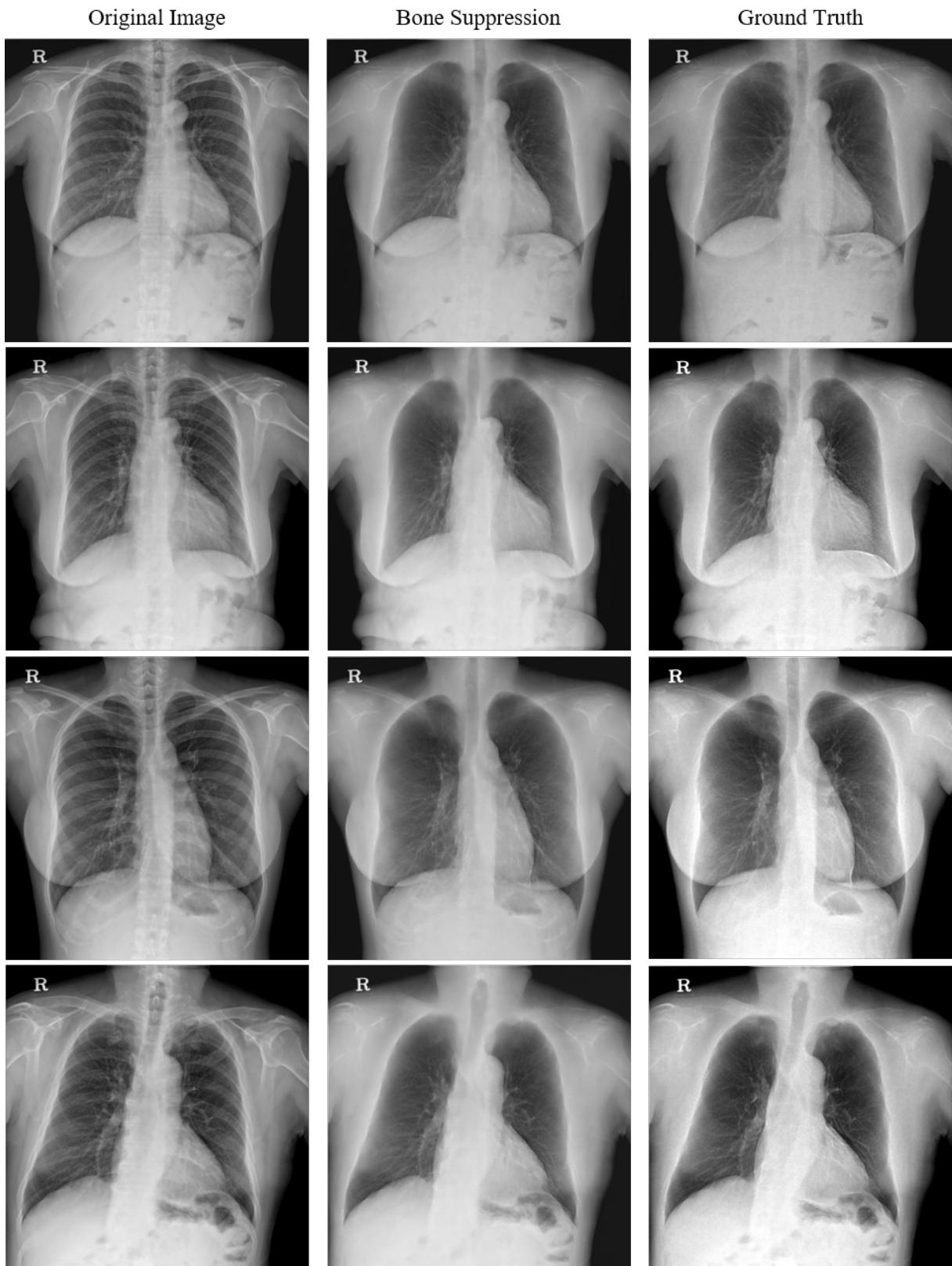


Figure 13. The figure shows the examples of original image (right column), bone suppressed with the proposed method (center column) and ground truth obtained via DES (left column).