DEEP FEATURE DISENTANGLEMENT LEARNING FOR BONE SUPPRESSION IN CHEST RADIOGRAPHS

Chunze Lin[⋆] *Ruixiang Tang*[⋆] *Darryl D. Lin*[†] *Langechuan Liu*[†] *Jiwen Lu*[⋆]
Yunqiang Chen[†] *Dashan Gao*[†] *Jie Zhou*[⋆]

[⋆]Tsinghua University, Beijing, China [†]12 Sigma Technologies, San Diego, USA

ABSTRACT

Suppression of bony structures in chest radiographs is essential for many computer-aided diagnosis tasks. In this paper, we propose a Disentanglement AutoEncoder (DAE) for bone suppression. As the projection of 3D structures of bones and soft tissues overlap in 2D radiographs, their features are interwoven and need to be disentangled for effective bone suppression. Our DAE progressively separates the features of soft-tissues from that of the bony structure during the encoder phase and reconstructs the soft-tissue image based on the disentangled features of soft-tissue. Bone segmentation can be performed concurrently using the separated bony features through a separate multi-task branch. By training the model with multi-task supervision, we explicitly encourage the autoencoder to pay more attention to the locations of bones in order to avoid loss of soft-tissue information. The proposed method is shown to be effective in suppressing bone structures from chest radiographs with very little visual artifacts.

Index Terms— Bone suppression, feature disentanglement, auto-encoders, deep learning.

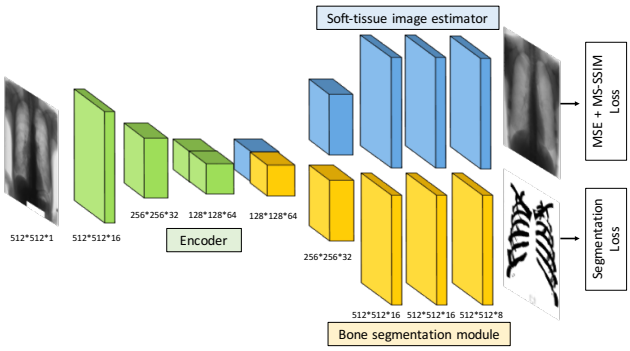

Fig. 1. Architecture of the DAE. The chest radiograph is first fed into the encoder that separates features of soft-tissue from bony structures. The decoders are composed of two branches: a soft-tissue image estimator attempts to reconstruct a soft-tissue image, and a bone segmentation module aims to highlight the bones.

1. INTRODUCTION

Chest radiography is a commonly used diagnostic imaging technique for identifying chest diseases since it is cost-effective, routinely available and safe. However, overlapping anatomical structures such as clavicles and ribs on top of soft tissues make the interpretation of chest radiographs difficult for radiologists. Accurate suppression of bones is therefore useful to simplify the tasks of radiologists [1] and improve the performance of computer-aided diagnosis methods [2].

Existing bones suppression methods can be mainly grouped into two categories: hardware based method such as dual-energy subtraction (DES) imaging [3] and software based method such as image processing techniques presented in [4]. DES radiography captures two radiographs with the use of two X-ray exposures at different energy levels. These radiographs are then combined to form a subtraction image that highlights either bone structures or soft tissues. Although DES systems produce high quality soft-tissue images, it requires specialized equipment with increased radiation dose and financial burden to the patients. On the other hand, the image processing techniques aim to suppress the bone structures in a given radiograph and estimate the soft-tissue image via a series of handcrafted features or learned features. Since these techniques do not require special equipment and can be applied as postprocessing to normal chest radiographs, they have attracted much attention over the past few years [4, 5, 6, 7, 8]. Most existing image processing methods treat bone suppression as a regression problem, where the regressors are trained to estimate the soft-tissue image from the chest radiograph. Suzuki *et al.* [4] proposed to suppress ribs in chest radiographs by means of massive training artificial neural network (MTANN). Loog *et al.* [7] presented a filter framework for suppression of bony structure by learning a complex non-linear filter directly from raw data. Recently, with the prevalence of deep learning, several methods employing deep convolutional neural networks (CNN) for bone suppression have been introduced. Yang *et al.* [6] proposed a cascade of multi-scale CNN for bone suppression in the gradient domain. Gusarev *et al.* [8] introduced two architecture of deep models to denoise bones from images.

In this paper, we propose a disentanglement autoencoder (DAE) model for bone suppression in chest radiographs. Unlike most existing methods which train the bone suppression model only based on reconstruction supervision, we simultaneously optimize our autoencoder to both segment the bony structures and estimate the soft-tissue image. An accurate bone suppression algorithm requires the model to identify the location of bony structures and discard them but retain the information about soft-tissue overlapping with the bones. The supervision from the segmentation branch guides our model to locate the bony structures and explicitly highlights the regions that require most attention for reconstruction. Moreover, in order to simultaneously succeed in both two tasks, our encoder should learn to disentangle the features of soft-tissues from the features of bones before feeding the encoded characteristics into the two decoder branches. Extensive experimental results on the publicly available JSRT dataset [9] demonstrate the effectiveness of the proposed method.

2. METHODOLOGY

Given a set of chest radiographs $\mathcal{X} = \{X_1, ..., X_N\}$ and corresponding soft-tissue images $\mathcal{S} = \{S_1, ..., S_N\}$, our goal is to learn an autoencoder to estimate the soft-tissue images by suppressing the bony structures while at the same time preserving the soft-tissues in the chest radiographs, including those that overlap with the bony structures.

2.1. Architecture overview

The bone suppression model is an adapted autoencoder architecture, composing of an encoder, a soft-tissue image estimator and a bone segmentation module. An overview of the architecture is illustrated in Fig. 1. Unlike conventional autoencoders, our model consists of two separate decoder branches: a soft-tissue image estimator to reconstruct a chest radiograph without anatomical structures, and a bone segmentation module to predicts the bone segmentation mask. More specifically, the encoder consists of 4 convolutional layers, where we progressively down-sample the spatial resolution via max-pooling operations and increasing the number of channels by a factor of 2 at each layer except the last one. The encoder takes as input chest radiographs of 512×512 spatial resolution and encodes the information into the feature maps of dimension $128 \times 128 \times 64$. During the decoding phase, we up-sample the spatial resolution through deconvolution layers to reconstruct a high resolution image and decrease the number of channels to match with image dimension.

2.2. Disentanglement Autoencoder (DAE)

A chest radiograph X can be seen as a composition of the bony structure B and the soft-tissue images S:

$$X = B + S$$

Our goal is to retrieve the soft-tissue image S from the chest radiograph. Most of existing methods force the networks to map X to S without any supplementary guidance except for the reconstruction supervision. The lack of precise supervision signal makes learning inefficient and leads to sub-optimal results. To address this issue, we propose a multi-task learning strategy to clearly guide the network to focus on the most important regions and separate the two components.

As shown in Fig. 1, we explicitly supervise the network to distinguish the model-extracted features of soft tissues from those of the bony structures. In order to succeed in both tasks, the soft-tissue image estimator would focus on information about soft tissues as input, while the bone segmentation module would concentrate on selecting features of bony structures. Consequently, multi-task supervision guides the autoencoder to perform feature disentanglement before reconstructing the target image. Thanks to the DAE architecture design, the model separates the features of these two components at the encoder phase, so that the decoders are able to select appropriate features to perform the segmentation and reconstruction tasks. Moreover, performing bone segmentation helps to explicitly highlight the regions that need to be reconstructed.

It is worth noting that the bone segmentation branch is an auxiliary branch for training that is turned off during inference. Thus the DAE does not require additional computation compared to the conventional autoencoder.

2.3. Optimization

Multiple loss functions are designed to supervise our disentanglement autoencoder:

$$L = L_{\text{MS-SSIM}} + \lambda_m L_{MSE} + \lambda_s L_{seg} \tag{1}$$

More specifically, $L_{\text{MS-SSIM}}$ and L_{MSE} are the loss functions designed for supervising the soft-tissue image estimator and L_{seg} corresponds to the supervision signal on the bone segmentation branch. λ_m and λ_s are the weight parameters to balance the loss terms. The formulation of these losses and design considerations are provided below.

MSE loss: A commonly used loss function for regression tasks is the mean-squared error (MSE), defined as:

$$L_{MSE} = \frac{1}{N} \sum_j ||S(j) - \tilde{S}(j)||_2 \tag{2}$$

where S and $\tilde{S}$ denote ground-truth and estimated soft-tissue image, respectively. N corresponds to the number of pixels in the image. The MSE loss aims to reduce the difference between the estimated and the true images at the pixel level. However, the MSE loss ignores the correlation between pixels and may lead to undesirable blurring of output images.

MS-SSIM loss: The multi-scale structural similarity index (MS-SSIM) [10] is used as complementary supervision

to improve the quality of image by considering local details. Specifically, the single-scale SSIM for a pixel j is defined as:

$$\text{SSIM}(j) = \frac{2\mu_S(j)\mu_{\tilde{S}}(j)}{\mu_S(j)^2 + \mu_{\tilde{S}}(j)^2} \cdot \frac{2\sigma_{S\tilde{S}}(j)}{\sigma_S(j)^2 + \sigma_{\tilde{S}}(j)^2} \quad (3)$$

$$= l(j) \cdot cs(j) \quad (4)$$

where the pairs $\{\mu_S(j), \sigma_S(j)\}$ and $\{\mu_{\tilde{S}}(j), \sigma_{\tilde{S}}(j)\}$ are the mean and standard deviation of two local neighborhoods centered at pixel index j, respectively. $\sigma_{S\tilde{S}}(j)$ is the covariance of the two local neighborhoods. Note that a small constant may be added to the denominators to avoid division by zero. SSIM compares the luminance l, contrast and structure cs between two local neighborhoods. In order to explore local details at multiple scales, SS-SSIM is extended to MS-SSIM:

$$\text{MS-SSIM}(j) = l_K^{\alpha}(j) \cdot \prod_{k=1}^{K} cs_k^{\beta_k}(j) \quad (5)$$

where l_K and cs_k are the luminance, contrast and structure at scale K and k, respectively. The loss function based on MS-SSIM is defined as:

$$L_{\text{MS-SSIM}} = \frac{1}{N} \sum_j 1 - \text{MS-SSIM}(j) \quad (6)$$

Segmentation loss: For training the network to estimate the bone segmentation mask, we employ the cross-entropy loss between the ground-truth segmentation mask M and the predicted mask $\tilde{M}$:

$$L_{seg} = -\frac{1}{N} \sum_j M_j \ln \tilde{M}_j + (1 - M_j) \ln(1 - \tilde{M}_j) \quad (7)$$

3. EXPERIMENTS AND RESULTS

3.1. Dataset

We conducted our experiments on the publicly available JSRT dataset [9], consisting of 247 pairs of chest radiographs. The corresponding soft-tissue images are provided by [11] and we obtain the bone segmentation mask by subtracting radiographs with soft-tissue images followed by binarization. We divided the data into 85% for training and 15% for testing. A series of data augmentation including rotations, horizontal and vertical shifts, shear, zoom, and horizontal flips were used to increase the number of training samples. In the MS-SSIM loss function, the hyperparameters were set to $\alpha = 1, \beta_1 = 0.0048, \beta_2 = 0.2856, \beta_3 = 0.3001, \beta_4 = 0.2363$ and $\beta_5 = 0.1333$ following [10].

3.2. Quantitative analysis

The performance of our bone suppression model was evaluated in terms of four commonly used metrics including bone suppression ratio (BSR), relative mean absolute error (RMAE), peak signal-to-noise ratio (PSNR), and structural similarity (SSIM) index. In order to demonstrate the effectiveness of the proposed architecture and the associated loss function, we trained four different bone suppression models: original autoencoder, autoencoder with MS-SSIM loss, disentanglement autoencoder, and disentanglement autoencoder with MS-SSIM loss.

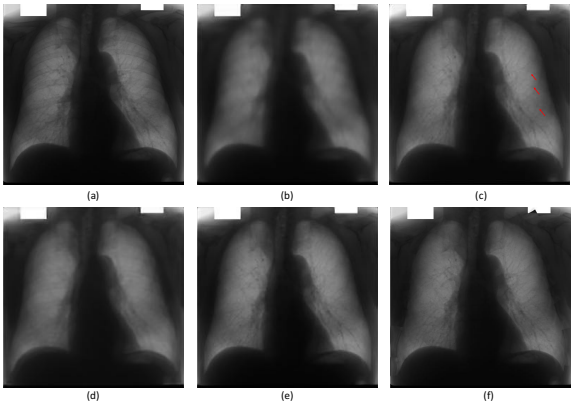

Fig. 2. Qualitative analysis of soft-tissue image generated by different models. (a) Chest radiograph. (b) Image estimated by original autoencoder. (c) Image estimated by standard autoencoder with MS-SSIM. (d) Image estimated by disentanglement autoencoder. (e) Image estimated by disentanglement autoencoder with MS-SSIM. (f) Ground-truth bone-free radiograph.

Table 1. The performance of bone suppression models with different design choices evaluated with four metrics.

Method	BSR	RMAE	PSNR	SSIM
AE	0.75	0.0309	25.9	93.5
AE + MS-SSIM	0.85	0.0109	29.0	96.1
DAE	0.82	0.0290	26.5	94.1
DAE + MS-SSIM	0.91	0.0091	30.8	97.0

The experimental results are reported in Table. 1. MS-SSIM supervision significantly improves the quality of the estimated image, reducing RMAE and increasing BSR, PSNR and SSIM for both standard autoencoder and disentanglement autoencoder. Using the same setting as the original autoencoder, the disentanglement autoencoder achieves better results, especially under the most important BSR criterion. This performance shows that the proposed method can more effectively suppress the bony structures from chest radiographs than the standard autoencoder.

3.3. Qualitative comparison

In order to better interpret the performance of different models, we qualitatively compared the estimated soft-tissue im-

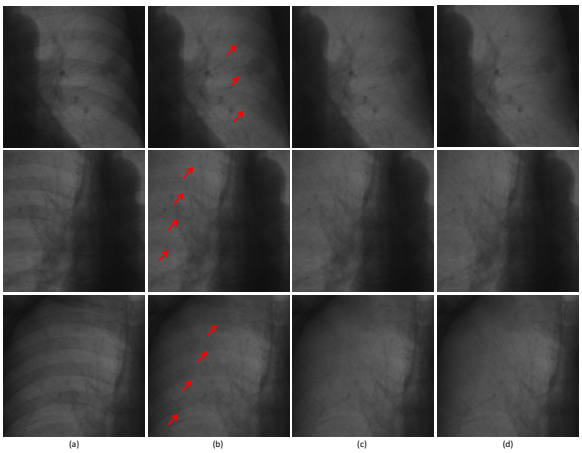

Fig. 3. Visualization of zoomed in images to display details. (a) Chest radiograph. (b) Estimated by standard autoencoder + MS-SSIM. (c) Estimated by our disentanglement autoencoder + MS-SSIM. (d) Ground-truth bone-free radiograph.

ages. As Fig. 2 shows, when using the MSE loss alone to supervise the networks, we obtain blurry output images (b) and (d). By adding the MS-SSIM supervision, the quality of images is significantly improved as shown in (c) and (e). As indicated by the red arrows, the standard autoencoder fails to completely suppress bony structures and the shadows of ribs are still visible. In contrast, the disentanglement autoencoder effectively suppress these shadows.

Zoomed-in patches on different images are illustrated in Fig. 3 to visualize the details of the estimated images. The bony structures are almost entirely removed from the radiographs while the soft-tissue information are preserved by our disentanglement autoencoder (Fig. 3(c)).

4. CONCLUSION

In this paper, we proposed a disentanglement autoencoder model for bone suppression in chest radiographs. The disentanglement autoencoder is trained in a multi-task learning manner with auxiliary supervision. During training, the model is encouraged to simultaneously estimate the soft-tissue image and predict the bone segmentation mask. In this manner, the autoencoder is guided to progressively disentangle the features of soft-tissues from bones and significantly improve the quality of the estimated bone-free chest radiographs. During inference, the bone segmentation mask can be disabled to reduce the computation burden to the same level of a conventional auto-encoder.

5. REFERENCES

[1] Feng Li, Takeshi Hara, Junji Shiraishi, Roger Engelmann, Heber MacMahon, and Kunio Doi, "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.*, 2011.

[2] Sheng Chen and Kenji Suzuki, "Computerized detection of lung nodules by means of "virtual dual-energy" radiography," *Trans. Biomed. Eng.*, 2013.

[3] Peter Vock and Zsolt Szucs-Farkas, "Dual energy subtraction: principles and clinical applications," *Eur. J. Radiol.*, 2009.

[4] Kenji Suzuki, Hiroyuki Abe, Heber MacMahon, and Kunio Doi, "Image-processing technique for suppressing ribs in chest radiographs by means of massive training artificial neural network (mtann)," *Trans. Med. Imaging*, 2006.

[5] Sheng Chen and Kenji Suzuki, "Separation of bones from chest radiographs by means of anatomically specific multiple massive-training anns combined with total variation minimization smoothing," *Trans. Med. Imaging*, 2014.

[6] Wei Yang, Yingyin Chen, Yunbi Liu, Liming Zhong, Genggeng Qin, Zhentai Lu, Qianjin Feng, and Wufan Chen, "Cascade of multi-scale convolutional neural networks for bone suppression of chest radiographs in gradient domain," *Medical image analysis*, 2017.

[7] Marco Loog, Bram van Ginneken, and Arnold MR Schilham, "Filter learning: application to suppression of bony structures from chest radiographs," *Med. Image Anal.*, 2006.

[8] Maxim Gusarev, Ramil Kuleev, Adil Khan, Adin Ramirez Rivera, and Asad Masood Khattak, "Deep learning models for bone suppression in chest radiographs," in *CIBCB*, 2017.

[9] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi, "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," *American Journal of Roentgenology*, 2000.

[10] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multiscale structural similarity for image quality assessment," in *Asilomar Conference on Signals, Systems & Computers*, 2003.

[11] S Juhász, Á Horváth, L Nikházy, and G Horváth, "Segmentation of anatomical structures on chest radiographs," in *MEDICON*, 2010, pp. 359–362.