Medical Image Analysis 67 (2021) 101839

Contents lists available at ScienceDirect

Medical Image Analysis

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

A disentangled generative model for disease decomposition in chest X-rays via normal image synthesis

Youbao Tang[a,*], Yuxing Tang[a], Yingying Zhu[a], Jing Xiao[b], Ronald M. Summers[a]

[a] Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Bethesda, MD 20892-1182, USA
[b] Ping An Insurance Company of China, Shenzhen, 510852, China

ARTICLE INFO

Article history:
Received 7 November 2019
Revised 28 September 2020
Accepted 2 October 2020
Available online 7 October 2020

Keywords:
Chest radiography (X-ray)
Medical image synthesis
Disentangled representation learning
Disease decomposition

ABSTRACT

The interpretation of medical images is a complex cognition procedure requiring cautious observation, precise understanding/parsing of the normal body anatomies, and combining knowledge of physiology and pathology. Interpreting chest X-ray (CXR) images is challenging since the 2D CXR images show the superimposition on internal organs/tissues with low resolution and poor boundaries. Unlike previous CXR computer-aided diagnosis works that focused on disease diagnosis/classification, we firstly propose a deep disentangled generative model (DGM) simultaneously generating abnormal disease residue maps and "radiorealistic" normal CXR images from an input abnormal CXR image. The intuition of our method is based on the assumption that disease regions usually superimpose upon or replace the pixels of normal tissues in an abnormal CXR. Thus, disease regions can be disentangled or decomposed from the abnormal CXR by comparing it with a generated patient-specific normal CXR. DGM consists of three encoder-decoder architecture branches: one for radiorealistic normal CXR image synthesis using adversarial learning, one for disease separation by generating a residue map to delineate the underlying abnormal region, and the other one for facilitating the training process and enhancing the model's robustness on noisy data. A self-reconstruction loss is adopted in the first two branches to enforce the generated normal CXR image to preserve similar visual structures as the original CXR. We evaluated our model on a large-scale chest X-ray dataset. The results show that our model can generate disease residue/saliency maps (coherent with radiologist annotations) along with radiorealistic and patient specific normal CXR images. The disease residue/saliency map can be used by radiologists to improve the CXR reading efficiency in clinical practice. The synthesized normal CXR can be used for data augmentation and normal control of personalized longitudinal disease study. Furthermore, DGM quantitatively boosts the diagnosis performance on several important clinical applications, including normal/abnormal CXR classification, and lung opacity classification/detection.

Published by Elsevier B.V.

1. Introduction

The chest radiography (chest X-ray, or CXR) has been one of the most common diagnostic imaging tests. It is frequently used for emergency diagnosis and treatment of thoracic diseases such as pneumonia, emphysema and cancer throughout the world, for its low cost, less scanning time (CXR: $<$ 15 min, CT: 15–30 min, MRI: 10–60 min) and less radiation exposure comparing to CT (about 200 times less than CT) (Smith-Bindman et al., 2008). Developing computer-aided diagnosis/detection (CADx/CADe) methods for CXR could potentially benefit radiologists/physicians by reducing their workload, improving their working efficiency and reducing the inter-observer variability in patient care that is currently widespread (Qin et al., 2018; Horváth et al., 2009). With the availability in recent years of CXR imaging data and accompanying text-mined labels using natural language processing (NLP) (Wang et al., 2017; Goldberger et al., 2000; Rajpurkar et al., 2017), many works are proposed to use deep neural networks to analyze CXR images. Published methods include supervised disease classification (Dunnmon et al., 2019; Tang et al., 2020), weakly supervised thoracic disease localization (Tang et al., 2018b; Cai et al., 2018), lung segmentation (Tang et al., 2019b) and disease ontology prediction (Chen et al., 2019b).

Although previous works show promising results, very limited study has been done for improving the interpretability of current deep learning models on CXR images (Yasaka and Abe, 2018). Recognition of abnormal patterns/regions associated with diseases

* Corresponding author.
E-mail address: youbao.tang@nih.gov (Y. Tang).

https://doi.org/10.1016/j.media.2020.101839
1361-8415/Published by Elsevier B.V.

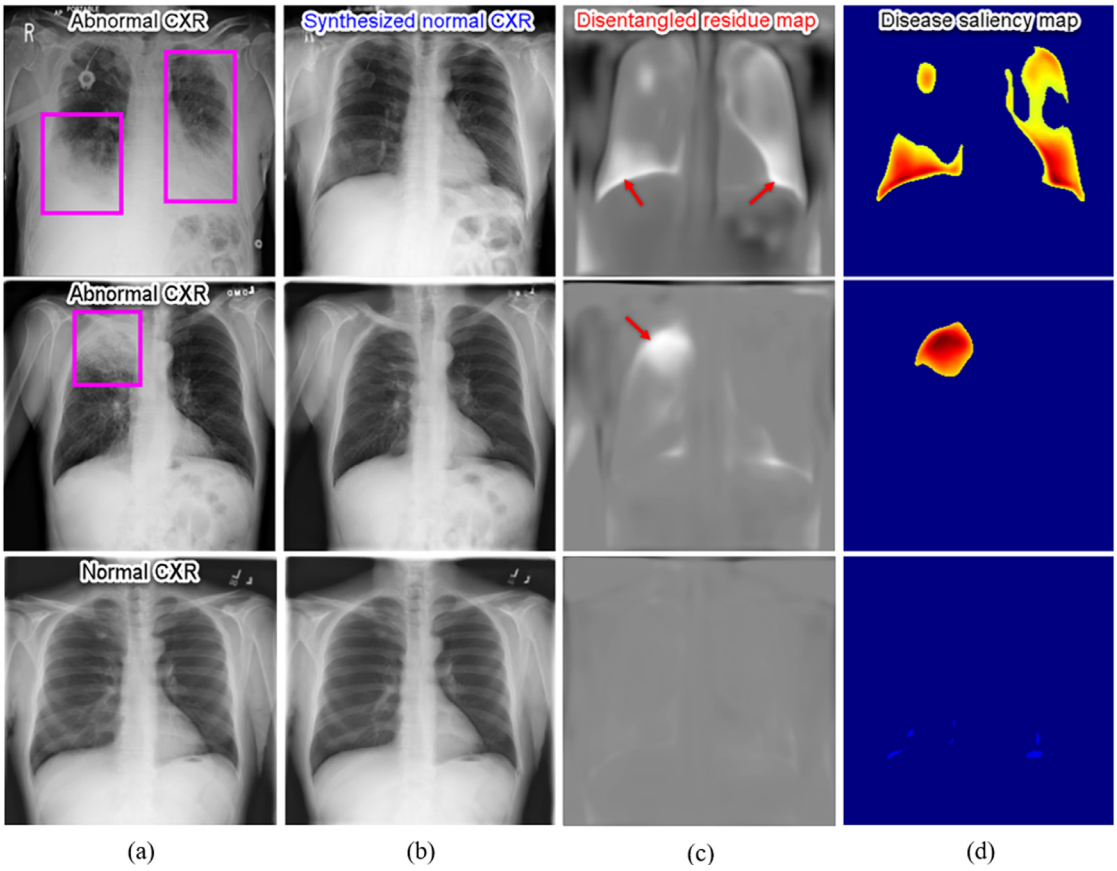

Fig. 1. An illustration of our proposed disentangled generative deep model for CXR synthesis and diagnosis. The model is capable of synthesizing radiorealistic and person-specific normal CXR images (b) and disentangling residue maps to delineate the potential disease patterns (c) from the input abnormal/normal CXR images (a). The disentangled residue maps (computed by our generative model) can be normalized and effectively visualized as disease saliency or attention maps (d), allowing radiologists to easily focus on underlying abnormalities for further depiction. Here, for the saliency map generation, we first define a bounding box from the segmented lung regions obtained by XLSor (Tang et al., 2019b). Then we binarize the residue map with an adaptive threshold. Only the connected components overlapped with the defined bounding box are shown for saliency visualization. It is worth noting that the extracted disease saliency map is consistent with the radiologist annotations (the pink boxes in (a)). Better viewed in color.

is a critical part of CXR interpretation. CXR is a complex cognition procedure that requires cautious observation, precise understanding/parsing of the normal thoracic anatomy, and knowledge of physiology and pathology. Adding to the challenge is the need to identify the abnormal disease regions and infer 3D anatomy from the 2D projection CXR images that have low resolution and contrast (Khan et al., 2009), intrinsically overlaid internal body/tissue structures and limited organ textures/boundaries (Fig. 1 (a)).

To visualize the spatial extent of diseases/pathologies in the abnormal CXRs, in this work, we develop a disentangled representation learning framework called deep *disentangled generative model* (DGM) to decompose the abnormal CXR images into a pair of normal synthetic images (Fig. 1 (b)) and their abnormal residue maps associated with diseases (Fig. 1 (c)). The extracted abnormal disease regions can also assist radiologists to interpret CXR images and further improve the disease diagnosis accuracy. When inputting a normal CXR, the synthesized normal image should be identical to the input, and the disentangled residue map should be minimal or nearly blank (the third row of Fig. 1). The assumption is that an abnormal CXR image is composed of a normal one and superimposed or replaced abnormal regions. This is depended on the fact that a CXR image is generated based on the absorption of X-ray beam passing through the human body. Generally, dark pixel (air) indicates low X-ray absorption and bright pixel indicated high X-ray absorption (bones). For example, normal lung is full of air and appears black (Fig. 1(b)). When someone has a lung opacity, one or both lungs is partially filled with foreign material (*e.g.*, fluid or pus: whiter pixels) which absorbs more x-rays than air. Therefore, the abnormal lung may appear whiter than normal due to the overlaid brighter foreign material (see Fig. 1(a)). Moreover, for many thoracic diseases, the pathological structures are replacing healthy tissues. For example, consolidation or infiltrate is typically an area of white lung (or a density that resembles fluffy clouds) on a standard X-ray, because consolidated tissue is more radio-opaque (looks brighter) than normally aerated lung parenchyma (looks darker). In this and many other similar cases, we assume that the brighter pixels superimpose upon or replace the pixels of normal tissues and the synthesis of a normal chest X-ray from an abnormal CXR image as an image inpainting process (Yu et al., 2018).

As collecting paired normal and abnormal CXR images from the same patient is impractical in real clinical applications, the proposed framework only requires unpaired CXR images with image-level labels. It is able to disentangle potential abnormal patterns and extents from any input CXR images (abnormal/normal) and synthesize high-quality, radiorealistic (i.e., a synthesized radiograph that appears anatomically correct and realistic on soft tissue structures such as the lungs and shape of the heart, other bony structures, and various disease patterns when applicable.) and patient specific normal CXR images. Therefore, when inputting an abnormal CXR with thoracic diseases, the synthesized normal CXR image is treated as the status when the patient is disease free or

healed completely. The disentangled (or decomposed, separated) abnormal regions reveal the disease patterns amd extents from the input CXR image. We apply our model on three different diagnostic tasks: normal vs. abnormal CXR classification, lung opacity vs. non-lung opacity CXR classification, and lung opacity detection. The experimental results show that the disentangled abnormal disease patterns improve the diagnostic performance.

The main contributions of this work can be summarized as below: 1) A novel disentangled generative deep model is designed for chest X-ray decomposition. 2) It can synthesize radiorealistic normal chest X-ray images and produce disease residue maps to indicate the underlying abnormal regions for interpretation explicitly and simultaneously. 3) Using the generated residue maps can quantitatively boost the diagnosis performance on several important clinical applications.

2. Related work

CADx/CADe on chest radiography. Many works (Rajpurkar et al., 2018; Li et al., 2018; Cai et al., 2018; Guan and Huang, 2020) have attempted to classify and localize 14 main thoracic diseases using deep learning on the large-scale NIH chest X-ray dataset (Wang et al., 2017). Most works utilize one or more deep CNN architectures with modifications for multi-label disease classification. Without bounding boxes for training, these methods commonly generate class-specific heatmap with global average pooling (Zhou et al., 2016) for disease localization using the same framework used for classification. TieNet (Wang et al., 2018b) uses a multi-level attention CNN-RNN architecture to learn a blend of image and text representations from both images and radiology reports, for classification and report generation. (Salehinejad et al., 2018) synthesizes CXR images with diseases (pathology) used for data augmentation in training deep CNNs. (Liang et al., 2020) generates bone suppressed radiographs that should be obtained by dual-energy imaging technique from conventional CXR images.

One criticism is that the performance of all these 14 diseases classification works are very low (about 82% of AUC), which limits its clinical relevance. A potentially more clinically practical application at the present state of the technology is to solve the two-class problem (normal/abnormal) (Tang et al., 2020). The radiologist can check these abnormal CXR images for more detailed analysis. Depending on the clinical setting, this can substantially reduce the work load. Our method adopts the normal/abnormal classification framework and achieves consistent performance improvements on disease classification and location compared to previous works (our two class classification AUC $>$ 96%).

Generative adversarial networks. The basic generative adversarial network (GAN) (Goodfellow et al., 2014) includes two parts: a generator which takes a random noise vector from a latent space as input and produces new samples in the data space and a discriminator which identifies whether the generated new samples come from the true data distribution. An adversarial loss balances the trade-off between generator and discriminator. GANs (Goodfellow et al., 2014; Radford et al., 2016; Reed et al., 2016; Ledig et al., 2017) have been successfully applied to generate artificial data indiscernible from their real counterparts in many applications such as image style transfer(Chang et al., 2018), image-to-image transfer (Zhu et al., 2017), image super-resolution (Ledig et al., 2017) and text to image synthesis (Reed et al., 2016).

The original GAN requires paired source and target domain data, which limits its application in this case because paired data from the same patient is not always available. To avoid this issue, Zhu et al. (2017) trained GANs using unpaired images by leveraging cycle consistency to regularize the training with unpaired images. Recently, many works proposed variants of GANs to synthesize medical images for data augmentation (Tang et al., 2018a; Jin et al., 2018; Tang et al., 2019c; 2019b), in order to alleviate the most commonly existing issues in medical image analysis: data scarcity (Tang et al., 2019e; 2019d), over-fitting, and domain divergence (Tang et al., 2019a; Zhu et al., 2020). Baumgartner et al. (2018) proposed a novel GAN-based feature attribution technique (VA-GAN) to produce compellingly realistic disease effect maps for Alzheimer's disease patients. In this work, we focus using GAN conditioned on an input CXR image. Our model produces normal CXR and disentangled potential disease map as outputs without requiring paired data.

Image-to-image translation and disentangled representation learning. The goal of Image-to-image (I2I) translation is to learn a mapping function from a source image domain to a target image domain. Pix2pix (Isola et al., 2017) adopts paired training data and applies a conditional GAN to model the mapping process. CycleGAN (Zhu et al., 2017), MUNIT (Huang et al., 2018) and DRIT (Lee et al., 2018) leverage cycle consistency to regularize the training with unpaired images. Many previous works address the I2I problem using disentangled representation learning framework. For example, DRIT (Lee et al., 2018) disentangles an image into domain-invariant and domain-specific representations to facilitate the learning diverse cross-domain mappings. Bao et al. (2018) disentangles the identity and attributes of faces for identity preserving face synthesis in an open-set (*i.e.*, a face with identity outside the training set). Prior to deep learning, researchers developed intensity-based methods and modeled geometric relationship among different subjects' anatomies to synthesize medical images. Ye et al. (2013) propose a general framework for modality synthesis with a local patch-based data-driven regularization. The synthetic image is applied for abnormality detection in multi-channel brain MRI. Liu et al. (2014) registered a healthy brain atlas with the low-rank and sparse components from the original MRI in an iterative manner, to segment brain tissue from patients with large pathologies. Bowles et al. (2016) propose a modality transformation technique to generate a subject-specific disease-free image that is not present in the input modality. Recently, disentangled representation learning has been also studied in medical image analysis. Chartsias et al. (2019) proposed a Spatial Decomposition Network (SDNet) which decomposes 2D cardiac images into spatial anatomical factors and non-spatial modality factors. SDNet could synthesize computed tomography (CT) from magnetic resonance (MR) and vice versa by swapping the modality factors. Qin et al. (2019) proposed an unsupervised image-to-image translation to reduce the multi-modal registration problem to a mono-modal one through image disentangling. To deal with missing imaging modalities for multimodal brain lesion segmentation, Chen et al. (2019a) decomposed the input modalities into modality-specific appearance code and modality-invariant content code, using feature disentanglement and gated feature fusion. Our work is conceptually similar to (Lee et al., 2018; Bao et al., 2018; Chartsias et al., 2019; Bowles et al., 2016; Qin et al., 2019; Chen et al., 2019a) for disentangled feature learning using encoder-decoder networks, but we proposed two extra decoders to make the generated normal CXRs more "radiorealistic" and also to facilitate training.

3. CXR decomposition using a disentangled generative model

In this section, we first define the problem with all related notations. Then we describe the detailed structure of DGM and details of each component. Lastly, we present a joint loss function for DGM to be trained end-to-end.

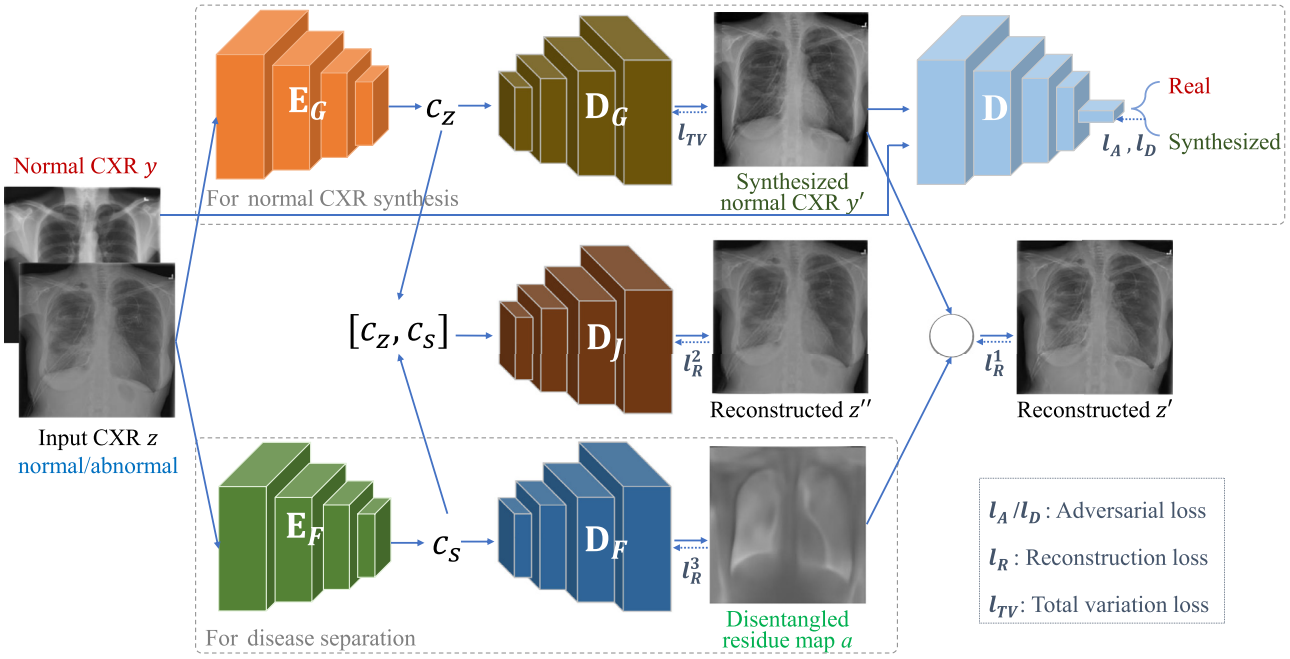

Fig. 2. Framework of the proposed disentangled generative model (DGM). An input chest X-ray image is decomposed into a normal CXR and a disease residue map by the DGM, which is trained in an end-to-end manner. The DGM training only requires image-level annotations of the input CXRs, *i.e.*, normal or abnormal. The input and output are shown in solid arrows, while the objective functions are shown in dashed arrows.

3.1. Problem formulation

Let $x \in \mathbb{X}$ and $y \in \mathbb{Y}$ denote images from two different domains, abnormal and normal. In the traditional image-to-image translation task (*e.g.*, CycleGAN (Zhu et al., 2017) and UNIT (Liu et al., 2017)), the goal is to learn a mapping $G : \mathbb{X} \mapsto \mathbb{Y}$, such that the distribution of images from $G(\mathbb{X})$ is indistinguishable from the distribution of images from $\mathbb{Y}$. Here, we are interested in performing this task in open domains, that is, the input image z may come from either domain, *i.e.*, $z \in \{\mathbb{X} \bigcup \mathbb{Y}\}$.

We assume that an abnormal CXR is composed of a normal one with abnormalities (disease part) superimposed upon or replaced with it. For a normal CXR, its disease part should be blank and its normal part should be similar to the original CXR, *e.g.*, the lung shape, the heart position, etc. Based on the above assumptions, the goals of our learning problem are twofold: 1) to synthesize a normal CXR by learning a mapping $G : \{\mathbb{X} \bigcup \mathbb{Y}\} \mapsto \mathbb{Y}$ such that the output $y' = G(z)$ is indistinguishable from images $y \in \mathbb{Y}$, and 2) to disentangle a residue map a by learning a mapping F that satisfies $a = F(z)$, $z' = G(z) + F(z)$ and $z' = z$. The residue map can be further normalized to a disease saliency/attention map visualizing underlying abnormalities (Fig. 1 (c) and (d)). Therefore, our learning problem can be formulated as a decomposition task.

Generally, for an abnormal CXR, it is difficult to get its normal version (before suffering diseases). Thus, it is not easy to explicitly learn the mappings, G and F. We first encode the input CXR z into a latent feature space $\mathcal{L} \in \mathbb{R}^{n \times k \times k}$ using disentangled representation learning, where n and k are the dimensions of feature maps in $\mathcal{L}$. In $\mathcal{L}$, the feature representations of normal CXR and its disease counterpart are denoted as c_z and c_s, respectively. Then we generate the normal CXR y' from c_z using adversarial learning and the disentangled residue map a from c_s with a constraint that the difference between $(y' + a)$ and z should be as small as possible. Additionally, we directly reconstruct the input CXR z from c_z and c_s to facilitate the learning process.

3.2. Network structure

To learn the mapping G and F, as shown in Fig. 2, the proposed DGM framework contains six components: 1) E_G, encodes a CXR z into a latent feature space c_z representing normal CXR attributes; 2) E_F, encodes z into another latent feature space c_s representing abnormal attributes (residue map); 3) D_G, decodes the normal part of z from c_z; 4) D_F, decodes the abnormal part of z from c_s; 5) D_J, decodes the input CXR z jointly from c_z and c_s; 6)D, discriminates real normal CXRs and synthesized ones generated by D_G. All six components are trained in an end-to-end fashion. Although the encoders and decoders have similar structures, their weights are not shared so that each module will avoid learning highly correlated information.

The encoders E_G and E_F contain several strided convolutional layers for downsampling the input CXR and several residual blocks for further processing it. All the convolutional layers are followed by Instance Normalization (IN) (Ulyanov et al., 2017) in E_G. We do not use IN layers in E_F, since IN removes the original feature mean and variance that represent important disease information.

The decoders D_G, D_F and D_J decode the latent feature representations c_z and c_s by several residual blocks, to produce the generated images (the normal CXR $y' = D_G(c_z)$, the residue map $a = D_F(c_s)$ and the reconstructed input CXR $z'' = D_J(c_z, c_s)$) by several upsampling and convolutional layers. The decoding process is the reverse process of encoding. Specifically, with encoded features $\{c_z, c_s\}$, the decoder D_J should jointly decode them back to the original input CXR z. That is, $z'' = D_J(E_G(z), E_F(z))$. Similar to MUNIT (Huang et al., 2018), the decoder D_J first uses a multi-layer perceptron (MLP) to produce a set of AdaIN (Huang and Belongie, 2017) parameters from c_s, based on which, the concatenation of c_z and c_s in channel dimension is then processed by residual blocks with AdaIN layers. When the input CXR is normal, the generated residue map a should be blank, *i.e.*, $a = \mathbf{0}$.

The discriminator D aims to distinguish the real normal CXRs from the fake ones produced by D_G, while D_G attempts to generate

radiorealistic normal CXRs. Here, we employ multi-scale discriminators (Wang et al., 2018a) to guide D_G to preserve both realistic details and correct global structure.

3.3. Loss functions

The objective is to jointly optimize all the components in an end-to-end manner. To achieve this, we design different loss functions:

Adversarial loss. To generate "radiorealistic" normal CXRs that are indistinguishable from real ones, the least squares loss (Mao et al., 2017), proved to be more stable than sigmoid cross entropy loss, is used as the adversarial learning loss:

$$\ell_A = \frac{1}{2}\mathbb{E}[(D(D_G(E_G(z))) - \mathbf{1})^2] \tag{1}$$

Reconstruction losses. DGM will output two reconstructed images, *i.e.*, z' and z'', from the input CXR z, namely, $z' = D_G(E_G(z)) + D_F(E_F(z))$ and $z'' = D_J(E_G(z), E_F(z))$. The reconstruction loss ℓ_R^1/ℓ_R^2 is defined as the $\mathcal{L}1$ distance:

$$\ell_R^1 = \|z' - z\|_1 = \|D_G(E_G(z)) + D_F(E_F(z)) - z\|_1 \tag{2}$$

$$\ell_R^2 = \|z'' - z\|_1 = \|D_J(E_G(z), E_F(z)) - z\|_1 \tag{3}$$

since it encourages sharper generated CXRs with less blurring than $\mathcal{L}2$.

DGM also generates a residue map a, which should be $\mathbf{0}$ when the input CXR is normal. Hence, we penalize a residue map for a normal CXR $z \in \mathbb{Y}$ by:

$$\ell_R^3 = \begin{cases} 0 & \text{if } z \in \mathbb{X} \\ \|a\|_1 = \|D_F(E_F(z))\|_1 & \text{if } z \in \mathbb{Y} \end{cases} \tag{4}$$

Total variation loss. To encourage spatial smoothness of the generated normal CXRs and reduce spike artifacts, we follow (Johnson et al., 2016) and use another auxiliary loss function called the total variation loss ℓ_{TV}. It performs total variation regularization on $y' = D_G(E_G(z))$ defined by

$$\ell_{TV} = \sum_{i,j} |y'_{i+1,j} - y'_{i,j}| + |y'_{i,j+1} - y'_{i,j}| \tag{5}$$

where (i, j) represents the image coordinate.

The final objective $\ell(G, F, D_J)$ for encoders and decoders optimization is a weighted sum of these losses:

$$\ell(G, F, D_J) = \lambda_A \ell_A + \lambda_R(\ell_R^1 + \ell_R^2 + \ell_R^3) + \lambda_{TV}\ell_{TV} \tag{6}$$

where λ_A, λ_R and λ_{TV} are the weights that control the importance of the different losses.

For discriminator D optimization, the LSGAN (Mao et al., 2017) objective is used:

$$\ell_D = \frac{1}{2}\mathbb{E}[(D(D_G(E_G(z))))^2] + \frac{1}{2}\mathbb{E}[(D(y) - \mathbf{1})^2] \tag{7}$$

4. Applications

We now describe in more detail about three typical and important clinical applications.

Normal versus abnormal classification. It is a binary classification problem. For this, a global average pooling layer, a fully connected layer with a single neuron and a sigmoid layer are added after the encoder E_F. The binary cross entropy loss is applied to compute the errors between the ground-truth CXR labels and the predictions. The added layers are jointly trained with the DGM framework as a multi-task learning manner.

Lung opacity versus no lung opacity classification. The generated residue maps of lung opacity related diseases should be different from the ones of other thoracic diseases (*e.g.*, cardiomegaly, effusion, etc.). We use a state-of-the-art classification model (DenseNet161 (Huang et al., 2017)) with slight modifications for lung opacity classification by concatenating the original CXR and its corresponding generated residue map as input and making the last fully connected layer have two output neurons for binary classification. The intuition behind these modifications is that the generated residue maps can provide some complementary and useful information to improve the performance.

Lung opacity detection.

Similar to lung opacity classification, we concatenate the original CXR and its corresponding generated residue map as input, then feed them into a state-of-the-art detection framework (RetinaNet (Lin et al., 2017)) for lung opacity detection.

5. Experiments

5.1. Implementation details

To optimize our defined objective functions, we alternately optimize the training of the discriminator and other parts (*i.e.*, encoders and decoders) until convergence. We use the Adam optimizer (Kingma and Ba, 2014) with exponential decay rates $(\beta_1, \beta_2) = (0.5, 0.999)$ and batch size of 1. The initial learning rate is 10^{-5} and the weight decay is 10^{-4}. We decay the learning rate by 0.5 every 10 epochs. The hyper-parameters are: $\lambda_A = 10$, $\lambda_R = 10$ and $\lambda_{TV} = 1$, which are determined empirically. DGM is trained from scratch for 30 epochs. The model is implemented using Pytorch on an NVIDIA DGX Station. The detailed structure information about the DGM' encoders and decoders is given in Tables 1 and 2. Please refer to the reference (Wang et al., 2018a) for details about the multi-scale discriminators.

5.2. Datasets and evaluation criteria

We evaluated DGM on a subset of the NIH Clinical Center Chest X-ray dataset[1] (Wang et al., 2017), which is the official dataset of the RSNA Kaggle pneumonia detection challenge[2]. Please refer to the reference (Shih et al., 2019) for more details about this subset (e.g. the annotation process). Cardiothoracic and pulmonary abnormalities include cardiomegaly, lung infiltrate, mass, nodule, pneumonia, pneumothorax, pulmonary atelectasis, consolidation, edema, emphysema, fibrosis, hernia, pleural effusion and thickening. The subset is composed of 26,684 frontal-view CXR images which were labeled into three categories: 1) normal (8851), 2) lung opacity (6012) (including pneumonia, infiltrate and consolidation) and 3) no lung opacity but not normal (11,821) (excluding normal, pneumonia, infiltrate and consolidation) (Shih et al., 2019). A hold-out set of 1,000 images were used for testing, the rest 25,684 CXRs were used for training and validation. We repeated the training, validation and testing process by randomly splitting training (95%) and validation (5%) sets at patient-level five times, and then calculated the means and standard deviations. All the lung opacity CXRs were annotated with rectangular bounding boxes indicating the locations of lung opacities. There were 1.6 boxes on average labeled on each lung opacity image.

For normal versus abnormal (CXRs other than normal) and lung opacity versus no lung opacity (normal + no lung opacity but not normal) CXR classification, the performance was evaluated using the area under the receiver operating characteristic (AUC) score, sensitivity (sen.), specificity (spe.), precision (pre.), F1-score (F1) and average precision (AP). The thresholds for binary classification

[1] https://nihcc.app.box.com/v/ChestXray-NIHCC

[2] https://www.kaggle.com/c/rsna-pneumonia-detection-challenge

Table 1
The detailed structure of encoders. IN means instance normalization used in E_G, while none means there is no normalization layer following a convolutional layer in E_F. The encoder E_G and E_F have identical structure except E_G with Instance Normalization (IN) layers while E_F without any normalization layer.

layer	input size	output size	filter size
Conv, IN/None, ReLU	$224 \times 224 \times 1$	$112 \times 112 \times 64$	7×7, 64, 2
Conv, IN/None, ReLU	$112 \times 112 \times 64$	$56 \times 56 \times 64$	4×4, 64, 2
Conv, IN/None, ReLU	$56 \times 56 \times 64$	$28 \times 28 \times 128$	4×4, 128, 2
Conv, IN/None, ReLU	$28 \times 28 \times 128$	$14 \times 14 \times 256$	4×4, 256, 2
Conv, IN/None, ReLU	$14 \times 14 \times 256$	$7 \times 7 \times 512$	4×4, 512, 2
$\begin{bmatrix}\text{Conv, IN/None, ReLU}\\ \text{Conv, IN/None}\end{bmatrix} \times 4$	$7 \times 7 \times 512$	$7 \times 7 \times 512$	$\begin{bmatrix}3 \times 3, 512, 1\\ 3 \times 3, 512, 1\end{bmatrix} \times 4$

Table 2
The detailed structure of the decoders. LN means layer normalization, while None means there is no normalization layer following a convolutional layer. The decoder D_G, D_F and D_J also have identical structure except different input sizes. For D_G and D_F, their input sizes are the same as the output sizes of their corresponding E_G and E_F, which are $7 \times 7 \times 512$. For D_J, the input size is $7 \times 7 \times 1024$.

layer	input size	output size	filter size
$\begin{bmatrix}\text{Conv, LN, ReLU}\\ \text{Conv, LN}\end{bmatrix} \times 4$	$7 \times 7 \times 512(1024)$	$7 \times 7 \times 512$	$\begin{bmatrix}3 \times 3, 512, 1\\ 3 \times 3, 512, 1\end{bmatrix} \times 4$
Upsample, Conv, LN, ReLU	$7 \times 7 \times 512$	$14 \times 14 \times 512$	5×5, 512, 1
Upsample, Conv, LN, ReLU	$14 \times 14 \times 512$	$28 \times 28 \times 256$	5×5, 256, 1
Upsample, Conv, LN, ReLU	$28 \times 28 \times 256$	$56 \times 56 \times 128$	5×5, 128, 1
Upsample, Conv, LN, ReLU	$56 \times 56 \times 128$	$112 \times 112 \times 64$	5×5, 64, 1
Upsample, Conv, Tanh	$112 \times 112 \times 64$	$224 \times 224 \times 1$	7×7, 1, 1

tasks are selected to minimize |sen. − spe.| on the validation set. For lung opacity detection, we evaluate the mean average precision (AP) using a series of intersection-over-union (IoU) threshold values range from 0.4 to 0.75 with a step size of 0.05.

5.3. Existing methods for comparisons

There is no previous work that can explicitly decompose an abnormal CXR into both a normal CXR and a residue map as done by DGM. Some existing state-of-the-art image-to-image translation methods (*e.g.*, CycleGAN (Zhu et al., 2017), MUNIT (Huang et al., 2018) and DRIT (Lee et al., 2018)) can only be used to explicitly obtain the normal CXR, but we can get the residue map by calculating the differences between the input abnormal CXR and the generated normal CXR. VA-GAN (Baumgartner et al., 2018) can only explicitly get the residue map and implicitly get the normal CXR by summing the input abnormal CXR and the generated residue map. Therefore, for CXR decomposition, we qualitatively and quantitatively compare the proposed method DGM with VA-GAN, CycleGAN, MUNIT and DRIT in this work. They are trained using the same data as DGM and their default hyper-parameters.

Additionally, to investigate the performance of DGM for normal versus abnormal CXR classification, we also train various well-known CNN classification models such as AlexNet (Krizhevsky et al., 2012), VGGNet19 (Simonyan and Zisserman, 2015), ResNet152 (He et al., 2016) and DenseNet161 (Huang et al., 2017), using the same training and validation sets as DGM.

5.4. Visual ablation study

To investigate the effect of the decoder D_J and the total variation loss ℓ_{TV}, we remove the decoder D_J from DGM (denoted as DGM w/o D_J) and train it without the self-reconstruction loss ℓ_R^2. We also train DGM without the total variation loss ℓ_{TV} (denoted as DGM w/o ℓ_{TV}). Fig. 3 shows two visual examples of the generated results produced by different models. From Fig. 3, we can see that 1) compared to DGM w/o D_J (Fig. 3(c)), the full model DGM with D_J can generate more radiorealistic normal CXRs having more

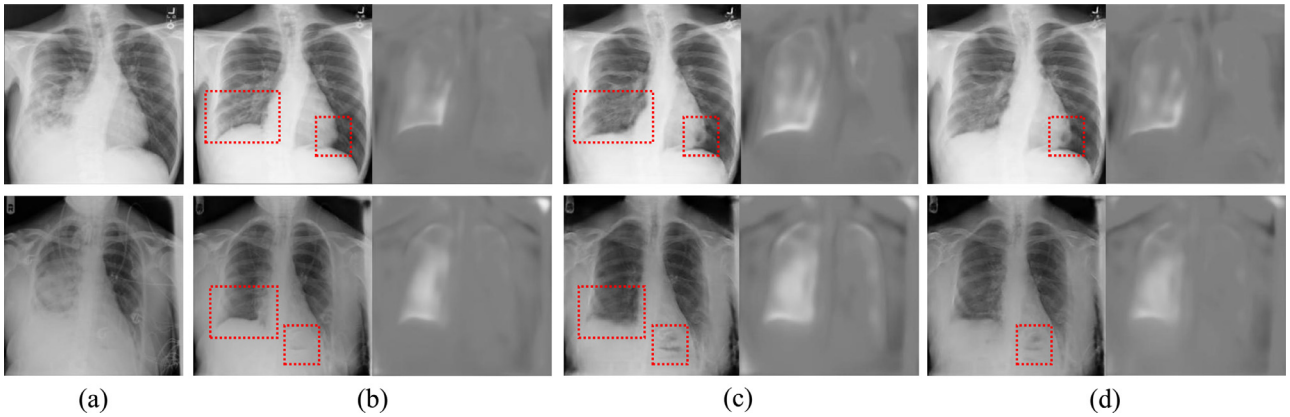

Fig. 3. Two examples of the generated normal CXRs (left) and disease residue maps (right) obtained from the input abnormal CXRs (a) using different models including (b) DGM, (c) DGM w/o D_J, and (d) DGM w/o ℓ_{TV}. The regions indicated by the red boxes show the biggest differences.

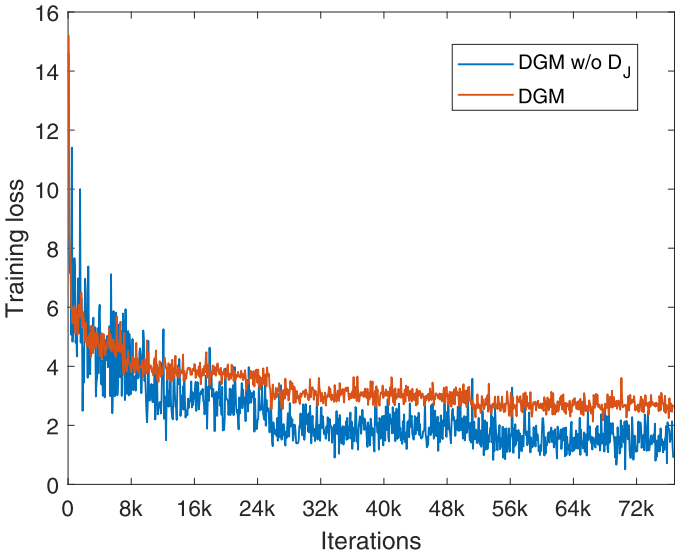

Fig. 4. Training loss of DGM and DGM w/o D_J.

similar structures to the input CXRs (Fig. 3(a)), so as to produce better residue maps to indicate the abnormalities. 2) Compared to DGM w/o ℓ_{TV} (Fig.e 3(d)), the full model DGM using ℓ_{TV} can generate normal CXRs with better quality and fewer artifacts (Fig. 3(a)). Fig. 4 shows the training loss of DGM with or without D_J. From Fig. 4, the training loss of the model with D_J is more stable and converges faster than the one of the model without D_J. All of these results demonstrate the effectiveness of the decoder D_J and the total variation loss ℓ_{TV}.

5.5. Qualitative comparisons of the disentangled disease residue maps and the generated normal CXRs

Fig. 5 shows four example results of the disentangled disease residue map and Fig. 6 shows four examples of the generated normal CXRs using different methods. The input abnormal CXRs in each row of Figs. 5(a) and 6(a) are the same.

From Fig. 5, the disease residue maps produced by our method are more meaningful and interepretable than other competing methods. Comparing the highlighted abnormal regions in the residue maps and the abnormal regions annotated by radiologist, our method shows most consistent abnormal regions compared to ground truth. Fig. 6 also shows that the normal CXRs generated by our method are more visually "radiorealistic" since it keeps more original information (*e.g.*, the intensity, the heart shape and position, the lung outline and the other soft tissues) than all other methods.

5.6. Human perceptual study for the generated images' quality evaluation

To quantitatively show whether the synthetic CXR images from different methods are radiorealistic, we perform human perceptual study and use the human preference to measure the image quality. We randomly select 100 abnormal CXRs from test set. For each abnormal CXR, we randomly combine two generated normal CXRs (or residue maps) from five results produced by different methods and ask a board-certificated radiologist to select which one looks more accurate. Therefore, we collect 1000 comparisons for generated normal CXR and residue map assessments, respectively. The percentage a method is preferred can be calculated as human preference score (HPS) for performance evaluation and ranking. As

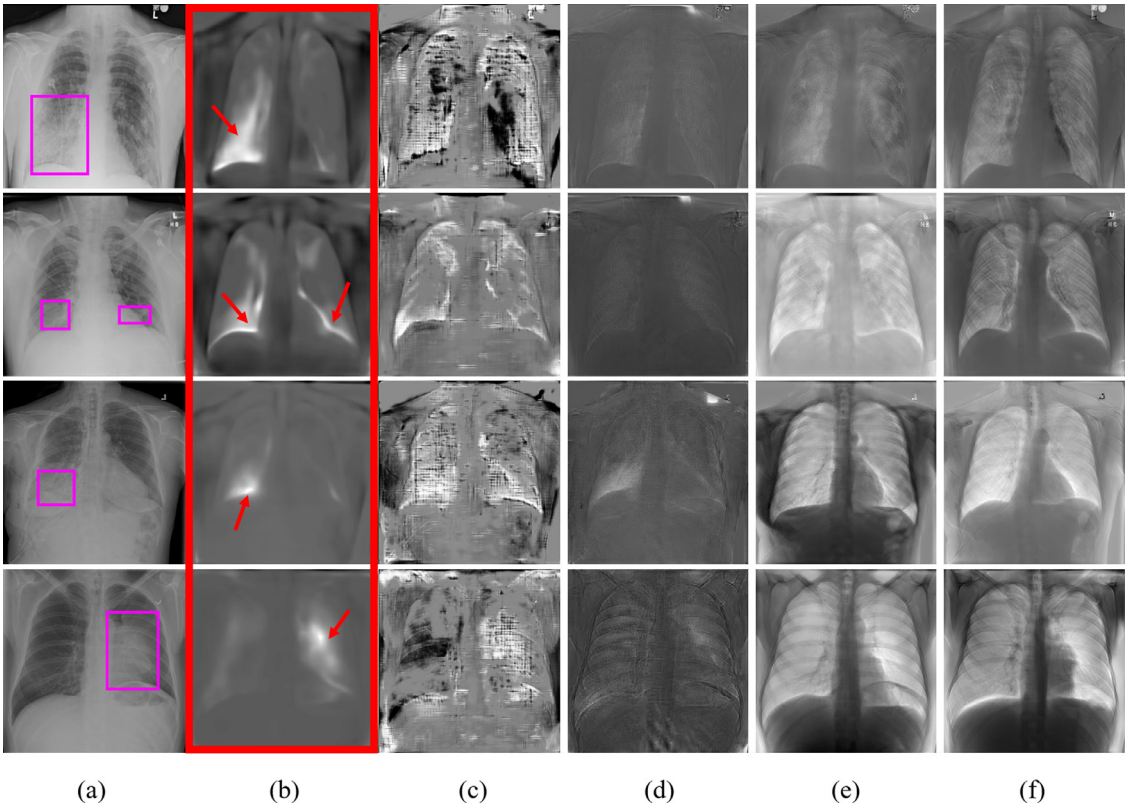

Fig. 5. Four examples of the disentangled disease residue maps obtained from the input abnormal CXRs (a) using different methods, including (b) DGM (red box), (c) VA-GAN, (d) CycleGAN, (e) MUNIT, (f) DRIT. The pink boxes in (a) indicate the abnormal regions annotated by radiologists. The red arrow shows the abnormal region highlighted by DGM. It is consistent with the radiologist's annotations (Better viewed in color).

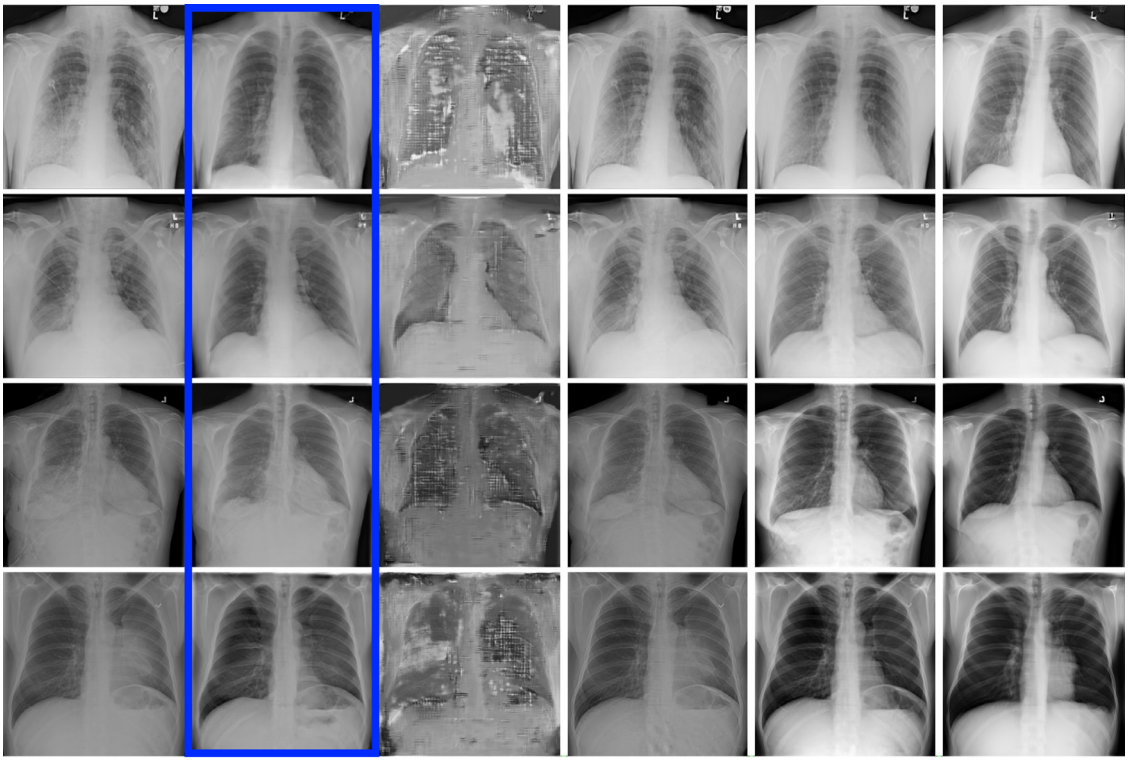

Fig. 6. Four examples of the synthesized normal CXRs obtained from the input abnormal CXRs (a) using different methods, including (b) DGM (blue box), (c) VA-GAN, (d) CycleGAN, (e) MUNIT, (f) DRIT. Here, the input abnormal CXRs are the same as the ones in Fig. 5(a).

Table 3
Human preference scores (HPSs) of different methods on generated normal CXR and residue map study.

Study	VA-GAN	CycleGAN	MUNIT	DRIT	DGM
Normal CXR	2.5%	12.5%	23.4%	27.1%	**34.5%**
Residue map	4.6%	9.2%	24.5%	26.3%	**35.4%**

shown in Table 3, DGM is significantly better than all other compared methods. This result suggests that our generated CXR images are more radiorealistic.

5.7. CXR classification performance evaluation with or without using the generated disease residue maps

When DGM is modified for normal and abnormal CXR classification, only the encoder E_F is required at the inference stage. To verify the discriminability of our generated residue maps, a DenseNet161 model is trained using the residue maps as the inputs for normal and abnormal CXR classification (denoted as DenseNet161*). Table 4 lists the results of different methods for normal and abnormal CXR classification in terms of inference model size, sensitivity, specificity, precision, F1, AP and AUC. From Table 4, we can see that: 1) DenseNet161 gets the best performance compared with the other models. 2) Although VGGNet19 has a much shallower structure compared to ResNet152, it gets better results. 3) DGM has the smallest inference model size, but still achieves comparable results with ResNet152 and better results than AlexNet, whose inference model sizes are much larger. 4) DenseNet161* achieves promising classification performance, suggesting that the residue maps contain some discriminative information to distinguish normal and abnormal CXRs.

Following the same processing as lung opacity classification, we also concatenate the original CXR and its corresponding generated residue maps produced by DGM, VA-GAN, CycleGAN, MUNIT and DRIT as input of DenseNet161 for normal and abnormal classification. Table 5 lists the results of DenseNet161 models using different input configurations for normal and abnormal CXR classification in terms of sensitivity, specificity, precision, F1, AP and AUC. From Table 5, we can see that DenseNet161 + DGM obtains better performance than DenseNet161. When using the residue maps produced by other methods (*i.e.*, VA-GAN, CycleGAN, MUNIT and DRIT), the classification performance becomes even worse.

Table 6 lists the results of DenseNet161 models using different input configurations for lung opacity classification in terms of different evaluation metrics. We can get similar conclusions as normal and abnormal classification. All results in Tables 5 and 6 implicitly demonstrate the effectiveness of DGM for CXR decomposition and explicitly demonstrate its effectiveness for CXR classification.

5.8. Lung opacity detection performance evaluation with or without using the generated disease residue maps

Based on the residue maps produced by DGM, it is straightforward to use them for lung opacity detection. The residue maps highlight the potential disease regions but does not preserve the contextual information, which is important for lung opacity related diseases detection. These contextual information can be obtained from the original CXRs. Therefore, we use the residue map along with original CXRs in order to extract more contextual information.

For comparison, we train the RetinaNet model with different inputs as done in CXR classification. The following configurations are used for training: 1) ResNet50 (He et al., 2016) is used as the backbone of RetinaNet, 2) SGD is used with learning rate of 0.01, mo-

Table 4
Size of inference model (MB), sensitivity (Sen.), specificity (Spe.), precision (Pre.), F1, AP and AUC results (%) of different methods for normal and abnormal CXR classification.

Method	Size	Sen.	Spe.	Pre.	F1	AP	AUC
AlexNet	57.0	87.12 ± 0.96	87.09 ± 0.92	76.55 ± 1.04	81.49 ± 0.76	87.36 ± 0.59	94.39 ± 0.37
VGGNet19	139.6	90.18 ± 0.71	90.21 ± 0.69	81.67 ± 0.64	85.71 ± 0.55	90.90 ± 0.29	96.11 ± 0.14
ResNet152	58.1	89.88 ± 0.73	89.91 ± 0.71	81.16 ± 0.67	85.30 ± 0.60	90.21 ± 0.35	95.95 ± 0.17
DenseNet161	26.5	**92.02 ± 0.51**	**91.99 ± 0.55**	**84.75 ± 0.47**	**88.24 ± 0.44**	**90.29 ± 0.34**	**96.43 ± 0.12**
DenseNet161*	26.5	88.48 ± 0.89	88.62 ± 0.85	79.63 ± 0.92	83.24 ± 0.68	88.58 ± 0.47	94.51 ± 0.35
DGM	**7.5**	90.18 ± 0.69	90.36 ± 0.67	81.89 ± 0.61	85.48 ± 0.57	89.92 ± 0.37	95.83 ± 0.19

Table 5
Sensitivity (Sen.), specificity (Spe.), precision (Pre.), F1, AP and AUC results (%) of DenseNet161 models using different input configurations for normal and abnormal CXR classification.

Method	Sen.	Spe.	Pre.	F1	AP	AUC
DenseNet161	92.02 ± 0.51	91.99 ± 0.55	84.75 ± 0.47	88.24 ± 0.44	90.29 ± 0.34	96.43 ± 0.12
+ VA-GAN	88.64 ± 0.89	88.59 ± 0.91	80.12 ± 0.83	84.18 ± 0.71	88.42 ± 0.57	94.31 ± 0.48
+ CycleGAN	89.35 ± 0.81	89.39 ± 0.80	80.47 ± 0.78	84.71 ± 0.67	89.75 ± 0.43	95.36 ± 0.30
+ MUNIT	89.79 ± 0.77	89.84 ± 0.75	81.43 ± 0.69	85.41 ± 0.61	89.61 ± 0.45	95.54 ± 0.26
+ DRIT	90.01 ± 0.72	90.09 ± 0.70	81.61 ± 0.66	85.44 ± 0.59	89.77 ± 0.41	95.68 ± 0.21
+ **DGM**	**92.08 ± 0.47**	**92.13 ± 0.45**	**84.85 ± 0.43**	**88.41 ± 0.42**	**92.02 ± 0.25**	**96.78 ± 0.10**

Table 6
Sensitivity (Sen.), specificity (Spe.), precision (Pre.), F1, AP and AUC results (%) of DenseNet161 models using different input configurations for lung opacity classification.

Method	Sen.	Spe.	Pre.	F1	AP	AUC
DenseNet161	82.15 ± 0.59	82.23 ± 0.52	71.60 ± 0.85	76.52 ± 0.78	85.04 ± 0.48	90.98 ± 0.25
+ VA-GAN	81.59 ± 0.71	81.61 ± 0.69	70.76 ± 0.97	75.79 ± 0.94	84.18 ± 0.60	90.00 ± 0.34
+ CycleGAN	81.87 ± 0.67	81.92 ± 0.66	71.18 ± 0.92	76.15 ± 0.90	84.52 ± 0.57	90.38 ± 0.32
+ MUNIT	81.92 ± 0.68	81.96 ± 0.65	71.24 ± 0.90	76.35 ± 0.89	84.86 ± 0.53	90.65 ± 0.28
+ DRIT	82.00 ± 0.63	82.05 ± 0.58	71.36 ± 0.88	76.41 ± 0.82	84.93 ± 0.52	90.74 ± 0.27
+ **DGM**	**84.70 ± 0.49**	**84.70 ± 0.50**	**75.13 ± 0.69**	**79.63 ± 0.61**	**86.11 ± 0.39**	**91.97 ± 0.19**

Table 7
Comparison of lung opacity detection results of RetinaNet models with different input configurations in terms of average precision (AP, in %) of at different IoUs. AP: mean average precision @ IoU = [0.4:0.05:0.75], AP_{40}: average precision @ IoU = 0.4, and so forth.

Method	AP	AP_{40}	AP_{50}	AP_{60}	AP_{70}
RetinaNet	19.09 ± 0.24	41.42 ± 0.19	14.22 ± 0.28	12.76 ± 0.33	10.72 ± 0.32
+ VA-GAN	18.68 ± 0.30	41.18 ± 0.21	13.17 ± 0.34	12.39 ± 0.39	10.48 ± 0.36
+ CycleGAN	18.95 ± 0.28	41.61 ± 0.18	13.96 ± 0.30	12.59 ± 0.37	10.61 ± 0.35
+ MUNIT	19.34 ± 0.21	41.91 ± 0.16	14.78 ± 0.26	12.95 ± 0.30	10.89 ± 0.31
+ DRIT	19.81 ± 0.19	42.03 ± 0.15	15.48 ± 0.23	13.44 ± 0.27	11.21 ± 0.29
+ **DGM**	**23.95 ± 0.15**	**44.65 ± 0.11**	**24.25 ± 0.14**	**16.64 ± 0.21**	**12.54 ± 0.25**

mentum of 0.9 and weight decay of 0.0001, 3) we train 25 epochs with 2500 steps per epoch, batch size of 8 and input CXR resolution of 224 × 224. After training, we select the models show the best performance on the validation set and apply it on the test set for evaluation.

Table 7 reports the lung opacity detection results of RetinaNet models with different input configurations on test set. As shown in Table 7, we can see that using the residue map produced by DGM obtains noticeable performance gains, especially at AP_{50}. For VA-GAN and CycleGAN, the detection performance drops. For MUNIT and DRIT, the performance is improved slightly but is still much worse than ours. These results demonstrate that the DGM residue maps are helpful for lung opacity detection. Overall, all results in Tables 5–7 indirectly indicate the effectiveness of DGM for CXR decomposition.

Fig. 7 shows five visual lung opacity detection results. As we can see that when using the generated residue maps of DGM, 1) the model is forced to produce tighter predictions. The predicted bounding boxes are closer to the ground truths, as shown in Fig. 7(a). 2) the model can give more correct predictions by reducing false negatives shown in Fig. 7(b)&(c) and false positives shown in Fig. 7(d). The results in Fig. 7(a)-(d) qualitatively verify the usefulness of the residue maps generated by DGM for lung opacity detection. But there are some difficult scenarios where the models may fail, *e.g.*, when the lung opacity severity is mild (Fig. 7(e)). Therefore, it is still very challenging to accurately detect the lung opacities in CXRs.

6. Discussion

The decoder D_J is an important component of the proposed framework DGM according to Figs. 3 and 4. That is because an extra self-reconstruction loss is directly introduced when using the decoder D_J, which can be considered as a strong supervision to guide the model optimization. As a result, the encoder E_G and E_F could be learned better and the model can converge more stably and faster during training. Therefore, using D_J to reconstruct the input CXR z from c_z and c_s can facilitate the learning process. Because of the spatial smoothness property of the total variation loss, it can guide the model to generate normal CXRs with fewer noises or artifacts.

For CXR decomposition, DGM achieves better qualitative and quantitative results than the other methods. The reason is that besides generating population-based normal CXRs with the adversar-

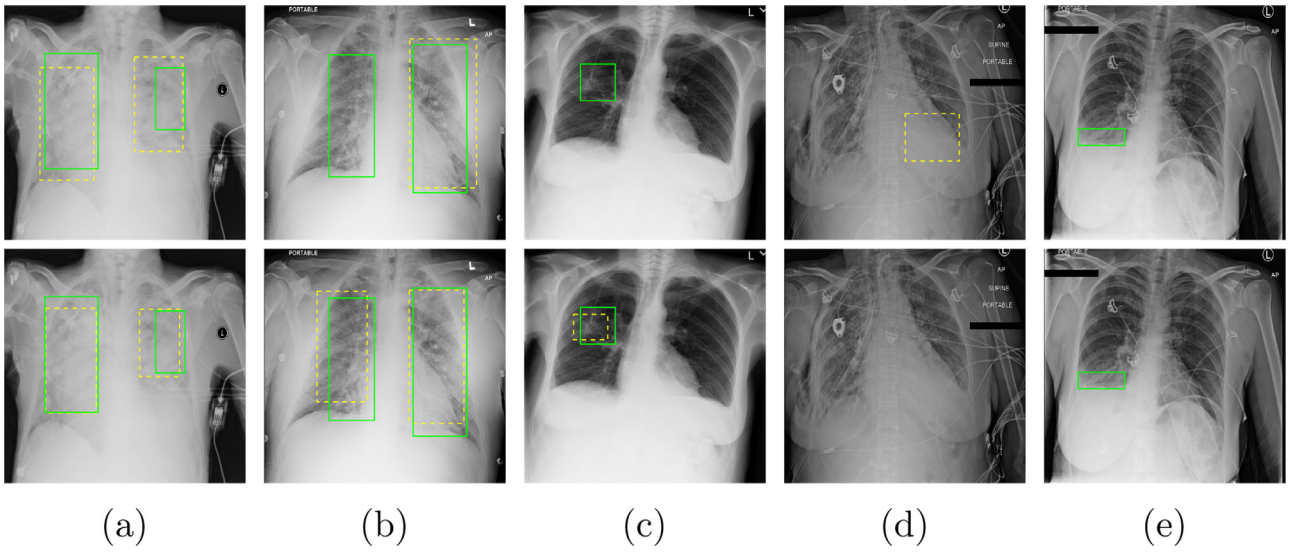

Fig. 7. Five visual examples of lung opacity detection results. The first row shows the results produced only using the original CXRs, while the second row shows the ones produced using extra residue maps generated by DGM. The green solid boxes represent the ground truths, and the yellow dashed boxes represent the prediction results.

ial loss like others (e.g. CycleGAN, MUNIT and DRIT), DGM introduces three individual-based self-reconstruction losses to guide the generator to preserve person-specific structures in the generated normal CXRs. Therefore, the normal CXRs generated by DGM not only are realistic but also have similar structures as the original abnormal CXRs, so as to produce more interpretable residue maps. Actually, we also investigate KL divergence loss and latent regression loss that encourages invertible mapping between the CXR and the latent feature space. Through experiments, we didn't find that they have provided any visual improvement. The most important difference from the other methods is DGM can output an abnormal residue map simultaneously and explicitly as an auxiliary of the generated normal CXR. This gives a comprehensive interpretation of the input abnormal CXR diagnosis, such as locations of the potential abnormal regions and the severity levels of the abnormalities. From Figs. 5 and 6, the results obtained by VA-GAN are much poorer than the others. The possible reasons are that 1) The network architecture of the generator used in VA-GAN is a compressed U-Net to reduce GPU memory consumption, which is smaller than the models used in other works and ours. It may not be powerful enough on CXRs. 2) VA-GAN uses a single scale discriminator while the others and ours use multi-scale discriminators that can guide the generators to preserve both local detail structures and correct global structure.

For CXR classification, the deepest CNN architecture DenseNet161 gets the best performance, suggesting that the model with deeper structure and dense connections is more powerful to extract discriminative features for classification. This finding matches the common concept that deeper networks tend to work better for image classification (He et al., 2016; Huang et al., 2017). But ResNet152 even performs worse than VGGNet19 shown in Table 4. The possible reason is that VGGNet19 has more parameters, which pushes it to learn more discriminative features for this simple binary classification task. DGM achieves promising results, suggesting that a shallower and smaller model may show powerful ability to learn discriminative features in a sophisticated framework, such as the multi-task (including CXR binary classification and decomposition) learning framework for normal and abnormal CXR classification application in this work. From Table 5 and Table 6, the classification performance improvement is not significant when using our generated residue maps, but they do help. Therefore, the residue maps could provide some useful and supplementary information for CXR classification. They can be considered as a different way of extracting features from the input CXRs that somewhat implicitly enhances the capacity of the classification network.

For lung opacity detection, better residue maps should provide more helpful information. The highlighted areas in VA-GAN and CycleGAN's residue maps cannot well represent the abnormal regions, resulting in harming the detection performance. For MUNIT and DRIT, their highlighted regions in the residue maps can somehow indicate the abnormalities while ours can model the underlying abnormal regions in a better way.

When integrating the generated residue maps produced by DGM for all applications, their performances are boosted compared with the ones only using original CXRs, meaning that our generated residue maps can provide some complementary and helpful information compared to the original CXRs and can well indicate the underlying abnormal regions. From Tables 5–7, when using the residue maps from MUNIT and DRIT, the classification performance becomes worse while the detection performance becomes better. One possible reason is that their residue maps highlight background regions sometimes, which may reduce the classification accuracy. As we know, the classification model considers the global contextual information including background, while the detection model focuses on local structures.

Here, we have demonstrated the effectiveness of the proposed method for CXR decomposition that can well interpret CXRs by outputting meaningful disease residue or saliency maps. Besides interpretation, our model can be used for disease detection, disease diagnosis, and data augmentation when limited normal CXR images available. This work shows that the proposed deep learning model DGM offers the potential for radiologists to use the generated disease residue/saliency maps to improve the CXR reading efficiency in clinical practice, thus improving radiology workflow and patient care.

7. Conclusions and future work

In this paper, we investigated how the abnormal disease patterns can be disentangled from abnormal CXRs. We proposed a simple, yet effective model for "radiorealistic" normal chest X-ray synthesis and normal-disease separation. The proposed model is end-to-end trainable using unpaired images. Experiments showed

that the proposed model was capable of providing an interpretation of CXR images by generating a potential disease saliency map. In addition, it quantitatively improved the binary disease classification and detection performance. Future work will explore nonlinear decomposition in normal-disease separation, multi-class and multi-label disentanglement to achieve more accurate disease classification and detection.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ronald M. Summers receives royalties from PingAn, iCAD, ScanMed and Philips, research support from PingAn, and GPU card donations from NVIDIA.

CRediT authorship contribution statement

Youbao Tang: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft, Visualization. **Yuxing Tang:** Formal analysis, Data curation, Writing - original draft. **Yingying Zhu:** Writing - review & editing. **Jing Xiao:** Writing - review & editing. **Ronald M. Summers:** Data curation, Writing - review & editing, Supervision.

Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health Clinical Center and by the Ping An Insurance Company through a Cooperative Research and Development Agreement. We thank Nvidia for GPU card donation.

References

Bao, J., Chen, D., Wen, F., Li, H., Hua, G., 2018. Towards open-set identity preserving face synthesis. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6713–6722.

Baumgartner, C.F., Koch, L.M., Can Tezcan, K., Xi Ang, J., Konukoglu, E., 2018. Visual feature attribution using wasserstein gans. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 8309–8319.

Bowles, C., Qin, C., Ledig, C., Guerrero, R., Gunn, R., Hammers, A., Sakka, E., Dickie, D.A., Hernández, M.V., Royle, N., et al., 2016. Pseudo-healthy image synthesis for white matter lesion segmentation. In: International Workshop on Simulation and Synthesis in Medical Imaging, pp. 87–96.

Cai, J., Lu, L., Harrison, A.P., Shi, X., Chen, P., Yang, L., 2018. Iterative attention mining for weakly supervised thoracic disease pattern localization in chest x-rays. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 589–598.

Chang, H., Lu, J., Yu, F., Finkelstein, A., 2018. Pairedcyclegan: asymmetric style transfer for applying and removing makeup. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 40–48.

Chartsias, A., Joyce, T., Papanastasiou, G., Semple, S., Williams, M., Newby, D.E., Dharmakumar, R., Tsaftaris, S.A., 2019. Disentangled representation learning in cardiac image analysis. Med. Image Anal. 58, 101535.

Chen, C., Dou, Q., Jin, Y., Chen, H., Qin, J., Heng, P.-A., 2019. Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 447–456.

Chen, H., Miao, S., Xu, D., Hager, G.D., Harrison, A.P., 2019. Deep hierarchical multi--label classification of chest x-ray images. In: International Conference on Medical Imaging with Deep Learning, pp. 109–120.

Dunnmon, J.A., Yi, D., Langlotz, C.P., Ré, C., Rubin, D.L., Lungren, M.P., 2019. Assessment of convolutional neural networks for automated classification of chest radiographs. Radiology 290 (2), 537–544.

Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K., Stanley, H.E., 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation 101 (23), 215–220.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Advances in neural information processing systems, pp. 2672–2680.

Guan, Q., Huang, Y., 2020. Multi-label chest x-ray image classification via category–wise residual attention learning. Pattern Recognition Letters 130, 259–266.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Horváth, G., Orbán, G., Horváth, Á., Simkó, G., Pataki, B., Máday, P., Juhász, S., 2009. A cad system for screening x-ray chest radiography. In: World Congress on Medical Physics and Biomedical Engineering, pp. 210–213.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708.

Huang, X., Belongie, S., 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In: International Conference on Computer Vision, pp. 1501–1510.

Huang, X., Liu, M.-Y., Belongie, S., Kautz, J., 2018. Multimodal unsupervised image–to-image translation. In: European Conference on Computer Vision, pp. 172–189.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134.

Jin, D., Xu, Z., Tang, Y., Harrison, A.P., Mollura, D.J., 2018. Ct-realistic lung nodule simulation from 3d conditional generative adversarial networks for robust lung segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 732–740.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision, pp. 694–711.

Khan, A.N., Al-Jahdali, H., Al-Ghanem, S., Gouda, A., 2009. Reading chest radiographs in the critically ill (part i): Normal chest radiographic appearance, instrumentation and complications from instrumentation. Ann. Thoracic Med. 4 (2), 75–87.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization arXiv:1412.6980.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al., 2017. Photo-realistic single image super-resolution using a generative adversarial network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4681–4690.

Lee, H.-Y., Tseng, H.-Y., Huang, J.-B., Singh, M., Yang, M.-H., 2018. Diverse image–to-image translation via disentangled representations. In: European Conference on Computer Vision, pp. 35–51.

Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L.-J., Fei-Fei, L., 2018. Thoracic disease identification and localization with limited supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8290–8299.

Liang, J., Tang, Y.-X., Tang, Y.-B., Xiao, J., Summers, R.M., 2020. Bone suppression on chest radiographs with adversarial learning. In: Medical Imaging: Computer-Aided Diagnosis, 11314, p. 1131409.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: IEEE International Conference on Computer Vision, pp. 2980–2988.

Liu, M.-Y., Breuel, T., Kautz, J., 2017. Unsupervised image-to-image translation networks. In: Advances in Neural Information Processing Systems, pp. 700–708.

Liu, X., Niethammer, M., Kwitt, R., McCormick, M., Aylward, S., 2014. Low-rank to the rescue–atlas-based analyses in the presence of pathologies. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 97–104.

Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S., 2017. Least squares generative adversarial networks. In: IEEE International Conference on Computer Vision, pp. 2794–2802.

Qin, C., Shi, B., Liao, R., Mansi, T., Rueckert, D., Kamen, A., 2019. Unsupervised deformable registration for multi-modal images via disentangled representations. In: International Conference on Information Processing in Medical Imaging, pp. 249–261.

Qin, C., Yao, D., Shi, Y., Song, Z., 2018. Computer-aided detection in chest radiography based on artificial intelligence: a survey. Biomed. Eng. Online 17 (1), 113.

Radford, A., Metz, L., Chintala, S., 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. In: International Conference on Learning Representations.

Rajpurkar, P., Irvin, J., Ball, R.L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C.P., et al., 2018. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. PLOS Medicine 15 (11), 1–17.

Rajpurkar, P., Irvin, J., Zhu, K., Brandon Yang, H.M., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M.P., Ng, A.Y., 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning arXiv:1711.05225.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H., 2016. Generative adversarial text to image synthesis. In: International Conference on Machine Learning, pp. 1060–1069.

Salehinejad, H., Colak, E., Dowdell, T., Barfett, J., Valaee, S., 2018. Synthesizing chest x-ray pathology for training deep convolutional neural networks. IEEE Transactions on Medical Imaging 38 (5), 1197–1206.

Shih, G., Wu, C.C., Halabi, S.S., Kohli, M.D., Prevedello, L.M., Cook, T.S., Sharma, A., Amorosa, J.K., Arteaga, V., Galperin-Aizenberg, M., et al., 2019. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiology: Artificial Intelligence 1 (1), e180041.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations.

Smith-Bindman, R., Miglioretti, D.L., Larson, E.B., 2008. Rising use of diagnostic medical imaging in a large integrated health system. Health affairs 27 (6), 1491–1502.

Tang, Y., Cai, J., Lu, L., Harrison, A.P., Yan, K., Xiao, J., Yang, L., Summers, R.M., 2018. Ct image enhancement using stacked generative adversarial networks and transfer

learning for lesion segmentation improvement. In: International Workshop on Machine Learning in Medical Imaging, pp. 46–54.
Tang, Y., Tang, Y., Sandfort, V., Xiao, J., Summers, R.M., 2019. Tuna-net: Task-oriented unsupervised adversarial network for disease recognition in cross-domain chest x-rays. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 431–440.
Tang, Y., Tang, Y., Xiao, J., Summers, R.M., 2019. Xlsor: A robust and accurate lung segmentor on chest x-rays using criss-cross attention and customized radiorealistic abnormalities generation. In: International Conference on Medical Imaging with Deep Learning, pp. 457–467.
Tang, Y., Wang, X., Harrison, A.P., Lu, L., Xiao, J., Summers, R.M., 2018. Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs. In: International Workshop on Machine Learning in Medical Imaging, pp. 249–258.
Tang, Y.-B., Oh, S., Tang, Y.-X., Xiao, J., Summers, R.M., 2019. Ct-realistic data augmentation using generative adversarial network for robust lymph node segmentation. In: Medical Imaging: Computer-Aided Diagnosis, 10950, p. 109503V.
Tang, Y.-X., Tang, Y.-B., Han, M., Xiao, J., Summers, R.M., 2019. Abnormal chest X-ray identification with generative adversarial one-class classifier. In: International Symposium on Biomedical Imaging, pp. 1358–1361.
Tang, Y.-X., Tang, Y.-B., Han, M., Xiao, J., Summers, R.M., 2019. Deep adversarial one–class learning for normal and abnormal chest radiograph classification. In: Medical Imaging: Computer-Aided Diagnosis, 10950, p. 1095018.
Tang, Y.-X., Tang, Y.-B., Peng, Y., Yan, K., Bagheri, M., Redd, B.A., Brandon, C.J., Lu, Z., Han, M., Xiao, J., et al., 2020. Automated abnormality classification of chest radiographs using deep convolutional neural networks. NPJ Digit. Med. 3 (1), 1–8.
Ulyanov, D., Vedaldi, A., Lempitsky, V., 2017. Improved texture networks: maximizing quality and diversity in feed-forward stylization and texture synthesis. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6924–6932.
Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., Catanzaro, B., 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 8798–8807.
Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017. Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2097–2106.
Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M., 2018. Tienet: text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 9049–9058.
Yasaka, K., Abe, O., 2018. Deep learning and artificial intelligence in radiology: Current applications and future directions. PLoS Med. 15 (11), e1002707.
Ye, D.H., Zikic, D., Glocker, B., Criminisi, A., Konukoglu, E., 2013. Modality propagation: coherent synthesis of subject-specific scans with data-driven regularization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 606–613.
Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S., 2018. Generative image inpainting with contextual attention. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5505–5514.
Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929.
Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision, pp. 2223–2232.
Zhu, Y., Tang, Y., Tang, Y., Elton, D.C., Lee, S., Pickhardt, P.J., Summers, R.M., 2020. Cross-domain medical image translation by shared latent gaussian mixture model arXiv:2007.07230.