Medical Image Analysis 35 (2017) 421–433

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Cascade of multi-scale convolutional neural networks for bone suppression of chest radiographs in gradient domain

Wei Yang[a], Yingyin Chen[a], Yunbi Liu[a], Liming Zhong[a], Genggeng Qin[b], Zhentai Lu[a], Qianjin Feng[a,*], Wufan Chen[a]

[a] Guangdong Provincial Key Laboratory of Medical Image Processing, School of Biomedical Engineering, Southern Medical University, Guangzhou, China
[b] Radiology Department, Nanfang Hospital, Southern Medical University, Guangzhou, China

ARTICLE INFO

Article history:
Received 3 March 2016
Revised 24 July 2016
Accepted 15 August 2016
Available online 16 August 2016

Keywords:
Chest radiography
Convolutional neural network
Bone suppression
Dual-energy subtraction

ABSTRACT

Suppression of bony structures in chest radiographs (CXRs) is potentially useful for radiologists and computer-aided diagnostic schemes. In this paper, we present an effective deep learning method for bone suppression in single conventional CXR using deep convolutional neural networks (ConvNets) as basic prediction units. The deep ConvNets were adapted to learn the mapping between the gradients of the CXRs and the corresponding bone images. We propose a cascade architecture of ConvNets (called CamsNet) to refine progressively the predicted bone gradients in which the ConvNets work at successively increased resolutions. The predicted bone gradients at different scales from the CamsNet are fused in a maximum-a-posteriori framework to produce the final estimation of a bone image. This estimation of a bone image is subtracted from the original CXR to produce a soft-tissue image in which the bone components are eliminated. Our method was evaluated on a dataset that consisted of 504 cases of real two-exposure dual-energy subtraction chest radiographs (404 cases for training and 100 cases for test). The results demonstrate that our method can produce high-quality and high-resolution bone and soft-tissue images. The average relative mean absolute error of the produced bone images and peak signal-to-noise ratio of the produced soft-tissue images were 3.83% and 38.7 dB, respectively. The average bone suppression ratio of our method was 83.8% for the CXRs with pixel sizes of nearly 0.194 mm. Furthermore, we apply the trained CamsNet model on the CXRs acquired by various types of X-ray machines, including scanned films, and our method can also produce visually appealing bone and soft-tissue images.

1. Introduction

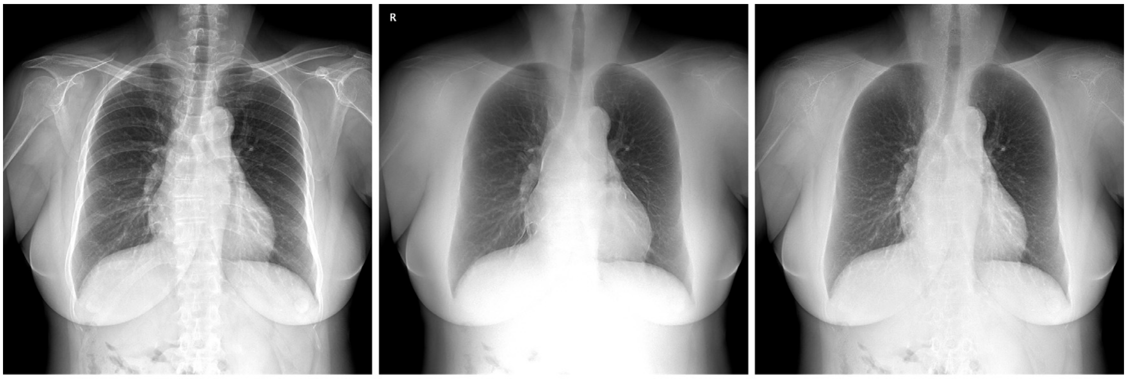
Chest radiography (chest X-ray: CXR) is a widely used diagnostic imaging technique for lung diseases because this method is inexpensive, routinely available, and relatively safe. However, overlying anatomical structures, such as ribs and clavicles, pose difficulty for radiologists to read and interpret CXRs. Suppression of bones in CXRs would be potentially useful for radiologists (Li et al., 2011) as well as computer-aided nodule detection performance (Chen and Suzuki, 2013).

One way to reduce the visual clutter of overlying anatomy in CXRs is dual-energy subtraction (DES) imaging (Vock and Szucs-Farkas, 2009). DES radiography involves capturing two radiographs with the use of two X-ray exposures at two different energy levels. These two radiographs are then combined to form a subtraction image that highlights either soft-tissue or bone components. The soft-tissue image can achieve improved visualization of soft-tissue because the ribs and clavicles become invisible, as shown in Fig. 1(b). DES radiography exhibits many advantages over conventional CXR in terms of facilitating image interpretation. However, only a few hospitals use a DES system because of the required specialized equipment. In addition, the motion artifacts caused by cardiac motion or breath during the interval between the two X-ray exposures seriously influences the imaging quality of DES soft-tissue and bone images.

Another way to suppress bones in CXRs is using an image processing technique that does not require specialized equipment for DES. The commercial software ClearRead Bone Suppress (formerly SoftView) of Riverain Technologies is a tool for bone suppression in CXRs that has been approved by FDA and CFDA (Riverain Technologies, 2016). Previous methods for bone suppression can generally be divided into two categories: supervised and unsupervised methods. The supervised methods treat bone suppression in CXRs as a regression prediction problem, and the regressors are trained or optimized by teaching DES radiographs to estimate the soft-tissue or bone images (Chen and Suzuki, 2014; Chen et al., 2016; Loog et al., 2006; Suzuki et al., 2004; Suzuki et al., 2006). In the

[*] Corresponding author.
E-mail address: qianjinfeng08@gmail.com (Q. Feng).

http://dx.doi.org/10.1016/j.media.2016.08.004
1361-8415/© 2016 Elsevier B.V. All rights reserved.



(a) chest radiograph (b) DES soft-tissue image (c) soft-tissue image produced by our method

Fig. 1. Illustration of (a) a standard chest radiograph, (b) corresponding real DES soft-tissue image, and (c) corresponding soft-tissue image produced by our method.

supervised methods, only local image features and information of the input CXRs can be used to predict the soft-tissue or bone images. Extracting the useful information and recognizing the characteristic structures from the CXR to estimate the corresponding soft-tissue or bone components are the key problems for the regressors. Given that the contents of bone images are more consistent and simple than components of soft-tissue images, the prediction target of the supervised models are usually bone images (Chen and Suzuki, 2014; Suzuki et al., 2004), which was the same approach used in our study. The unsupervised methods for bone suppression do not require teaching DES with radiographs, but these methods need segmentation and the border locations of bony structures as intermediate results (Hogeweg et al., 2013; Lee et al., 2012; Rasheed et al., 2007; Simko et al., 2009). The bone-free images are reconstructed with the blind-source separation approach or the gradient images modified according to the intermediate results. The effectiveness of unsupervised methods highly depends on the accuracy of segmentation and the border locations of bony structures.

In this work, we aim to develop an effective bone suppression method for conventional CXRs in a supervised way using a large number of real two-exposure DES CXRs as training data. We sought to build a regression prediction model with the training data. This model can predict the bone components utilizing only image features from a single CXR and should have the capability to automatically extract useful image features and context information from the input CXR. Meanwhile, deep learning (LeCun et al., 2015) and deep convolutional neural networks (ConvNet) have recently been successfully applied to dense prediction problems such as rain drop removal (Eigen et al., 2013), super-resolution (Dong et al., 2016), and edge-aware filtering (Xu et al., 2015). Deep ConvNets can directly learn end-to-end mapping, and the levels of features can be enriched by the number of stacked layers (depth) (Zeiler and Fergus, 2014). The task of predicting bone or soft-tissue components from a CXR is analogous to image denoising, rain drop removal, and edge-aware filtering. Motivated by these similarities, we considered using ConvNets as the basic prediction models to learn the mapping between the CXRs and the corresponding bone images.

However, obtaining a good prediction of bone images is still difficult using ConvNets. Contextual and structural information of a large spatial range (receptive field) in the CXRs should be extracted by the ConvNets as much as possible to determine whether bony structures are present and to determine the proportion of bone components. If the ConvNet for the fine-scale prediction is in a fully convolutional form, then the sizes of the convolution kernels or the number of convolution layers should be large enough to capture the information in a large receptive field. The ConvNet would become very large with an excessive number of parameters to learn and more training samples would be required. The training tasks for large deep neural networks can be very difficult (Glorot and Bengio, 2009).

To avoid training very large ConvNets, we present a cascade architecture of ConvNets to predict the bone components at a fine resolution. We call this model **CamsNet** (**Ca**scade of **m**ulti-**s**cale Conv**Net**s). The ConvNets with fixed receptive fields in a CamsNet function at successively increased resolutions. Using the upscaled output of a ConvNet for a coarser scale as a part of input feature maps, the context and structure information of a large spatial range with respect to the finer resolution can propagate to the successive ConvNet for a finer scale. Although the receptive field of each ConvNet is relatively small, the receptive field of cascaded ConvNets can be very large. The result depends on the number and the receptive field of ConvNets in the cascade. Thus, the context and structure information of a large spatial support is involved for the final prediction at the finest scale. The proposed cascade architecture can lead to significantly improved prediction accuracy of bone images related to a single-scale ConvNet. By fusing the output of the ConvNets at multiple scales, the quality of predicted bone images can be further improved.

Instead of learning the mapping between the intensities of CXRs and their bone components, we train the ConvNets for different scales to learn the mappings between CXRs and their bone components in the gradient domain. In fact, the gradients contain enough information for reconstructing the images. Moreover, the gradients are sparse, and the probability of both soft-tissue and bone components having large gradients of the same orientation at the same location is low (Chen et al., 2009). Therefore, the soft-tissue and bone components of CXRs in the gradient domain are more separable than in the intensity domain. The mapping between the gradients of CXRs and their bone components may be less complicated than the mapping between the intensities. This mapping can be learned relatively easier by the ConvNets. Furthermore, the gradients of CXRs can be conveniently normalized because they obey a nearly identical distribution.

The remainder of this paper is organized as follows. First, we briefly introduce the related works in Section 2. The framework and details of our method are described in Section 3. The experimental results are provided in Section 4. The results are discussed in Section 5. Finally, a summary of the results is presented in Section 6.

2. Related works

One of the early works addressing bone suppression in CXRs was MTANN proposed in (Suzuki et al., 2004; Suzuki et al., 2006). These studies employed traditional artificial neural networks with one hidden layer to predict the bone image from a CXR, which could be subtracted to yield an image similar to a soft-tissue image. MTANN was trained under a multi-resolution framework on the CXRs with known bone components obtained by a computed radiography (CR) system with a single-exposure DES unit. Each MTANN was trained independently for a certain resolution with corresponding resolution images using multi-resolution decomposition, and the prediction targets of MTANN were the values of single pixels extracted from the bone images. The trained MTANNs produced different-resolution bone images, and then these images were combined to provide a complete high-resolution bone image by using multi-resolution composition. Five downsampled (440×440 pixels) DES chest radiographs were used as the training set. The MTANN method was improved to separate bones from soft-tissues by training MTANN in different anatomic segments and combining total variation (TV) minimization for the bone images (Chen and Suzuki, 2014). Nine DES chest radiographs were used as the training data in that study. The artificial neural networks (e.g., multi-layer perceptron) were used to predict the pixel values or the wavelet coefficients of bone images in the software Riverain's ClearRead Bone Suppress, and the inputs of the multi-layer perceptrons were well-designed image features such as multi-scale harmonics derivatives and shape indices (Knapp and Worrell, 2012). Recently, Chen et al. extended their method for bone suppression in portable CXRs (Chen et al., 2016). An earlier initial version (Suzuki et al., 2006) was nonexclusively licensed and commercialized by Riverain Technologies. Another type of supervised method for bone suppression of CXRs is the k-nearest neighbor (kNN) regression method with optimized local features (Loog et al., 2006). Using eight downsampled (512×512 pixels) CXRs with known soft-tissue and bone components obtained by a single-exposure CR DES system as the training data, Loog et al. proposed a linear dimensionality reduction method for local image features to optimize the performance of kNN regression (Loog et al., 2006). Despite promising results in (Loog et al., 2006), the ribs could not be completely filtered out from the CXRs by kNN regression. Additionally, kNN regression has the disadvantage of having a relatively long running time for bone suppression. In contrast, we use a large number of DES chest radiographs to train the prediction model that can produce high-resolution bone and soft-tissue images.

Aside from the supervised learning methods, several unsupervised methods for suppressing bony structures without the need of teaching DES radiographs have also been studied. Simko et al. suppressed clavicles by creating a bone image from a gradient map modified along the bone border direction, and then created a clavicle-free image by subtracting the bone image (Simko et al., 2009). This method could also be applied to suppress ribs. Their method needed to detect the bone borders first. Hogeweg et al. proposed blind-source separation techniques to suppress bony structures in CXRs (Hogeweg et al., 2013). The segmentation results of ribs and clavicles are used in the suppression procedure. These unsupervised methods need accurate segmentation and border locations for the targeted structures. However, this task is also challenging (Ginneken et al., 2006).

Recently, deep learning and deep ConvNets have been successfully applied in many applications such as image classification, image labeling, and image denoising (LeCun et al., 2015). The progresses of hardware and algorithms for deep ConvNets cause the possibility of training ConvNets with many convolution layers on the large-scale samples. The works closely related to our method are those on the deep ConvNets of a fully convolutional form for removing noisy patterns (dirt/rain) (Eigen et al., 2013), single image super-resolution (Dong et al., 2016), and edge-aware filtering (Xu et al., 2015). Eigen et al. used a ConvNet to predict clean patches, given dirty ones as input (Eigen et al., 2013). Deep ConvNets were also applied to learn the mapping between the low- and high-resolution images for image super-resolution (Dong et al., 2016). Dong et al. showed that the sparse coding-based super-resolution methods could be viewed as a deep ConvNet. Xu et al. used the deep ConvNets with a gradient domain training procedure to approximate many types of edge-aware filtering effects (Xu et al., 2015). Our method for bone suppression benefits from these excellent works on the application of ConvNet. The deep ConvNets are adopted as the basic units for predicting the bone components of CXRs in the gradient domain.

3. Methods

We predicted the bone components using a CamsNet for a given CXR and then subtracted the bone image from the original CXR to obtain a bone-suppressed image (soft-tissue image). We first introduced the ConvNets to predict the gradients of the bone image at a single scale. The method to combine multiple ConvNets into a CamsNet and the solutions for some related issues are described in the following sections.

3.1. ConvNet for predicting bone images in the gradient domain

We first resized an input CXR to a fixed resolution by using bicubic interpolation and then normalized the CXR such that the 90th percentile of the normalized gradient magnitudes became constant (e.g., 10). Let us denote the normalized CXR as $\mathbf{I}$ and its horizontal and vertical gradients as $\mathbf{I}_x$ and $\mathbf{I}_y$, respectively. Our goal was to predict the gradient images $\mathbf{G}_x$ and $\mathbf{G}_y$, which were as close as possible to the gradients $\mathbf{B}_x$ and $\mathbf{B}_y$ of the ground truth bone image. We needed to learn a mapping F between ($\mathbf{I}_x$, $\mathbf{I}_y$) and ($\mathbf{B}_x$, $\mathbf{B}_y$), or between ($\mathbf{I}_x$, $\mathbf{I}_y$, $\mathbf{B}_x$', $\mathbf{B}_y$') and ($\mathbf{B}_x$, $\mathbf{B}_y$) if a coarse estimation of the bone image is given as the input. $\mathbf{B}_x$' and $\mathbf{B}_y$' are denoted as the estimated gradients of the coarse bone image. We accomplish this task using a ConvNet with a fully convolutional form without the pooling layers or the full-connected layers, $(\mathbf{G}_x, \mathbf{G}_y) = F(\mathbf{I}_x, \mathbf{I}_y)$, or $(\mathbf{G}_x, \mathbf{G}_y) = F(\mathbf{I}_x, \mathbf{I}_y, \mathbf{B}_x', \mathbf{B}_y')$. The ConvNet is composed of a series of convolution layers. Each of these convolution layers applies a linear convolution to its input, typically followed by an element-wise nonlinear transform, as shown in Fig. 2.

The input gradient images and the intermediate output of convolution layers are called the feature maps, which are referred to as F_l. The input feature map is $F_0 = (\mathbf{I}_x, \mathbf{I}_y)$ or $(\mathbf{I}_x, \mathbf{I}_y, \mathbf{B}_x', \mathbf{B}_y')$. The convolution layers can typically be expressed as

$$F_l = \sigma\left(\mathbf{W}_l * F_{l-1} + B_l\right), l = 1, 2, .., \mathrm{L} - 1, \tag{1}$$

where l indexes the layer; L is the number of convolution layers of the ConvNet; $\mathbf{W}_l$ and B_l represent the convolution filters and biases, respectively; $\sigma()$ is a nonlinear function that could be a rectified linear unit (ReLU) (Nair and Hinton, 2010), a hyperbolic tangent, or a Sigmoid function; and $*$ denotes the convolution operation. $\mathbf{W}_l$ corresponds to c_l filters of size $p_l \times p_l \times c_{l-1}$, where c_{l-1} is the number of channels in the input feature map, and p_l is the spatial size of the convolution kernel. $\mathbf{W}_l$ applies c_l convolutions on the input feature map, and each convolution filter has a kernel size of $p_l \times p_l \times c_{l-1}$. The output F_l is composed of c_l feature maps. B_l is a c_l-dimensional vector with elements that are associated with a filter. The same bias is used at each spatial location. We chose ReLU as the nonlinear function $\sigma()$ because ReLU can realize fast training while still presenting good quality without unsupervised pre-training (Krizhevsky et al., 2012). The ReLU function can be expressed as $\mathrm{ReLU}(t) = \begin{cases} t, & \text{if } t > 0 \\ 0, & \text{otherwise} \end{cases}$.

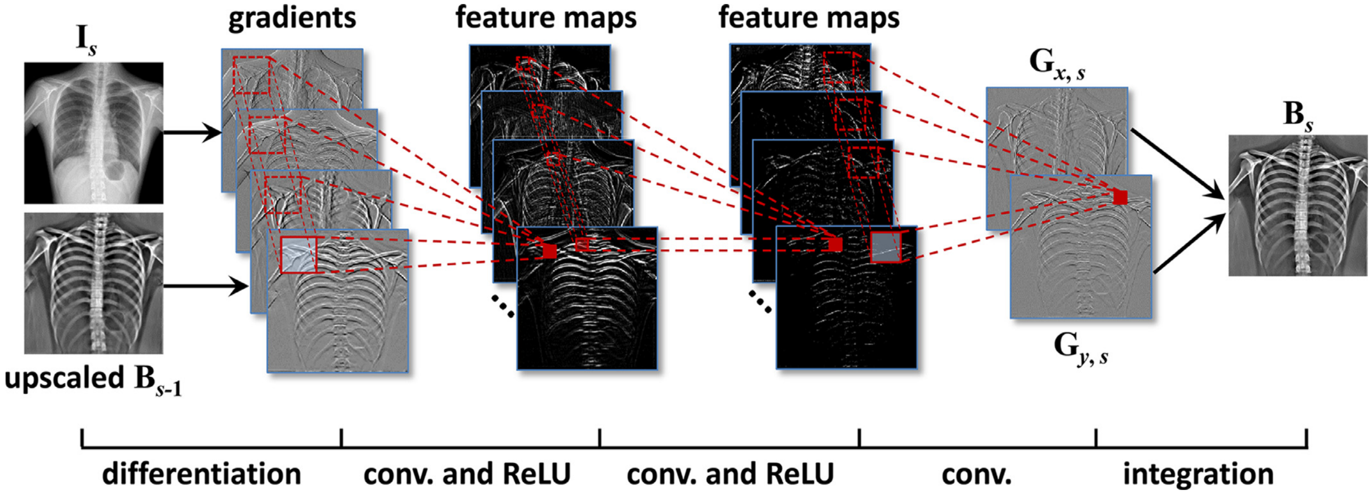


Fig. 2. Basic prediction pipeline of bone images using a ConvNet for a certain scale. The input feature maps of the ConvNet are the gradients of the downscaled input CXR $\mathbf{I}_s$ and the upscaled bone image $\mathbf{B}_{s\text{-}1}$ predicted by a unit for a coarse scale. The outputs of the ConvNet are the predicted gradients of the bone image at a finer scale, which were integrated to reconstruct the bone image $\mathbf{B}_s$.

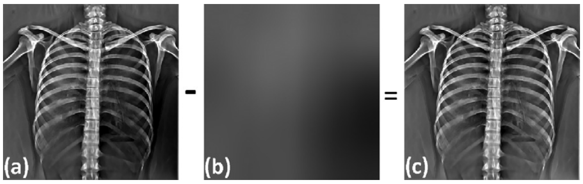


Fig. 3. (a) Integration bone image with inconsistent background intensity. (b) Background intensity estimated by a guided image filter. (c) Corrected bone image in which the background intensity is removed.

The final output of the ConvNet is produced by a convolution layer without nonlinear transform

$$F_{\mathrm{L}} = \mathbf{W}_{\mathrm{L}} * F_{\mathrm{L}-1} + B_{\mathrm{L}}. \tag{2}$$

F_{L} represents the two-channel feature maps. Each channel of F_{L} corresponds to the predicted gradient image of one direction ($\mathbf{G}_x$ and $\mathbf{G}_y$) of the bone image. The number of convolution layers of the ConvNets in this work is fixed to 3 (i.e., L=3) as in (Dong et al., 2016; Eigen et al., 2013). Each convolution layer in the three-layer ConvNets for predicting bone gradients can perform a specific function. The first layer performs feature extraction. The second layer performs nonlinear mapping. The last layer performs reconstruction of the bone gradients.

Reconstruction of a bone image from the predicted gradients: After obtaining the predicted gradient maps $\mathbf{G}$=($\mathbf{G}_x$, $\mathbf{G}_y$) of a bone image, we still need to reconstruct the bone image from the gradient maps. An estimation of the bone image can be reconstructed from the gradients through 2D integration by solving the Poisson equation:

$$\nabla^2 \hat{\mathbf{B}} = \nabla \bullet \mathbf{G}, \tag{3}$$

where ∇^2 is the Laplacian operator,$\nabla^2 \hat{\mathbf{B}} = \frac{\partial^2 \hat{\mathbf{B}}}{\partial x^2} + \frac{\partial^2 \hat{\mathbf{B}}}{\partial y^2}$, and $\nabla \bullet \mathbf{G}$ is the divergence of the gradients, $\nabla \bullet \mathbf{G} = \frac{\partial \mathbf{G}_x}{\partial x} + \frac{\partial \mathbf{G}_y}{\partial y}$. To solve the Poisson equation Eq. (3), Neumann boundary conditions are used. However, 2D integration of the predicted gradients may cause inconsistent background intensity as shown in Fig. 3(a) because the predicted gradients are possibly not integrable. We treated the background intensity as a very smooth component of the integration image $\hat{\mathbf{B}}$, and then the smooth component was removed from $\hat{\mathbf{B}}$ to reduce the intensity inconsistency and maintain global contrast in the corresponding bone-suppressed result. We adopted a guided image filter (He et al., 2013) with large kernels to smooth $\hat{\mathbf{B}}$ as an estimation of the background intensity. An example of an integration bone image, the estimated background intensity, and the corrected integration image are shown in Fig. 3.

3.2. CamsNet: cascade of multi-scale ConvNets for predicting bone images

As mentioned previously, accurately predicting a fine bone image using one ConvNet at a fine scale is difficult because only limited local image information can be utilized. In fact, predicting a coarser bone image is relatively more easy and accurate than predicting a finer one. The reason is that large image structures are useful to reduce the uncertainty of prediction. As demonstrated in (Eigen and Fergus, 2015; Schmidt et al., 2016), cascade is an effective architecture to refine prediction from basic units, which was employed to integrate the ConvNets for multiple scales in our work. As shown in Fig. 4, we can start with the coarsest prediction of bone image from the downscaled input CXR so that sufficient and useful information from a large spatial support can be contained in the prediction. Then we used the intermediate coarse prediction as a part of the input for the fine-scale ConvNet. Therefore, the information from the coarse prediction propagates successively to the fine-scale ConvNet in the pipeline of CamsNet. The fine-scale ConvNets progressively refine the bone prediction by integrating the coarser prediction and the information from finer features in the CXR. The coarse or inaccurate intermediate predictions can provide useful context information to help predict finer bone images. The work pipeline of CamsNet is somewhat similar to the auto-context model (Tu and Bai, 2010) that works on a single scale in a cascade framework.

From another perspective, CamsNet can be viewed as a single large, deep ConvNet. The successive ConvNets from coarser to finer scales are connected by additional three layers. These layers perform 2D integration, upscale, and differentiation operations. The upscale and differentiation operations are performed on the integration image. Thus, the upscaled bone gradients in the input feature maps are integrable and are the proper estimation of the bone gradients at a finer scale. In the actual implementation, the estimation of bone images at different scales is reconstructed by solving Eq. (3). The upscale operation is implemented by bicubic interpolation, and the upscale factor of each intermediate prediction is set to 2. For a ConvNet at a certain scale, the input feature maps have the same resolution by upscaling the output of a previous prediction unit and downscaling the original input CXR. If

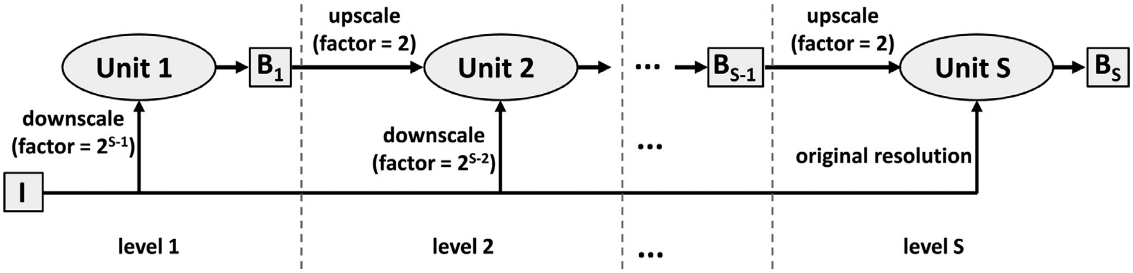
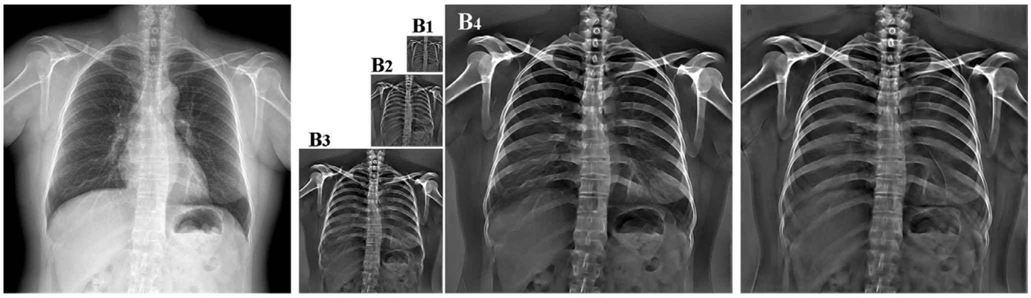


Fig. 4. Prediction cascade for bone images. **I** is an input CXR. $\mathbf{B}_s$ is the predicted bone image at scale $1/2^{S-s}$ ($s=1, 2,\ldots, S$). S is the level number of the cascade.



(a) an input chest radiograph (b) predicted bone images at different scales (c) multi-scale fusion

Fig. 5. Example of predicted bone images. (a) Chest radiograph. (b) Bone images at different scales predicted by a 4-level CamsNet. (c) Bone image produced by fusing the predicted multi-scale bone gradients from the CamsNet.

the effect of 2D integration is ignored, then the receptive field of a CamsNet can be computed as $1+\sum_{s=1}^{S} 2^{s-1}(p_s-1)$, where S is the number of ConvNets for different scales and p_s is the receptive field of each ConvNet in the prediction unit s. The CamsNet can simultaneously use the context information of a large receptive field and the fine structure information for prediction. Each ConvNet is trained individually, and the need for training a single very large and deep ConvNet is avoided by cascading the ConvNets for multiple scales. Therefore, the CamsNet can achieve more efficient and easier learning of the complicated mapping between the CXRs and their bone components.

In this study, the level number of a CamsNet is set to 4 (i.e., the number of ConvNets in the CamsNet $S=4$). We found that four levels was enough to achieve satisfactory performance. If the level number was significantly large, then the training samples for the ConvNets for coarse scales would be insufficient. Fig. 5(b) shows four bone images at different scales predicted by the units at different levels in a 4-level CamsNet.

3.3. Fusion of the multi-scale predictions from CamsNet

Each ConvNet in a CamsNet can yield the predicted bone gradients at different scales. The information predicted by the ConvNets for different scales may be complementary. Proper combination of the predicted bone gradients of multiple scales may lead to a better estimation of bone images than the final outputs of the CamsNet. We integrated the predicted bone gradients at different scales in the maximum-a-posteriori (MAP) framework. The corresponding energy function was defined as:

$$E(\mathbf{B}) = \lambda\|\mathbf{B}\|^2 + \sum_{s=1}^{S}\left(\frac{1}{4^{S-s}\sigma_{x,s}^2}\|\mathbf{D}_x\mathbf{A}_s\mathbf{B}-\mathbf{G}_{x,s}\|^2 + \frac{1}{4^{S-s}\sigma_{y,s}^2}\|\mathbf{D}_y\mathbf{A}_s\mathbf{B}-\mathbf{G}_{y,s}\|^2\right), \tag{4}$$

where $\mathbf{D}_x$ and $\mathbf{D}_y$ denote the forward difference operators; $\mathbf{G}_{x,\ s}$ and $\mathbf{G}_{y,\ s}$ are the gradients predicted by the ConvNet in the unit of level s; $\sigma_{x,\ s}$ and $\sigma_{y,\ s}$ are the roots of the mean-squared error of bone gradients predicted by the ConvNet in the unit of level s; $\mathbf{A}_s$ denotes the downscaling operation with the factor of 2^{S-s} via bicubic interpolation; and λ is a tunable parameter. The energy function $E(\mathbf{B})$ is quadratic with respect to the bone image **B**. The optimal solution of **B** in Eq. (4) can be obtained by solving the following linear equation through minimum residual method (e.g., function “minres” built in Matlab):

$$\left(\sum_{s=1}^{S}\frac{1}{4^{S-s}}\mathbf{A}_s^T\left(\frac{1}{\sigma_{x,s}^2}\mathbf{D}_x^T\mathbf{D}_x+\frac{1}{\sigma_{y,s}^2}\mathbf{D}_y^T\mathbf{D}_y\right)\mathbf{A}_s+\lambda\mathbf{E}\right)\mathbf{B} = \sum_{s=1}^{S}\frac{1}{4^{S-s}}\mathbf{A}_s^T\left(\frac{1}{\sigma_{x,s}^2}\mathbf{D}_x^T\mathbf{G}_{x,s}+\frac{1}{\sigma_{y,s}^2}\mathbf{D}_y^T\mathbf{G}_{y,s}\right), \tag{5}$$

where **E** denotes an identity matrix.

3.4. Generation of ground truth from real DES data

Each case of the DES CXR in our collected dataset included a standard CXR (denoted by **I**), a DES soft-tissue image (denoted by $\mathbf{S}^0$), and a DES bone image (denoted by $\mathbf{B}^0$). Given the sophisticated nonlinear post-processing (e.g., contrast enhancement and sharpening) of the raw image data, the relationship $\mathbf{I}=\mathbf{S}^0+\mathbf{B}^0$ was not eventually satisfied, and $\mathbf{S}^0/\mathbf{B}^0$ was not exactly the soft-tissue/bone component in the CXR **I**. If DES bone images are used as the prediction target, the soft-tissue image cannot be produced by directly subtracting the predicted bone image from the corresponding chest radiograph. Therefore, we needed to separate the bone component or soft-tissue component from the CXR **I** using the information contained in $\mathbf{S}^0$ and $\mathbf{B}^0$ as the ground truth for training and quantitative evaluation of the models.

We treated the bone components as shadows and removed their edges from the CXR by transforming its gradients using cross projection tensors (Agrawal et al., 2006) obtained from $\mathbf{B}^0$ that contains the structure information of the bone component. First, the noise in the images was reduced by the BM3D algorithm (Dabov et al., 2007) to obtain more smooth and reliable gradients. Then the intensity ranges of $\mathbf{S}^0$ were calibrated by multiplying a coefficient α such that the correlation between $\mathbf{I}-\alpha\mathbf{S}^0$ and $\mathbf{S}^0$ was minimized. An extremely smoothed $\mathbf{S}^0$ from the guided image

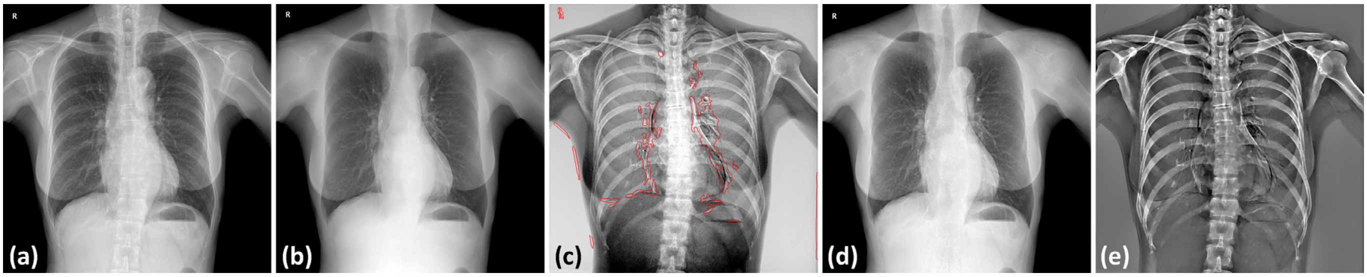


Fig. 6. Example of DES chest radiography and the corresponding decomposition results through cross projection tensors. (a) Standard chest radiograph. (b) DES soft-tissue image. (c) DES bone image. (d) Soft-tissue image reconstructed from transformed gradients. (e) Bone image reconstructed from transformed gradients. The detected regions with motion artifacts are outlined by the red lines in (c). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

filter was subtracted from **I** to generate an intermediate result $\mathbf{I}^1$. The base layer of the soft-tissue components was removed from **I**, and only the details of the soft-tissue components and the bone components are kept in $\mathbf{I}^1$. Finally, the gradients **G** of the bone components in **I** were obtained as the transformed gradient field of $\mathbf{I}^1$ using cross projection tensors from $\mathbf{B}^0$. The bone component **B** in **I** was ultimately reconstructed from **G** through 2D integration. The corresponding soft-tissue component can be obtained as **I** – **B**, as shown in Fig. 6(d).

Detection of motion artifacts: Usually, a few motion artifacts are present in the soft-tissue and bone images of two-exposure DES as a result of cardiac motion and breathe. The soft-tissue and bone components were not successfully separated in the regions with motion artifacts. To reduce the effect of obvious motion artifacts on the prediction models, we masked out the regions with motion artifacts and excluded them from the training samples. We found that the correlation coefficients between the gradients of soft-tissue and bone images in the motion artifact regions were significantly higher than the coefficients in the other regions. Therefore, these correlation coefficients were computed and thresholded to generate the masks of motion artifacts. An example of the detected motion artifacts is shown in Fig. 6(c).

3.5. Training ConvNets

The parameters of a single ConvNet are represented as $\Theta=\{\mathbf{W}_l, B_l;\ l=1, 2,\ldots, \mathrm{L}\}$, which determined the behavior and performance of the ConvNet. The parameters Θ of ConvNets can be optimized by minimizing the loss between the predicted gradient images $(\mathbf{G}_x, \mathbf{G}_y)=F(\mathbf{I}_x, \mathbf{I}_y;\ \Theta)$ and the corresponding ground truth gradient images of bone $(\mathbf{B}_x, \mathbf{B}_y)$. Given a set of pairs $\{(\mathbf{G}_{x,\ i}, \mathbf{G}_{y,\ i}), (\mathbf{B}_{\ x,\ i}, \mathbf{B}_{y,\ i});\ (i=1, 2, \ldots, n)\}$, the mean-squared error (MSE), a commonly used loss function for regression tasks, is defined as

$$\ell(\Theta) = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{G}_{x,i} - \mathbf{B}_{x,i}\|^2 + \left\|\mathbf{G}_{y,i} - \mathbf{B}_{y,i}\right\|^2, \tag{6}$$

where n is the number of training samples. Other loss functions can also be used to train the ConvNets.

The ConvNets for the coarsest scale to the finest scale in CamsNet were trained successively. The intermediate predictions of the coarse bone images at medium scales were reconstructed through 2D integration, and the background intensity was removed by the method described in Section 3.1. After completing the training of a ConvNet at a coarse scale, the coarse bone images were reconstructed from its outputs and upscaled to generate samples for training the successive ConvNet at a finer scale.

Considering that the bony structures in the right and left chest are roughly symmetrical, we trained the ConvNets to predict the bone gradients of the right chest and the horizontally flipped left chest. This trick was helpful to reduce the complexity of the learning task because more samples could be collected for similar bony structures. The spinal column was approximately located at the maximum peak of the smoothed vertical intensity projection of a CXR, as shown in Fig. 7. The CXRs were divided into the overlapping right and left parts according to the detected spinal column. Then, the right and horizontally flipped left parts of CXRs were used to produce the training sample and were fed to the ConvNets. The predicted left bone gradients were horizontally flipped back and corrected. Finally, the predicted gradients of the overlapped right and left chest were merged into the entire predicted bone gradients using weighted averaging of the overlapped regions.

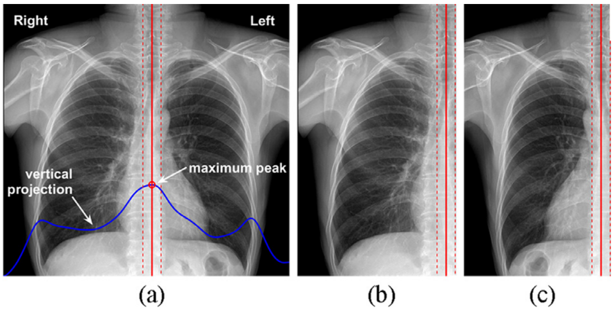


Fig. 7. Location of the spinal column. The smoothed vertical intensity projection of the chest radiograph is represented by the blue line in (a). The location of the maximum peak in the vertical intensity projection roughly indicates the location of the spinal column. The chest radiograph is divided into overlapping right (b) and horizontally flipped left parts (c) by the two dashed lines, which are located at the edges of spinal column. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

An effective way to improve the prediction performance of ConvNets is to augment the training samples. We augmented the dataset by scale (with factors in range of [0.95, 1.05]) and rotation (in a range of [−5, 5] degrees) jittering on the training images at coarse scales, and made the ConvNets not sensitive to the resolution and rotation of CXRs. During the back-propagation iterations of ConvNet training, the pairs of gradient patches were rescaled with a factor that is randomly uniform drawn in the range of [0.96, 1.04]. The mappings learned by the ConvNets are expected to be resistant to small perturbations in the magnitudes of the input feature maps.

All the convolution layers have no padding to avoid border effects during training, and ConvNets produced small outputs. In the training phase, we randomly collected 0.2 million 54×54 gradient patches of CXRs and the upscaled coarse prediction of bone images, which were paired with the centered $(54-p_s+1)\times(54-p_s+1)$ gradient patches of the corresponding ground truth of the bone images, were used as the training dataset to train the ConvNets in processing unit s. p_s is denoted as the receptive field of the

ConvNet in the processing unit *s*. Although a fixed size for patches was used in the training, the ConvNets and the CamsNet can be applied to the images of arbitrary sizes during testing. According to the mirror reflection padding of size $(p_s-1)/2$ along the two dimensions of the input feature maps, the size of the output feature maps of each ConvNet was the same as the size of the input feature maps.

The loss function in Eq. (6) was minimized with mini-batch stochastic gradient descent (SGD) based on back-propagation with momentum. The batch size, momentum, and weight decay for the mini-batch SGD were set to 8, 0.9, and 10^{-5}, respectively. The learning rate of the last convolution layer was set to 10^{-3} during the first 5 epochs and 10^{-4} during the next 20 epochs. The learning rates of the first and middle convolution layers were four times and two times higher than the previous layer, respectively. The learning was stopped after 374 K back-propagation iterations. The filter weights of each layer were initialized by drawing randomly from a Gaussian distribution with a zero mean and standard deviation of$\sqrt{2/M}$, where *M* denotes the number of incoming nodes of one neuron (He et al., 2015). The initial biases of each convolution layer were set to 0.

We implemented our method with MatConvNet toolbox (Vedaldi and Lenc, 2015) in MATLAB 2014b. On a workstation equipped with a GPU (NVIDIA Quadro K4000), training a single ConvNet took approximately 1–2 days depending on the ConvNet architecture.

3.6. Evaluation metrics

In addition to the widely used relative mean absolute error (RMAE) and peak signal-to-noise ratio (PSNR), we adopted another two metrics to quantitatively evaluate the prediction performance of the models, specifically the structure similarity (SSIM) index (Wang et al., 2004) and bone suppression ratio (BSR) (Hogeweg et al., 2013). These metrics are defined and computed as follows.

Let us denote $\mathbf{Z}$ as a ground truth image and $\hat{\mathbf{Z}}$ as a prediction of $\mathbf{Z}$. RMAE is estimated as

$$RMAE = \sqrt{\frac{1}{N}\sum_{(x,y)\in\Omega}\left|\hat{\mathbf{Z}}(x,y)-\mathbf{Z}(x,y)\right|}/(Z_{\max}-Z_{\min}),$$

where Ω denotes the valid regions in the image $\mathbf{Z}$, N is the number of pixels in Ω, (x, y) denotes the pixel locations in $\mathbf{Z}$, and $Z_{\max}$ and $Z_{\min}$ are the maximum value and the minimum value of pixels in Ω, respectively. We set 0.1 and 99.9 percentiles in the ground truth image $\mathbf{Z}$ as the values of $Z_{\max}$ and $Z_{\min}$, respectively, to reduce the effect of outliers on the quantitative metrics. A small value of RMAE indicates a good prediction. PSNR (in dB) is defined as:

$$PSNR = 20\log_{10}\left((Z_{\max}-Z_{\min})/\sqrt{\frac{1}{N}\sum_{(x,y)\in\Omega}\left(\hat{\mathbf{Z}}(x,y)-\mathbf{Z}(x,y)\right)^2}\right).$$

PSNR is partially related to perceptual quality. BSR is defined as (Hogeweg et al., 2013):

$$BSR = 1-\sum_{(x,y)\in\Omega}\left(\hat{\mathbf{S}}(x,y)-\mathbf{S}(x,y)\right)^2/\sum_{(x,y)\in\Omega}\mathbf{B}(x,y)^2,$$

where $\hat{\mathbf{S}}$ is an estimation of a ground truth soft-tissue image $\mathbf{S}$, and $\mathbf{B}$ is the corresponding ground truth bone image. $BSR=1$ indicates perfect performance.

If the bone component is treated as a type of structural noise, then the bone suppression procedure of CXR can be considered denoising or filtering. SSIM (Wang et al., 2004), a well-known denoising performance metric, can be also used to evaluate the quality of the predicted soft-tissue images. The intensity ranges of bone or soft-tissue images were rescaled into the range of [0, 255], and the default setting parameters in the implementation of SSIM were used to compute the values of SSIM indices.

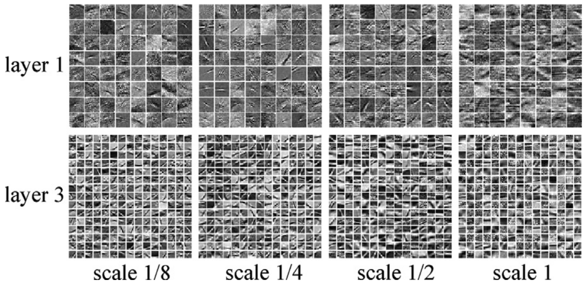


Fig. 8. Visualization of the learned filters of a CamsNet. The first row shows a subset of filters (16 × 16 pixels) for the input channel of CXR vertical gradients in the first convolution layers of the trained ConvNets for different scales. The second row shows a subset of filters (8 × 8 pixels) in the last convolution layers of the ConvNets to reconstruct vertical bone gradients at different scales.

4. Experiments

We first examined the learned filters of ConvNets. Then we investigated the effect of the number of filters and filter sizes of ConvNets on prediction performance. Subsequently, we compared the proposed CamsNet with the single-scale ConvNets in the gradient domain and in the intensity domain both quantitatively and qualitatively. We also applied the trained CamsNet model to the CXRs acquired with different X-ray machines to validate the generalization ability of the model.

4.1. Experimental data

We collected 646 posterior-anterior DES chest radiographs acquired with a digital radiography (DR) machine with a two-exposure DES unit (Discovery XR656, GE Healthcare) at Nanfang Hospital, Guangzhou, China. The X-ray tube voltages for the two exposures were 120 and 60 kV. The sizes of the CXRs ranged from 1800 × 1800 to 2022 × 2022 pixels, and the pixel sizes ranged from 0.1931 mm to 0.1943 mm. The images were restored in DICOM format with a 14-bit depth. A total of 142 radiographs with serious motion artifacts in the DES soft-tissue and bone images were excluded from the training and test dataset. The final dataset used in the experiments consisted of 504 radiographs. A total of 404 radiographs were randomly selected and used as the training set, and the remaining 100 cases were used as the test set. A detailed description of the dataset is listed in Table 1.

4.2. Learned filters and feature maps for bone prediction

Fig. 8 shows examples of learned filters in the first and third layers of the trained ConvNets for different scales. The filters in the first layers have obvious structures, and most of the structures were similar to edge detectors or Gabor filters from different directions. The majority of the filters in the third layers were similar to edge detectors of different scales from different directions. These edge detectors were used to reconstruct the gradients of the bone image, and they were considered to be the base for reconstruction.

We visualized the feature maps of a ConvNet for scale 1/4 from the right part of a CXR in Fig. 9. The bone gradients among the input feature maps were slightly blurry because of the upscale operation. The output feature maps of the first convolution layer contained different structures, which represented the details of soft-tissues and bones at different locations. The output feature maps of the second convolution layer looked very sparse. In these maps, the details related to soft-tissue structures were canceled out. The

Table 1
Training and test datasets comprised of DES chest radiographs.

Dataset	Total	Gender		Age						
		Male	Female	<=20	20~29	30~39	40~49	50~59	60~69	>=70
Training	404	305	99	18	97	71	88	86	36	8
Test	100	82	18	3	17	24	23	22	10	1

Table 2
Performance of CamsNets with different numbers of filters. The filter sizes of the three convolution layers are fixed as: $p_1 = 16$, $p_2 = 1$, and $p_3 = 8$.

Number of filters	RMAE-**B** (%)		PSNR-**S** (dB)		SSIM-**S** (%)		BSR (%)	
	CamsNet	Fusion	CamsNet	Fusion	CamsNet	Fusion	CamsNet	Fusion
64	7.31 ± 1.43	4.46 ± 0.66	33.6 ± 1.2	37.5 ± 1.0	96.1 ± 0.6	97.0 ± 0.5	46.2 ± 37.5	78.9 ± 9.3
128	6.05 ± 1.11	4.18 ± 0.63	35.2 ± 1.1	38.0 ± 1.0	96.8 ± 0.5	97.3 ± 0.5	63.0 ± 21.8	81.2 ± 8.7
256	5.33 ± 1.09	3.83 ± 0.60	36.2 ± 1.2	38.7 ± 1.1	97.2 ± 0.5	97.6 ± 0.4	69.7 ± 24.4	83.8 ± 8.0

Table 3
Performance of CamsNets with different filter sizes. The number of filters in each convolution layer is fixed as 256.

Filter size	RMAE-**B** (%)		PSNR-**S** (dB)		SSIM-**S** (%)		BSR (%)	
	CamsNet	Fusion	CamsNet	Fusion	CamsNet	Fusion	CamsNet	Fusion
8-1-4	7.05 ± 1.53	4.60 ± 0.75	34.0 ± 1.3	37.3 ± 1.0	96.6 ± 0.7	97.1 ± 0.5	50.7 ± 33.3	77.7 ± 10.3
12-1-6	5.57 ± 1.28	3.96 ± 0.60	35.9 ± 1.4	38.5 ± 1.0	97.2 ± 0.5	97.5 ± 0.4	66.7 ± 31.4	83.0 ± 7.6
16-1-8	5.33 ± 1.09	3.83 ± 0.60	36.2 ± 1.2	38.7 ± 1.1	97.2 ± 0.5	97.6 ± 0.4	69.7 ± 24.4	83.8 ± 8.0

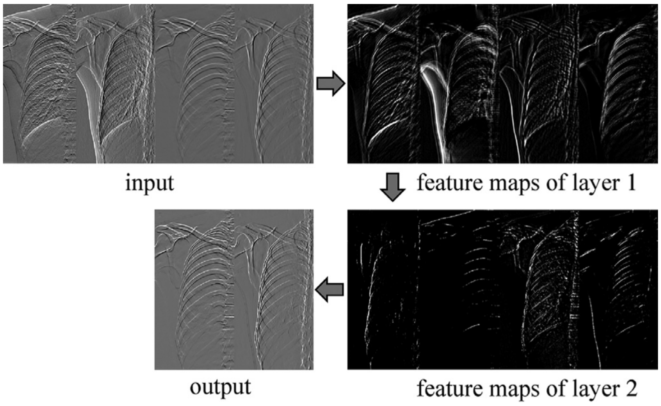


Fig. 9. Visualization of the feature maps for different convolution layers. Four of 256 output feature maps of the first and second layers of a ConvNet for scale 1/4 in a 4-level CamsNet are displayed.

final output was the predicted bone gradients, which were reconstructed with the convolution filters as the reconstruction base and the output feature maps of the second convolution layer as the coefficients. Fig. 9 shows that the output bone gradients were sharper than the input bone gradients, and the gradients in the regions without bone borders were suppressed to near zero.

4.3. Effect of the number of filters, filter sizes, and multi-scale fusion

In general, the performance of ConvNets can improve if the number of filters and the receptive field are increased at the cost of running time. We conducted experiments to investigate the effect of the numbers of filters and the filter sizes on the prediction performance of CamsNets. First, we fixed the filter sizes as $p_1 = 16$, $p_2 = 1$, and $p_3 = 8$ (denoted as 16-1-8), and the number of filters for the ConvNets were varied to train the 4-level CamsNet models. Table 2 lists the quantitative evaluation metrics of three CamsNets on the test dataset, among which the number of filters for each convolution layer in the ConvNets was set to 64, 128, and 256. We determined the RMAE of the predicted bone images (denoted as RMAE-B), the PSNR and SSIM of the produced soft-tissue images (denoted as PSNR-S and SSIM-S, respectively), and the BSR of different models. Table 2 indicates that superior performance can be achieved by increasing the number of filters. In another experiment, we fixed the number of filters in each convolution layer as 256, and the filter sizes were varied. The quantitative evaluation metrics of the corresponding CamsNets are listed in Table 3. As expected, the reasonably larger filter size can grasp rich structural information and lead to better results. In particular, the average RMAE values of bone images achieved by 8-1-4, 12-1-6, and 16-1-8 were 7.05%, 5.57%, and 5.33%, respectively. The average PSNR value of the soft-tissue images produced with a CamsNet with the number of filters set at 256 and the filter size set as 16-1-8 was 36.2 dB The results suggest that a large number of filters and a large filter size are beneficial to the CamsNets. However, the deployment speed also decreases with a large number of filters and large filter sizes.

Tables 2 and 3 also list the evaluation metrics for the results of multi-scale fusion of bone gradients at different scales from the CamsNets in the MAP framework, as described in Section 3.3. For the multi-scale fusion procedure, λ in Eq. (4) is set to 10^{-7}. We found that a smaller value of λ led to better results. The performance of the CamsNets with different numbers of filters and different filter sizes can be consistently improved with the multi-scale fusion procedure. In particular, the average BSR of CamsNets was largely improved with multi-scale fusion. For example, the average BSR of the CamsNet for 256 filters and a filter size of 16-1-8 improved from 69.7% to 83.8% via the multi-scale fusion procedure. The corresponding running time for our GPU implementation for the full bone suppression procedure was approximately 66 s for an input CXR of 2000 × 2000 pixels, among which approximately 12 s were spent on the multi-scale fusion stage. The results indicate that the MAP estimation by fusing multi-scale bone gradients can adequately utilize the output information of the ConvNets in the CamsNet to produce improved results. An example of bone images produced by different models for a CXR is shown in Fig. 10. We can see that the visual quality of bone images obtained via the multi-scale fusion procedure is significantly superior to the CamsNet output.

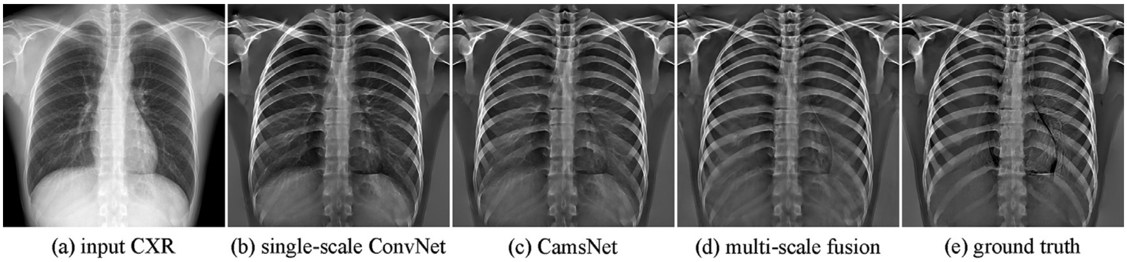


Fig. 10. The bone images produced with different models for an input chest radiograph (a). (b), (c), and (d) are the bone images produced with the single-scale ConvNet, the CamsNet, and the multi-scale fusion procedure. (e) shows the corresponding ground truth of the bone image.

Table 4
Comparison of performance of single-scale ConvNet to CamsNet.

Model	RMAE-**B** (%)	PSNR-**S** (dB)	SSIM-**S** (%)	BSR (%)
Single-scale ConvNet	7.01 ± 1.18	34.0 ± 1.0	96.0 ± 0.5	51.6 ± 23.4
CamsNet	5.33 ± 1.09	36.2 ± 1.2	97.2 ± 0.5	69.7 ± 24.4
Multi-scale Fusion	3.83 ± 0.60	38.7 ± 1.1	97.6 ± 0.4	83.8 ± 8.0

4.4. CamsNet vs. single-scale ConvNet

We trained a single-scale ConvNet with only the gradients of CXRs at the finest scale as the input feature maps to investigate the contribution of the cascade architecture to prediction performance. The architecture of the single-scale ConvNet was the same as the ConvNets in CamsNet, as shown in Fig. 2, except there was no predicted coarse bone image as the input and no bone gradients were used as the input feature maps. The number of filters for each convolution layer in the single-scale ConvNet and the CamsNet was set to 256, and the filter size was set to 16-1-8. The same procedures for normalization, 2D integration, and post-processing for the CamsNet were implemented for the single-scale ConvNet to compare the performance of bone image prediction of the two models.

Table 4 shows the quantitative results of the single-scale ConvNet, the CamsNet, and the multi-scale fusion procedure. The average RMAE of the bone images predicted by the single-scale ConvNet on the test dataset was significantly higher than the RMAE of the CamsNet (7.01% to 3.83%), as shown in Table 4. The average PSNR of the soft-tissue images produced by the single-scale ConvNet was 2 dB lower than the PSNR of the CamsNet. A good prediction of bone image and the bone suppression results were obtained with the CamsNet. The prediction quality of the CamsNet was further improved by the multi-scale fusion procedure. Qualitative results can be found in Fig. 10. Fig. 10 shows that the CamsNet produced cleaner and sharper bone images than the single-scale ConvNet, and the multi-scale fusion procedure could further remove textures in the bone image and could gave the bony structures a cleaner appearance.

4.5. CamsNets in the gradient domain vs. the intensity domain

In this section, we compared the CamsNets working in the gradient domain to the CamsNets working in the intensity domain, which were denoted as CamsNet-G and CamsNet-I, respectively. For CamsNet-I, the input feature maps of each ConvNet were the downscaled CXR and the upscaled predication of the bone image, and the training algorithm was similar to the algorithm used for CamsNet-G. The output of CamsNet-I was the directly predicted bone image at a certain scale without 2D integration. The number of filters for each convolution layer in CamsNet-I was set to 256, and the filter size was set to 16-1-8. We determined that the training for CamsNet-I was more difficult than training for CamsNet-G. The inappropriate initialization of ConvNets and the inappropriate learning rates would result in divergence, particularly for the ConvNet predictions of the fine-scale bone image in the intensity domain. Significant effort was dedicated to varying the initialization and the learning rates, and only a two-level CamsNet-I with a reasonable performance was successfully trained. Table 5 lists the quantitative results for scale 1/8 and 1/4 of the 2-level CamsNet-I and CamsNet-G. As shown in Table 5, CamsNet-G can produce better results than CamsNet-I in terms of all quantitative evaluation metrics at two scales.

Table 5
Results from 2-level CamsNets trained in the intensity domain and the gradient domain.

Scale	Domain	RMAE-**B** (%)	PSNR-**S** (dB)	SSIM-**S** (%)	BSR (%)
1/8	Gradient	3.86 ± 0.67	38.19 ± 1.73	96.86 ± 0.90	86.72 ± 6.43
1/8	Intensity	4.01 ± 0.74	37.73 ± 1.66	96.01 ± 1.08	85.18 ± 7.52
1/4	Gradient	3.66 ± 0.69	39.01 ± 1.27	97.63 ± 0.54	86.29 ± 7.33
1/4	Intensity	4.00 ± 0.64	38.03 ± 1.12	96.44 ± 0.67	83.12 ± 7.52

Fig. 11 shows the bone images at scale 1/4 produced by the 2-level CamsNets-I and CamsNets-G for a CXR. The regions without the bone borders in the bone image produced by CamsNets-G were smoother and cleaner than the regions produced by CamsNets-I, as shown in Fig. 11. CamsNets-G could also produce sharper edges than CamsNets-I. These results confirmed that the soft-tissue and bone components were more separable in the gradient domain, and the mapping between them can be learned easily in the gradient domain. This fact makes CamsNets-G more effective than CamsNets-I for predicting the bone images.

4.6. Qualitative evaluation

In this section, we applied a trained 4-level CamsNet in the gradient domain to perform bone suppression on the test dataset, and the results were evaluated qualitatively. The number of filters for each convolution layer of the ConvNets in the CamsNet was set to 256, and the filter size was set to 16-1-8. The final output bone image was produced by fusing the four-scale bone gradients in the MAP framework as described in Section 3.3. A soft-tissue image was simply obtained by subtracting the predicted bone image from the input CXR.

Figs. 12 and 13 illustrate two examples of bone suppression results using our method. Our method provided visually appealing soft-tissue images, and the bone components, including ribs, clavicles, and spines, were effectively eliminated in the produced soft-tissue images while maintaining visibility of the textures and detailed structures of the soft-tissues. Few obvious bony structures can be observed in the lung fields of the produced soft-tissue images. For the large scale, the soft-tissue images produced by our method and the DES system were visually similar. Some residual pieces of the scapula and humeri remained in both of the predicted soft-tissue images and the DES soft-tissue images,

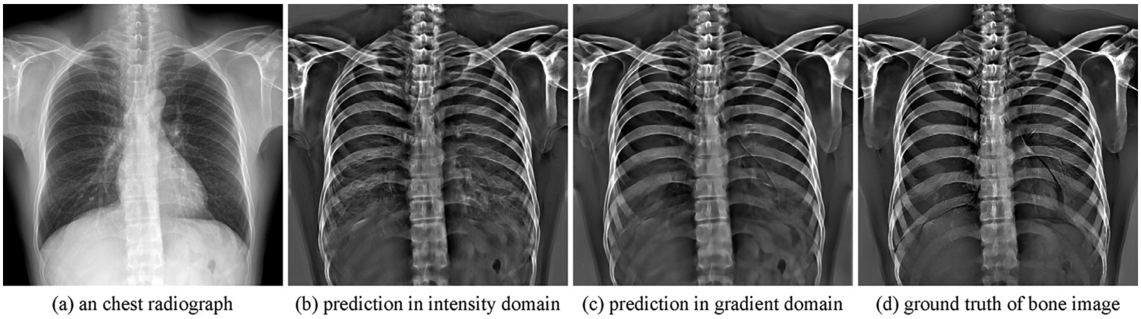

(a) an chest radiograph
(b) prediction in intensity domain
(c) prediction in gradient domain
(d) ground truth of bone image

Fig. 11. Bone images at scale 1/4 predicted by the 2-level CamsNets in the intensity domain and the gradient domain for an input CXR (a). (b) and (c) are the bone images produced with the CamsNets in the intensity domain and the gradient domain, respectively. (d) shows the corresponding ground truth of the bone image.

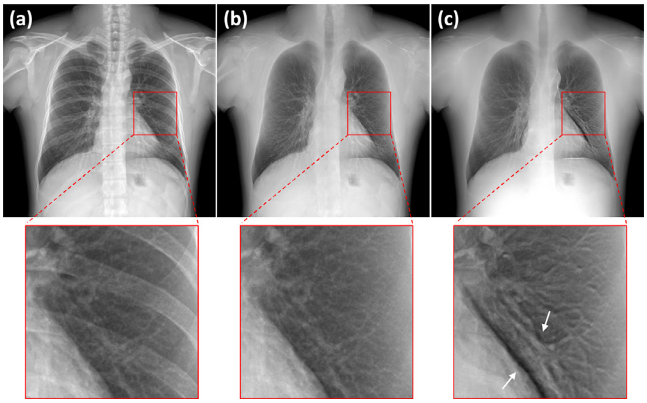


Fig. 12. Comparison of the results of bone suppression with the DES soft-tissue image. (a) shows a standard chest radiograph. (b) is the soft-tissue image produced with our method. (c) is the DES soft-tissue image. The regions in the red rectangles are zoomed in to display details. Two blurred and distorted structures caused by motion are indicated by white arrows in (c). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

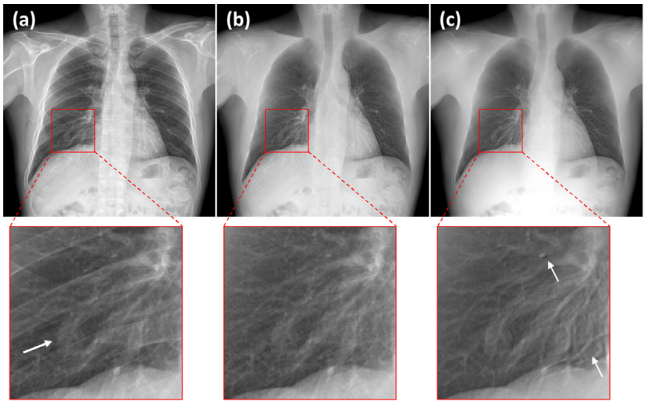


Fig. 13. Comparison of the results of bone suppression with the DES soft-tissue image. (a) shows a standard chest radiograph. (b) is the soft-tissue image produced with our method. (c) is the DES soft-tissue image. The regions in the red rectangles contain a nodule indicated by the white arrow in (a). Two blurred and distorted structures caused by motion are indicated by white arrows in (c). These regions are zoomed in to display details. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

but these residual pieces would not prevent clinicians from diagnosing lung diseases. Overall, the quality of the soft-tissue images produced by our method was comparable to DES soft-tissue images.

An advantage of our method is avoiding motion artifacts, which were usually present in the soft-tissue and bone images of two-exposure DES. As shown in the zoomed-in regions in Figs. 12(c) and 13(c), obvious motion artifacts were seen in the DES soft-tissue images. In contrast, few motion artifacts were seen in the soft-tissue images produced by our method because the regions with obvious motion artifacts were excluded from the training samples for the CamsNet, and the behavior of motion artifacts in real DES radiographs could not be learned and mimicked with the CamsNet. Therefore, the contents of the soft-tissue images produced by our method were more consistent with the CXR than the DES soft-tissue images. Fig. 13 shows the result of bone suppression for a nodule case. Fig. 13(b) shows that the rib edges across the nodule were suppressed, and the contrast and details of the nodule were maintained.

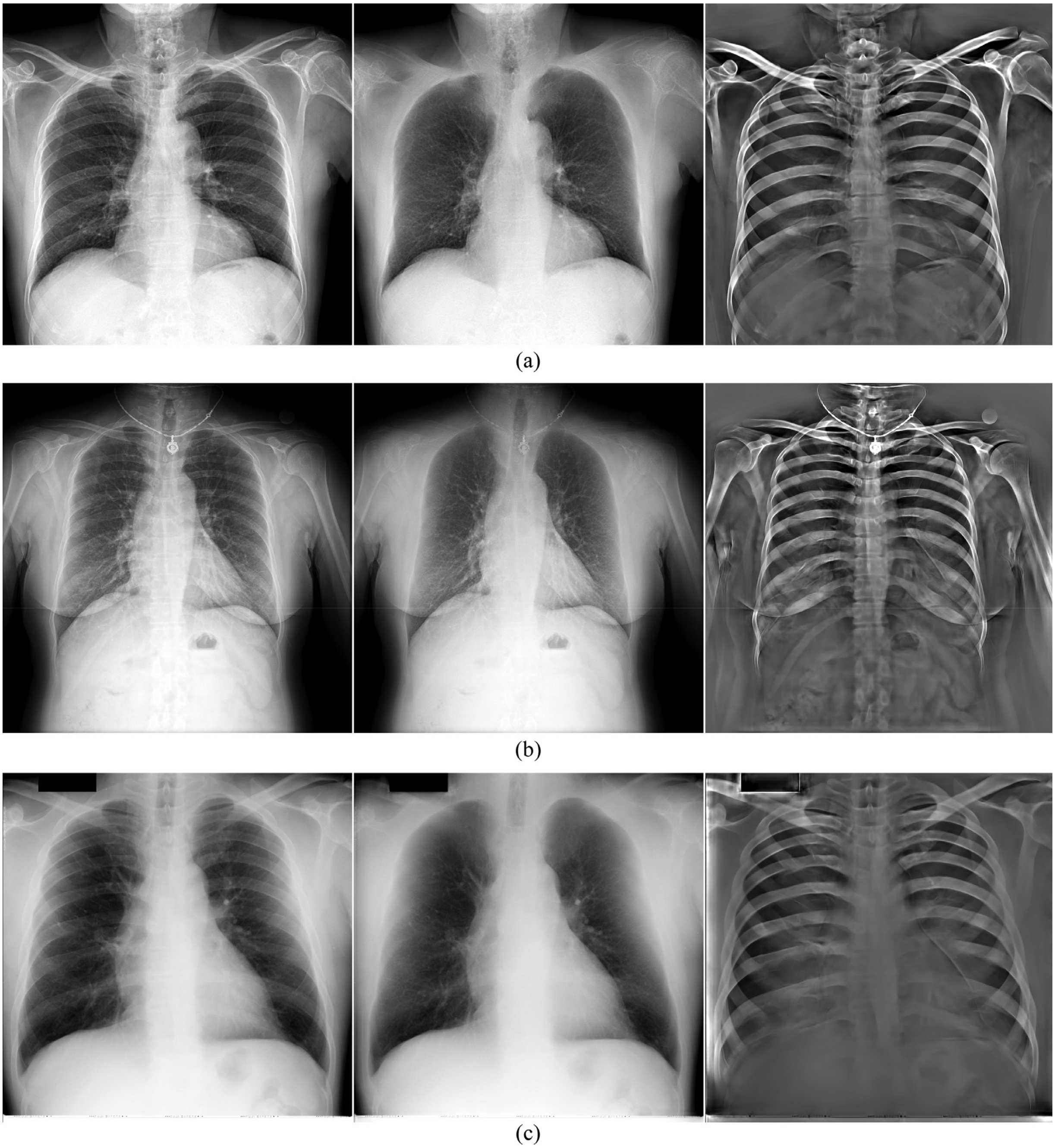
4.7. Generalization of CamsNet on CXRs acquired with other X-ray machines

In this section, we applied our method to predict the bone images and to produce the soft-tissue images from the CXRs acquired with different types of X-ray machines. The cross-dataset generalization ability of the CamsNet was evaluated quantitatively. The CXRs used in this study included the DR images acquired with Siemens FD-X (Siemens Healthcare) and SUNTO T-D3000 (SONTU Medical Imaging, Shenzhen, China). Scanned films from the publicly available JSRT dataset were included as well (Shiraishi et al., 2000). The CXRs were resized to a spatial resolution of 0.194 mm × 0.194 mm as the input for the CamsNet. Given the differences in X-ray tubes, flat detectors, and post-processing algorithms, the appearance of the CXRs were different, as shown in Fig. 14. In particular, the DR CXRs and the scanned films were significantly different.

The same CamsNet model and the multi-scale fusion procedure described in Section 4.6 were used in this section. Fig. 14. shows the soft-tissue and bone images produced with our method from three CXRs. Even the appearance of CXRs in Fig. 14 was different from the appearance of the training CXRs. Our method could provide visually appealing soft-tissue and bone images. In particular, clean and sharp bone images could be predicted from the scanned films of CXRs, in which the contrast was significantly lower than the DR CXRs. The bone components in the CXRs were effectively eliminated to produce the soft-tissue images, as shown in Fig. 14. These results proved the CamsNet model has good generalization ability. One reason underlying this outcome was that the gradients of CXRs acquired with different X-ray machines and post-processing algorithms were apparently similar.

5. Discussion

According to the quantitative and qualitative results, our proposed method, which predicts the bone gradients at multiple scales through the CamsNets in the gradient domain and merges



(a)

(b)

(c)

Fig. 14. Illustration of cross-dataset generalization of CamsNet model. From left to right: the input CXRs, the corresponding soft-tissue images, and bone images produced with our method. The input DR CXRs in (a) and (b) are acquired with Siemens FD-X and SUNTO T-D3000 systems, respectively. The input CXR in (c) is a scanned film from the JSRT dataset.

the predictions to estimate the bone images for bone suppression of the CXRs, is effective. Several strategies may contribute to the effectiveness of the full bone suppression procedure. First, ConvNet is the basic unit for predicting bone components. The end-to-end learning for large-scale samples of ConvNets allows the basic unit to extract effective image features and to have good prediction ability. Second, the cascade architecture of ConvNets makes the receptive field of the CamsNet large enough for information extraction and prediction of the corresponding bone components. Third, working in the gradient domain makes the ConvNets learn the mapping between the CXR and the corresponding bone component easily. Fourth, the fusion of multi-scale bone gradients in the MAP framework to estimate the bone image can adequately utilize the prediction information and further improve the quality of the estimated bone images.

Compared with the two-exposure DES, our bone suppression method has few motion artifacts in the soft-tissue and bone images. Given that the regions with obvious motion artifacts were excluded from the training samples for the CamsNet model, only the mapping between the regions of CXR and the corresponding bone component without motion artifacts was learned by the ConvNets. However, even the soft-tissue and bone components cannot be separated perfectly through using a DES system. Therefore, the soft-tissue images produced with our method were also not perfect. In some cases, some residual rib edges remained and can be observed in the soft-tissue images produced with our method.

Generally speaking, the trained models without extra knowledge and strategies cannot surpass the system for generating the training data.

Considering the generation method of the ground truth for training the models, the bone-suppressed results produced by the trained models can preserve most contents and details of the corresponding chest radiograph but cannot achieve the same result as DES soft-tissue images. Additionally, the quantitative evaluation metrics of different models from the experiments only reflect the learning and generalization ability of the models for the ground truth dataset. It is preferable that the raw data of DES chest radiographs be used as the ground truth for clinical use because the DES soft-tissue and bone images of raw data would linearly correlate to the corresponding chest radiographs. However, the raw data of DES chest radiographs are not currently available to us from commercial DES systems.

Compared with previous work on bone suppression through image processing techniques, our models were trained on a large number (404 cases) of DES radiographs. For example, the numbers of training cases in (Suzuki et al., 2006) and (Chen and Suzuki, 2014) were only five and nine. The models of deep ConvNets trained on large-scale samples can produce more reliable predictions. MTANN, which was proposed in (Suzuki et al., 2006) and extended in (Chen and Suzuki, 2014; Chen et al., 2016), learns the mapping between the patches of CXRs and the values of centered pixels in the corresponding bone images. Our ConvNets learn the mapping between the gradient patches of CXRs and bone images in the convolutional form. In addition, the predicted bone images can be progressively refined with ConvNets in CamsNet. Our method can achieve the finest resolution (pixel size: approximately 0.194 mm) for the predicted bone images. Techniques from previous studies could only produce the bone images with a relatively coarse resolution (e.g., approximately 0.8 mm in (Suzuki et al., 2006) and (Loog et al., 2006)). Moreover, our method for bone suppression does not require segmentation of lung fields or the complicated contrast normalization procedure for the input CXRs.

The prediction performance of the CamsNet model can be further improved in several ways. One direct way is to increase the depth (the number of convolution layers) or the width (the number of filters for the convolution layers) of ConvNets in a CamsNet to make the model more powerful. We can also train multiple CamsNet models with different settings, and their outputs can be combined as the final prediction. However, the running time might also increase for a bone suppression procedure using either a large or increased number of ConvNets. If we aim to produce smooth bone images and to maintain textural details in the predicted soft-tissue images, we can inject some prior terms into the MAP energy function in Eq. (4), such as sparsity and TV. It was observed that the quality of bone-suppressed results produced with a CamsNet model varied in different regions. Given that the characteristics in the different anatomical regions of CXRs and the corresponding bone images were significantly different, we can train specific CamsNets to predict the bone gradients in specific regions, such as lung fields, spines, and clavicles, which is similar to the methods proposed in (Chen and Suzuki, 2014) and (Chen et al., 2016). In principle, the complexity and uncertainty of the underlying mappings to be learned for a specific region should be lower than the whole CXR. Training specific CamsNets for different regions would improve the prediction performance to some extent. If we emphasize the contrast of nodules or abnormal regions, we can sample more training patches in these regions or set relatively large weights for these regions to train the CamsNet models.

Although visually appealing results can be produced with our method, the clinical usefulness of our method should be further evaluated. For example, clinical examination of a comparison of the bone suppression results with the standard CXRs for lung nodule detection and lung disease diagnosis should be conducted with radiologists. Another feasible way to qualitatively evaluate the bone suppression results is to rate image quality and clinical utility by radiologists. We also plan to implement some methods for computerized detection of lung nodules in the bone-suppressed images and the CXRs to compare detection performance. In the experiments, the quantitative evaluation of model performance was performed through a holdout procedure. A more reliable estimate for the performance could be obtained through a cross-validation procedure or using more samples as the test dataset to reduce a bias from case variations. The hyper-parameters of CamsNet, such as the number of filters and the learning rate, can be further optimized through the cross-validation procedure with the training dataset.

6. Summary

In this study, we applied deep ConvNets to predict the bone images from CXRs in the gradient domain. We proposed a cascade architecture for the multi-scale ConvNets to prompt prediction performance. The prediction quality of bone images was further improved by fusing the predicted multi-scale bone gradients in the MAP framework. Our method was validated with CXRs acquired with different X-ray machines through a large number of two-exposure DES radiographs to train the CamsNet models. The experimental results demonstrated that our proposed method for bone suppression of CXRs was effective and could produce high-quality results.

Acknowledgments

This work was supported by the grants from National Natural Science Foundation of China (no. 61471187), Natural Science Foundation of Guangdong Province (no. 2015A030313280), Guangdong Provincial Key Laboratory of Medical Image Processing (no. 2014B030301042), and the Excellent Young Teachers Program of Guangdong Colleges.

References

Agrawal, A., Raskar, R., Chellappa, R., 2006. Edge suppression by gradient field transformation using cross-projection tensors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2006), pp. 2301–2308.

Chen, S., Suzuki, K., 2013. Computerized detection of lung nodules by means of "virtual dual-energy" radiography. IEEE Trans. Biomed. Eng 60, 369–378.

Chen, S., Suzuki, K., 2014. Separation of bones from chest radiographs by means of anatomically specific multiple massive-training ANNs combined with total variation minimization smoothing. IEEE Trans. Med. Imaging 33, 246–257.

Chen, S., Zhong, S., Yao, L., Shang, Y., Suzuki, K., 2016. Enhancement of chest radiographs obtained in the intensive care unit through bone suppression and consistent processing. Phys. Med. Biol 61, 2283–2301.

Chen, Y., Chang, T.-c., Zhou, C., Fang, T., 2009. Gradient domain layer separation under independent motion. In: IEEE 12th International Conference on Computer Vision (ICCV). Kyoto, pp. 694–701.

Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K., 2007. Image denoising by sparse 3-d transform-domain collaborative filtering. IEEE Trans. Image Process. 16, 2080–2095.

Dong, C., Loy, C.C., He, K., Tang, X., 2016. Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. 38, 295–307.

Eigen, D., Fergus, R., 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. IEEE International Conference on Computer Vision (ICCV 2015).

Eigen, D., Krishnan, D., Fergus, R., 2013. Restoring an image taken through a window covered with dirt or rain. In: IEEE International Conference on Computer Vision 2013 (ICCV 2013), pp. 633–640.

Glorot, X., Bengio, Y., 2009. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256.

He, K., Sun, J., Tang, X., 2013. Guided image filtering. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1397–1409.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. IEEE International Conference on Computer Vision (ICCV 2015).

Hogeweg, L., Sanchez, C.I., Ginneken, B.V., 2013. Suppression of translucent elongated structures: applications in chest radiography. IEEE Trans. Med. Imaging 32, 2099–2113.

Knapp, J., Worrell, S., 2012. Feature based neural network regression for feature suppression, US8204292 B2. Riverain Medical Graoup, LLC, USA.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 1097–1105.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

Lee, J.-S., Wang, J.-W., Wu, H.-H., Yuan, M.-Z., 2012. A nonparametric-based rib suppression method for chest radiographs. Comput. Math. Appl. 64, 1390–1399.

Li, F., Hara, T., Shiraishi, J., Engelmann, R., MacMahon, H., Doi, K., 2011. Improved detection of subtle lung nodules by use of chest radiographs with bone suppression imaging: receiver operating characteristic analysis with and without localization. Am. J. Roentgenol. 196, W535–W541.

Loog, M., van Ginneken, B., Schilham, A.M.R., 2006. Filter learning: application to suppression of bony structures from chest radiographs. Med. Image Anal 10, 826–840.

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted Boltzmann machines. In: International Conference on Machine Learning, pp. 807–814.

Rasheed, T., Ahmed, B., Khan, M.A.U., Bettayeb, M., Lee, S., Kim, T.-S., 2007. Rib suppression in frontal chest radiographs: a blind source separation approach. 9th International Symposium on Signal Processing and Its Applications.

Riverain Technologies 2016, http://www.riveraintech.com/products/bone-suppression/

Schmidt, U., Jancsary, J., Nowozin, S., Roth, S., Rother, C., 2016. Cascades of regression tree fields for image restoration. IEEE Trans. Pattern Anal. Mach. Intell. 38, 677–689.

Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K., Matsui, M., Fujita, H., Kodera, Y., Doi, K., 2000. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. Am. J. Roentgenol. 174, 71–74.

Simko, G., Orban, G., Maday, P., Horvath, G., 2009. Elimination of clavicle shadows to help automatic lung nodule detection on chest radiographs. 4th European Conference of the International Federation for Medical and Biological Engineering (EMBEC).

Suzuki, K., Abe, H., Li, F., Doi, K., 2004. Suppression of the contrast of ribs in chest radiographs by means of massive training artificial neural network. In: Proc. SPIE, Medical Imaging 2004: Image Processing, pp. 1109–1119.

Suzuki, K., Abe, H., MacMahon, H., Doi, K., 2006. Image-processing technique for suppressing ribs in chest radiographs by means of massive training artificial neural network (MTANN). IEEE Trans. Med. Imag. 25, 406–416.

Tu, Z., Bai, X., 2010. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 32, 1744–1757.

van Ginneken, B., Stegmann, M.B., Loog, M., 2006. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. Med. Image Anal 10, 19–40.

Vedaldi, A., Lenc, K., 2015. MatConvNet - convolutional neural networks for MATLAB. the ACM International Conference on Multimedia.

Vock, P., Szucs-Farkas, Z., 2009. Dual energy subtraction: principles and clinical applications. Eur. J. Radiol. 72, 231–237.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13, 600–612.

Xu, L., Ren, J., Yan, Q., Liao, R., Jia, J., 2015. Deep edge-aware filters. The 32nd International Conference on Machine Learning.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. European Conference on Computer Vision.