Image-to-Images Translation for Multi-Task Organ Segmentation and Bone Suppression in Chest X-Ray Radiography

Mohammad Eslami, *Member, IEEE*, Solale Tabarestani, *Student Member, IEEE*, Shadi Albarqouni, *Member, IEEE*, Ehsan Adeli, *Member, IEEE*, Nassir Navab, *Member, IEEE*, and Malek Adjouadi

***Abstract*— Chest X-ray radiography is one of the earliest medical imaging technologies and remains one of the most widely-used for diagnosis, screening, and treatment follow up of diseases related to lungs and heart. The literature in this field of research reports many interesting studies dealing with the challenging tasks of bone suppression and organ segmentation but performed separately, limiting any learning that comes with the consolidation of parameters that could optimize both processes. This study, and for the first time, introduces a multitask deep learning model that generates simultaneously the bone-suppressed image and the organ-segmented image, enhancing the accuracy of tasks, minimizing the number of parameters needed by the model and optimizing the processing time, all by exploiting the interplay between the network parameters to benefit the performance of both tasks. The architectural design of this model, which relies on a conditional generative adversarial network, reveals the process on how the well-established *pix2pix* network (*image-to-image* network) is modified to fit the need for multitasking and extending it to the new *image-to-images* architecture. The developed source code of this multitask model is shared publicly on Github as the first attempt for providing the two-task *pix2pix* extension, a supervised/paired/aligned/registered *image-to-images* translation which would be useful in many multitask applications. Dilated convolutions are also used to improve the results through a more effective receptive field assessment. The comparison with state-of-the-art algorithms along with ablation study and a demonstration video[1] are provided to evaluate the efficacy and gauge the merits of the proposed approach.**

***Index Terms*— Bone suppression, chest X-Ray, CXR imaging, image-to-image translation, image-to-images translation, multitask deep learning, organ segmentation, pix2pix.**

I. Introduction

CHEST radiography, also called chest X-ray or CXR, is one of the most affordable and widely used medical imaging modality, which has significant practical implications in the diagnosis and screening of the thorax region, the organs and bone structure within it. Over 2 billion procedures per year are performed using this technology for the purpose of medical diagnosis of a variety of diseases, such as pneumonia, tuberculosis, lung cancer, and heart failure. Moreover, chest radiography remains the most prevalent screening test for pulmonary disorders [1]–[5]. However, due to overlapping organs, low resolution and subtle anatomical shape and size variations, interpreting CXR images accurately remains challenging and requires a well-trained staff. On the other hand, managing a large number of CXR images each day results in high workloads for the radiography staff, yielding a tedious process fraught with setbacks and errors in diagnosis and in planning for adequate treatment follow up. It is reported that almost 90 percent of mistakes in pulmonary tumor diagnosis could be associated with the CXR screening of images [6]. Therefore, many efforts have been devoted to the development of automated computer-based methods to improve accuracy in diagnosis and in finding any abnormalities that may otherwise be left undetected [7]–[10].

There is considerable literature focusing on CXR image analysis. Among the more recent work on chest radiography,

Manuscript received December 16, 2019; revised February 4, 2020; accepted February 8, 2020. Date of publication February 14, 2020; date of current version June 30, 2020. This work was supported by the National Science Foundation (NSF) under NSF Grant CNS-1920182, Grant CNS-1532061, and Grant CNS-1551221. The work of Mohammad Eslami was also supported by ESLA White Labs GmbH. The work of Shadi Albarqouni was supported by the PRIME Programme of the German Academic Exchange Service (DAAD) with funds from the German Federal Ministry of Education and Research (BMBF). *(Corresponding author: Mohammad Eslami.)*

Mohammad Eslami was with CAMP-TUM, 80333 Munich, Germany. He is now with the Center for Advanced Technology and Education (CATE), Florida International University, Miami, FL 33174 USA (e-mail: meslami@fiu.edu).

Solale Tabarestani and Malek Adjouadi are with the Center for Advanced Technology and Education (CATE), Florida International University, Miami, FL 33174 USA (e-mail: staba006@fiu.edu; adjouadi@fiu.edu).

Shadi Albarqouni is with the Computer Aided Medical Procedures and Augmented Reality (CAMP), Technical University of Munich (TUM), 80333 Munich, Germany, and also with the Computer Vision Lab (CVL), Department of Information Technology and Electrical Engineering, ETH Zürich, 8092 Zürich, Switzerland (e-mail: shadi.albarqouni@tum.de).

Ehsan Adeli is with the School of Medicine and Computer Science Department, Stanford University, Stanford, CA 94305 USA (e-mail: eadeli@stanford.edu).

Nassir Navab is with the Computer Aided Medical Procedures and Augmented Reality (CAMP), Technical University of Munich (TUM), 80333 Munich, Germany, and also with Computer Aided Medical Procedures (CAMP), Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: nassir.navab@tum.de).

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

Digital Object Identifier 10.1109/TMI.2020.2974159

[1] https://youtu.be/J8Uth26_7rQhttps://youtu.be/J8Uth26_7rQ

a team from Stanford [11] proposed a convolutional neural network called *CheXNeXt* as a deep learning algorithm to concurrently detect the presence of 14 different pathologies such as pneumonia, fibrosis, emphysema, and nodules in frontal-view chest radiographs, among others. The *CheXNeXt* algorithm achieved promising results in identifying abnormalities at a performance level that was comparable with the diagnostic accuracy of radiologist practitioners. Four different deep learning based methods are investigated in [12] and compared with radiology experts. In another study [6], Gozes and Greespan proposed a method to improve the contrast of lung structures in CXR images leading to better accuracy in nodule(s) detection. Wang and Chia proposed a deep neural network they named *ChestNet* [13] for enhanced diagnosis of diseases on chest radiography. Moreover, Li et al. developed an interesting multi-resolution convolutional network on chest X-ray radiograph for lung nodule detection [14].

The aim of this work is to construct a multitask learning framework using deep learning techniques that address in an effective way the two challenging tasks of organ segmentation and bone suppression simultaneously. Organ segmentation is used for computer-aided detection and diagnosis, while bone suppression enhances the visibility of the disease effects, e.g. nodules particularly on the lungs. Although similar architectures are reported to perform successfully each task separately, e.g. *U-net* [15], [16], addressing the two tasks of bone suppression and organ segmentation simultaneously seem to reinforce the filters of the network to look for and learn about the relevant features of both organs and bones. Multitask learning could hence be applied to those problems where two or more tasks are found to be inherently interrelated. Learning the relationship between tasks and their respective feature space could lead to a more unified feature learning process and hence better prospects for multitasking and generalization [17]. In our case, the two tasks are spatially correlated. Indeed, for the single task problem of bone suppression, the network should distinguish rib bones located over the chest organs. Similarly, for the single task problem of organ segmentation, the network should learn those features that can represent the chest organs as situated in context to the rib bones in that area.

In order to incorporate the multitask objective, a new *image-to-images* translation machine is proposed based on the *pix2pix* network which is known for its promising results in the domain of *image-to-image* translation and segmentation [18]. For this reason, the *pix2pix* network and its implementation are modified to fit the need for multitasking (*pix2pix MT*). As far as the authors know, the proposed network is a first attempt at expanding the application of *image-to-image* network to *image-to-images* with the ability to generate more than one desired output at once. Furthermore, the dilated convolution technique [19] is employed in specific layers of the generator, which is shown to improve further the results. Hence, this design is referred to as the *pix2pix MTdG* model given the implication of *multitask pix2pix* and the inclusion of the dilation property in the generator. More specifically, by feeding a CXR image to the *pix2pix MTdG* network, the proposed model will generate automatically two output images simultaneously, which are the image of the bone suppressed lungs and the image containing the segmentation masks of the heart and lungs. Experimental results show that the *MTdG* network yields comparable results to the state-of-the-art methods that deal with these tasks individually. Results which are evaluated with several metrics and verified using 5-fold cross-validation along with the significance test exhibit promising outputs for both tasks. Moreover, the conceptual design of the model can be generalized to extend to other applications. To confirm this assertion, two different applications are implemented involving 1) neuroimage modality conversion for cross-modality generation of T2-flair and T1-inverse from the T1 input image, and 2) low-dose Computed Tomography (LDCT) for image enhancement and segmentation of kidneys. The results obtained come in support of the merits of the proposed *image-to-images* network in such critical applications where two desired outputs can be obtained simultaneously with improved accuracy and with better system efficiency.

The contributions of this work can be summarized as follows: 1) Design and implementation of a multitask network that for the first time augments the traditional *image-to-image* translation model to an *image-to-images* translation model while improving both accuracy and computational efficiency of the *multitask pix2pix* model. 2) An architectural design of the model that allows for two critical tasks of CXR image analysis, namely bone suppression and organ segmentation, to be performed simultaneously through the use of efficient network parameters verified and augmented by an ablation study. 3) A generalized construct of the model making it more amenable to other application domains, where the results of two more medical applications for brain MRI cross-modality generation and low-dose CT image enhancement and segmentation are provided. 4) All the software code for the different variations of this work are publicly shared online including *multitask pix2pix* for the research community to replicate such work or extend its research potential to other applications. [2]

The rest of this paper is organized as follows: Section II reports the literature review of related work. Section III explains the proposed method, specifies the material used in conducting this study and provides the evaluation strategy used to assess its merits. The experimental results are presented and discussed in section IV which include a section on method generalization and future work. Finally, the conclusion section V provides a retrospective of what was accomplished through this proposed novel approach.

II. Related Work

A. Task 1: Organ Segmentation

Organ segmentation is one of the most difficult tasks in medical imaging due in large part to the elusive thresholding process and the ubiquitous presence of noise [20], [21], but remains an essential task for delineating the anatomical structures of organs and for detecting abnormalities such as enlarged heart or collapsed lungs. Moreover, when performing segmentation in chest radiography, there is need to contend

[2]https://github.com/mohaEs/image-to-images-translationhttps://github.com/mohaEs/image-to-images-translation

with the different shape variations in organs due to age, gender, disease and other health-related issues.

Mansoor et al. [22] presented a comprehensive survey discussing the challenges and accomplishments of the different segmentation methods for lungs which are reported in the literature. By considering the CXR as the imaging modality, several deep learning models based on fully convolutional networks have also been investigated. For instance, a network called *InvertedNet* is proposed to segment the heart, left and right clavicles, and lungs [15]. The well-established *U-Net* architecture has been utilized for segmenting the chest region yielding promising results [23]–[25]. A model called structure correcting adversarial network (*SCAN*) was proposed as a generative adversarial network that uses convolutional layers for heart and lungs segmentation [26]. Another method which incorporates two networks is proposed by [27] with one network used for the initial segmentation process and the second for fine-tuning and correcting the initial results. Moreover, traditional feature extraction methods are widely used for CXR imaging applications [22]. In [28], Ibragimov et al. proposed an approach for lung segmentation and landmark detection based on Haar-like features, a random forest classifier, and spatial relationships among landmarks. A hierarchical lung field segmentation based on the joint shape and appearance sparse learning is proposed in [29], and an atlas-based method is presented in [30].

B. Task 2: Bone and Rib Suppression

In chest X-ray images, the bone structure in the chest area is usually visible, which makes it hard for a radiologist to examine thoroughly the organs and assess any effects of a given disease accurately. Organs' visibility is effective for pulmonary abnormalities screening and detection [5]. Consequently, bone suppression is an essential pre-processing step for suppressing the appearance of bones in the chest X-ray images. One way to tackle this problem is to utilize dual-energy subtraction (DES) imaging [31]. The DES imaging technique captures two or three radiography scans with two or three different energy level of X-ray exposures. The captured images either highlight the soft tissues or bones based on the energy levels. Thus, the suppressed bone image will be estimated by combining the acquired images which include both the soft tissue-selective images and the bone-selective images [32]. Although effective in delineating the bone structure in the chest area, the DES imaging process has a number of shortcomings, among them is its more invasive nature due to the higher radiation dose and the presence of artifacts introduced in the acquisition process due to the effect of heartbeats.

Because of these aforementioned reasons, suppressing the bones in CXR images via traditional image processing techniques is considered safer and is shown to be more effective at overcoming the main challenges faced in CXR images [33]. Along this line of research, a cascaded convolutional neural networks architecture (called *CamsNet*) [34] is proposed to predict the bone gradients in CXR images progressively with the ability of suppressing the bones as a consequence of these determined gradients. Convolutional neural filters are exploited in [35] and are shown to be effective for bone suppression as well. Another recent method is developed by Chen *et al.* [36] which anatomically compensates for the ribs and clavicles by specific multiple massive-training artificial neural networks (*MTANNs*) combined with total variation (TV) minimization smoothing along with a post-processing by histogram-matching. In another study, Gusarev et al. proposed two deep learning architectures that perform bone suppression and create a soft tissue image. Considering bones as a noise level that is affecting these chest images [37], they tried to minimize the presence of this noise (i.e., bone) while still preserving the sharpness of the image for the eventual organ segmentation. In [20], many of the noise suppressing methods reviewed shared the objective of removing as much of the noise as possible while preserving most of the relevant details in the image. Another bone suppression method, based on deep adversarial networks and 2D Haar wavelet decomposition, has been proposed in [38]. Their method was mostly based on the theory of *pix2pix* network [18], a well-known conditional generative adversarial network. The *pix2pix* network is also used as the cornerstone of our proposed multitask model, which will be described in section III. Bone suppression is also used as a pre-processing step, where bone suppressed CXR images are then fed as input images to algorithms such as *CheXNet* in order to enhance the segmentation process and improve the results of the machine (automated) diagnosis [39]. The impact of bone suppression on machine diagnosis using deep learning networks have been thoroughly investigated and detailed in [40] and [41]. There are also some commercially available computer-aided detection (CAD) systems such as *Phillips* [42], *ClearRead* [43] and *Caresteam* [44].

C. Joint Tasks via Multitask Learning

In multitask learning, multiple tasks are solved at the same time by exploiting commonalities and differences across them. In contrast to training separate single task models, the multitask scheme can result in the following improvements [45]: 1) Coupling tasks makes the overall system achieve more accurate results. For example, in [46], a multitask learning approach based on deep convolutional networks is proposed for facial landmark detection, with the auxiliary tasks of head pose estimation, gender classification and facial visibility, yielding more accurate results for each of the tasks. 2) Efficiency in learning optimizes the number of required parameters, memory or storage requirements, computational time and training convergence rate. Hence, fewer but optimal parameters and lower memory requirements are desirable in deploying such algorithms on conventional devices like mobile phones and personal computers [47].

Aside from the improvement in learning efficiency by exploiting the interplay between tasks, and the need for only half the weights required of the multitask model in contrast to the two tasks run separately, other benefits acquired through the intrinsic functions of the multitasking model do not have straightforward or intuitive reasons. However, an additional function can be used to act as a regularization mechanism in other machine learning problems and use the algorithm

to find a solution on a smaller area of representations at the intersection of all tasks. Also, the feature selection and filter values can be reassessed and made more sufficient to understand the nature of the inputs to the model. The overall motivation here is to be able to perform the two tasks jointly via one deep network with the ability for improving both system efficiency and accuracy in the results.

D. Image-to-Image Translation: Applications and Methods

In general, the *image-to-image* translation (I2I) is the process of translating an input image X to a corresponding output image Y, and this correspondence could mean different things for the different context of the application at hand. Such I2I techniques could involve translations such as low-resolution ⇔ high resolution, blurry ⇔ sharp, thermal or grayscale ⇔ color, synthetic ⇔ real, low-dose rate (LDR) ⇔ high-dose rate (HDR), noisy ⇔ clean, image ⇔ painting, day ⇔ night, summer ⇔ winter, bad weather ⇔ good, foggy ⇔ clear, semantic labeling (segmentation) ⇔ realistic photo, aerial ⇔ map, edges and sketch ⇔ photo and so on, in which symbol ⇔ shows the bidirectionality between desired task and context [18], [48]. I2I techniques are also used in medical imaging for segmentation, denoising, super-resolution, modality conversion, CT and MRI reconstruction, among others [48].

I2I has been studied for decades, and different approaches are reported on the basis of filtering, optimization, dictionary learning, deep learning, and more recently generative adversarial network (GAN). While deep learning methods omit the hand-crafted features and GAN methods omit the hand-crafted objective functions, they both remain the most promising methods in data science. GAN-based I2I research in computer vision has yielded different learning models, with a myriad of applications and promising outcomes. In general, there are two categories of methods and applications based on the relation between input and output images: *1) Unsupervised/Unpaired/Unaligned/Unregistered* such as style changing, photo to painting, hair/face and color-changing, weather changing, and *2) Supervised/Paired/Aligned/Registered* such as supervised segmentation and labeling, denoising and super-resolution.

The unpaired category is not relevant in the proposed application of CXR image analysis since the problem at hand is supervised with paired and aligned input/output images. Until now, *pix2pix* [18], *CRN* [49], *BicycleGAN* [50], *SIMS* [51], *SPADE* [52] and *pix2pixHD* [53] remain the most important methods for the paired category. While *pix2pix* and *BicycleGAN* are dealing with a family of applications, others are just considering semantic labels to realistic photo translation. *BicycleGAN* is an *image-to-image* translation with potentially multiple outputs. For example, *BicycleGAN* is able to analyze and translate a given night image to synthesized day images with different types of lighting, sky, and clouds. Each different possibility is generated by feeding a random noise sampled from a known distribution (e.g., a standard normal distribution) along with the input image. Therefore, *pix2pix* is the only general method relevant to our problem, with bone suppression and organ segmentation being a paired/supervised/aligned/registered problem with no randomized output possibilities. The *pix2pix* and its variation are also used for image segmentation and landmark localization in vocal tract area [54], [55].

III. Materials & Methods

A. Methodology

1) Background: In recent years, generative adversarial networks (GANs) and conditional generative adversarial networks (cGANs) have gained a lot of attention because of their superior performance in generation, segmentation, and translation empowered by an adversarial scheme [56]. The GAN architecture consists of two 'adversarial' models trained to work against each other: the generator aiming at generating an output and deceiving the discriminator and a discriminator component aiming at segregating the real output from the fake ones. In conditional mode (cGAN), both generator and discriminator are conditioned on ground truth labels or images. For example, in this study, the segmented organs and the bone suppressed images are the conditions and the generator is set up to generate this type of images.

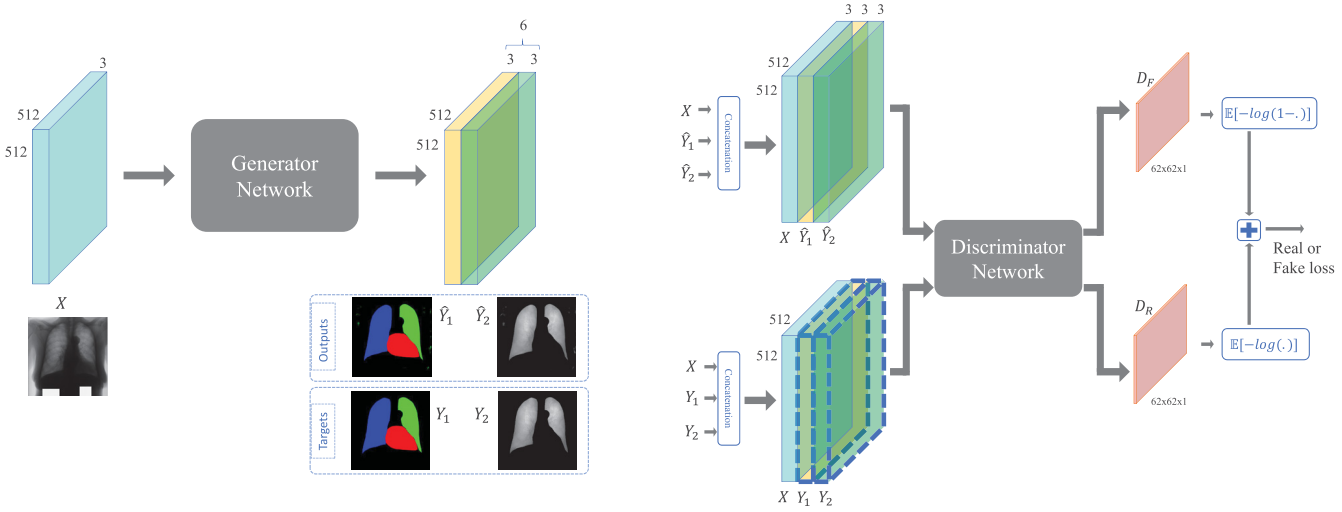
Generators of GANs are intended to learn the mapping from a random noise vector z to an output image y, *i.e.,* $G : z \Rightarrow y$ while cGANs are conditioned by an observed image x *i.e.,* $G : \{x, z\} \Rightarrow y$. The generator G would learn to produce outputs, which could not be distinguished as "fake" images by an adversarially trained discriminator, D. The objective of a GAN and of a conditional GAN can be expressed through equations (1) and (2) respectively, where $\mathbb{E}$ is the *Expectation* over the population. Generator G tries to minimize an objective function against an adversarial D which tries to maximize it, *i.e.,* a *minimax* game as $\hat{G} = arg\ min_G\ max_D\ \mathcal{L}_{GAN}$ and similarly $\hat{G} = arg\ min_G\ max_D\ \mathcal{L}_{cGAN}$.

$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}\left[log D(x, y)\right] + \mathbb{E}\left[log(1 - D(G(x, z)))\right] \quad (1)$$

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}\left[log D(x, y)\right] + \mathbb{E}\left[log(1 - D(x, G(x, z)))\right] \quad (2)$$

2) Proposed Model: A conditional generative adversarial network, called *pix2pix*, is selected as an *image-to-image* translation network to use and modify to meet the intended objectives of bone suppression and image segmentation [18]. The *pix2pix* is used in the proposed method, because 1) It is the main general *image-to-image* translation method; 2) It shows promising prospects for accurate organ segmentation [18], [57]–[59]; 3) It is intrinsically a collection of filters and would be reasonable to perform bone and rib suppression as an *image-to-image* translation task [38].

The generator of *pix2pix* contains an auto-encoding network of convolutional layers with skip connections. The discriminator is also a convolutional neural network (CNN) called *PatchGAN* discriminator as introduced in [18], which attempts to determine whether each patch with size $n \times n$ in an image is real or fake, where n can be much smaller than the full size of the image N. Specifically, *PatchGAN* discriminator is a CNN



a) Generator. b) Discriminator.

Fig. 1. Architecture of the presented image to images translation, *multitask pix2pix*. In this figure, X, Y_1, Y_2, $\hat{Y}_1$, $\hat{Y}_1$ are the images of input CXR, targets of task 1 and task 2, output for task 1 and task 2, respectively. Notice that, all the images, input, output and target have three channels.

which produces a matrix of size $k \times k \times 1$ from an input tensor (or image) of size $N \times N \times *$ where $k = N/n$ and each element in the output matrix indicates the status of the corresponding receptive field on the input tensor (*i.e.* a $k \times k$ *PatchGAN* classifies $k \times k$ patches of the input image/tensor as real or fake). The input tensor for the *PatchGAN* discriminator is a tensor built by concatenation of the input-target pair and the input-output pair for the discriminator to produce an estimation on how realistic they look [18].

Figure 1 shows our model aiming to convert the input CXR image (X) into the desired output (Y), which is the concatenation of desired targets, Y_1 as the organs' segmentation masks and Y_2 as the bone suppressed CXR image, *i.e.,* $Y : Y_1 \parallel Y_2$ where $\parallel$ shows concatenation in the channel axis. The input and output tensors of the generator network are X and $\hat{Y} : \hat{Y}_1 \parallel \hat{Y}_2$ where $\hat{Y}_2$ is the output image corresponding to the bone suppression task and $\hat{Y}_1$ is the output image corresponding to the organ segmentation task which include the masks for the heart (colored red), left lung (colored blue), right lung (colored green) and background (colored black). In fact, the generator creates a tensor with 6 channels, which are the concatenation of $\hat{Y}_1$ and $\hat{Y}_2$.

The discriminator network acts in a similar fashion to *PatchGAN* in order to produce two output matrices, D_R and D_F corresponding to the real and fake input tensors (R and F). The fake input tensor (F) is a concatenation of CXR input image and outputs ($F : X \parallel \hat{Y}_1 \parallel \hat{Y}_2$), and the real input tensor (R) is the concatenation of the CXR input image and targets ($R : X \parallel Y_1 \parallel Y_2$). If the discriminator is trained perfectly, it will create D_R matrix of 1 values and D_F matrix of 0 values. On the other hand, if the generator is successful in fooling the discriminator, D_F would be a matrix of 1 values. The loss functions for training the generator and discriminator are as expressed in equations (3) and (4) where $|\ |_1$ defines the $L1$ distance or norm. In the training phase of the networks, for each batch feeding step: 1) The generator generates output images, 2) The discriminator looks at the real pair tensor (R) and the fake pair tensor (F) and produces an estimate on how realistic they look (D_R and D_F), 3) The weights of the discriminator are then adjusted based on the $\mathcal{L}_D$, and 4) The generator's weights are then adjusted based on $\mathcal{L}_G$.

$$\mathcal{L}_G = \mathbb{E}[-log(D_F + \epsilon)] + \lambda\ \mathbb{E}[|\ Y - \hat{Y}\ |_1] \quad (3)$$

$$\mathcal{L}_D = \mathbb{E}[-(\ log(D_R + \epsilon) + log(1 - D_F + \epsilon)\ )] \quad (4)$$

Furthermore, in order to to produce more efficient receptive fields, dilated convolutions [19] are utilized in some specific layers of the generator. The encoder of the generator consists of 8 layers and dilated convolutions are used with dilation rate 2 in layers 2 through 7 in the proposed structure of the *MTdG* network. The effects of using and not using the dilated convolutional layers are contrasted in the results section.

B. Data

The *Japanese Society of Radiological Technology (JSRT)* is the only publicly available database where both desired tasks are available and hence most suitable for training and evaluating the proposed model. This dataset consists of CXR images collected by *JSRT* [60] and is publicly available in [61]. Segmentation masks for lungs and the heart were created later by [62] and are now available in [63]. The *JSRT* dataset comprises 247 CXRs, including images with and without lung nodules. All images have a resolution of 2048×2048 in gray scale with a color depth of 12 bits. While there is no publicly available dataset for bone and rib suppression based on DES, Juhasz et al. developed a method for bone suppression [64]. Their results on the *JSRT* dataset have become publicly available in [65]. This dataset is used to accomplish the second task of bone suppression as well.

As noted in *CheXNeXt* and through other investigations, the 512×512 resolution is sufficient for classifying lung-related diseases and for localizing nodule(s) [6], [11], [66], [67]. This resizing of images helped to significantly minimize the computational requirements and, as the results

will prove, high accuracy is maintained. Moreover, we anticipate that the resizing of images to 512 × 512 pixels could help in their effective use towards the development of new pre-processing methods for improving computer-aided diagnosis. The image intensities are set up with an 8-bit grayscale resolution in the range from 0 to 255. In order to train the machine learning and especially the deep learning networks, it is essential to have enough number of samples that cover the different variations [68], [69]. Therefore, the original images along with their corresponding masks and suppressed bone images were augmented by rotating them via 10 and -5 degrees, along with translations of (30, 10) and (-20,-10) pixels in reference to the (x,y) coordinates. Through this process, the size of the dataset has been increased by 5 times, to a total number of 1,235 images along with their corresponding ground-truths for the two tasks.

Furthermore, in order to assess the effectiveness of multitask *image-to-images* translation in terms of its generalization to other applications, two additional experiments are included in the *Generalization and Future Work* section. The dataset used for the low-dose CT (LDCT) experiment is from *"Multi-Atlas labeling beyond the cranial vault"* challenge containing CT scans and corresponding segmentation labels of 13 abdominal organs of 50 subjects [70]. In order to simulate the LDCT scans from CT scans, the method based on additive Poisson noise on sonograms of CT scans is used [71]. The dataset for the neuroimaging experiment is *MRBrainS18* challenge dataset containing multi-modal MRI brain images (T1, T1-inverse and T2-flair) with segmentation labels of gray matter, white matter, cerebrospinal fluid, and other structures on 3T MRI scans of the brain of seven subjects [72]. Augmentation with the same configuration is also exploited for the LDCT and neuroimaging experiments.

C. Implementation & Evaluation

The proposed model has been implemented and modified to comply with the multitasking scenario based on the publicly available *pix2pix* code [73]. For validation purposes, both the extended code that supports all these different variations and the video showing this process at work are made available through the Internet to the research community. The intensity channel of the input CXR image is replicated to support the CNN 3-channel RGB input data expectations. The size of the input/output images and kernel (or filter) are $512 \times 512 \times 3$ and 4×4, respectively and 5-fold cross-validation are considered as default. To make a fair comparison with the results of the current state-of-art-techniques, the resolution of 256×256 and 3-fold cross-validation has been considered in this study as well. The network has been implemented using Python and the Tensorflow library. All computations for training the network have been performed on a system equipped by NVIDIA GPU Quadro M6000 with 24 GB memory. The parameter λ in equation 3 is set to 10 and the learning rate of the Adam optimizer is set to 0.0002. The training is stopped when the L1 loss reaches almost 0.005. Because of a limited number of subjects, leaving-one-subject-out cross-validation scheme is used for the LDCT and neuroimaging experiments. For comparative assessment of the results obtained with the different methods, the average, standard deviation, box plots including median, percentiles, and outliers along with the t-test for statistical significance are considered regarding the different metrics.

IV. Results and Discussion

In this section, the experimental results for both tasks are reported and compared with state-of-the-art methods in the first two subsections. This is followed by a subsection which provides the performance and qualitative analysis, investigates the ablation study and demonstrates how this model is amenable to other application domains (i.e. generalization of the model) with preliminary results obtained on low-dose CT image processing and on a neuroimage translation problem. The generalization subsection shows the benefits of multitasking in comparison with single tasking and the potential for future work of *pix2pix MTdG* towards resolving other *image-to-images* translation problems. Notice that, for simplification in plotting the figures *p2p* notation is used instead of *pix2pix*.

A. Task 1: Organ Segmentation

The segmented regions of the heart, the left lung and right lung generated by the model as output masks, are associated with the same regions in the ground-truth by using standard evaluation metrics in image processing, namely the Dice and Jaccard scores, false-negative rate (FNR) and false-positive rate (FPR). The Jaccard index is a metric that measures the percent overlap between the target ground-truth mask (GT) and our prediction mask (PM). The overlap area between GT and PM would be the true positive area (TP). The area which is predicted in PM but is not in PM is a false-positive (FP) and inversely for false-negative (FN) defining the area which is in GT but not in PM. Jaccard metric is closely related to the Dice coefficient, which is not as easily described geometrically. The false-positive rate indicates the area ratio of predicted mask which had no associated ground truth mask, and similarly false-negative indicates the area ratio of the ground truth mask which had no associated predicted mask. These standard evaluations metrics can be expressed mathematically as follow where symbol ! defines the binary negation operator:

$$Dice = \frac{2 \times (PM \cap GT)}{PM + GT} = \frac{2 \times TP}{TP + FP + TP + FN} \tag{5}$$

$$Jaccard = \frac{PM \cap GT}{PM \cup GT} = \frac{TP}{TP + FP + FN} \tag{6}$$

$$FNR = \frac{!PM \cap GT}{GT} = \frac{FN}{TP + FN} \tag{7}$$

$$FPR = \frac{PM \cap !GT}{GT} = \frac{FP}{TP + FN} \tag{8}$$

The well-known U-net method [74] is used as the competitor due to its state-of-the-art performance in various segmentation applications [75] as well as for the reported CXR segmentation studies with promising results [23]–[25]. The average and standard deviation of the segmentation results, by all metrics, are summarized in Table I. The best

TABLE I
SEGMENTATION RESULTS OF DIFFERENT METHODS WHILE THE BEST SCORES ARE COLORED BLUE

		u-net	pix2pix MTdG	p-value
Dice	Left lung	0.983 ± 0.005	0.990 ± 0.006	1.2 e-93
	Heart	0.965 ± 0.014	0.977 ± 0.015	3.6 e-57
	Right lung	0.980 ± 0.009	0.988 ± 0.006	5.5 e-99
	Average	0.976 ± 0.007	0.985 ± 0.007	-
Jaccard	Left lung	0.967 ± 0.010	0.980 ± 0.011	7.7 e-97
	Heart	0.933 ± 0.026	0.956 ± 0.027	3.8 e-59
	Right lung	0.961 ± 0.017	0.977 ± 0.012	6.9 e-105
	Average	0.953 ± 0.013	0.971 ± 0.013	-
FNR	Left lung	0.010 ± 0.005	0.008 ± 0.005	8.8 e-18
	Heart	0.028 ± 0.017	0.018 ± 0.017	7.1 e-24
	Right lung	0.015 ± 0.008	0.010 ± 0.006	3.0 e-45
	Average	0.017 ± 0.006	0.012 ± 0.007	-
FPR	Left lung	0.024 ± 0.013	0.013 ± 0.013	3.0 e-73
	Heart	0.043 ± 0.033	0.028 ± 0.029	2.3 e-22
	Right lung	0.026 ± 0.017	0.013 ± 0.012	8.3 e-67
	Average	0.031 ± 0.015	0.018 ± 0.013	-

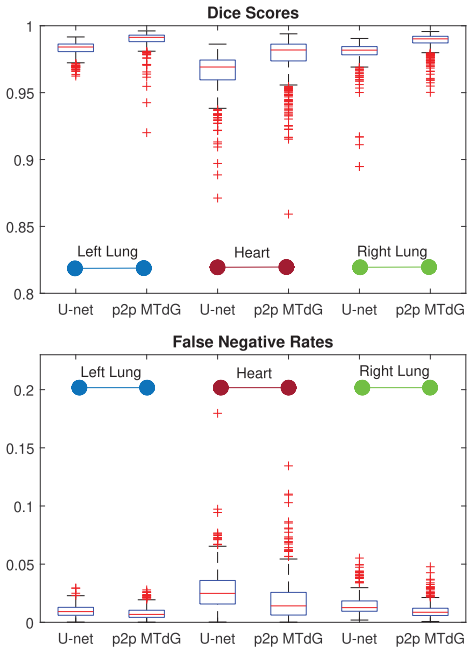


Fig. 2. The box plot of the segmentation accuracy achieved by *u-net* and *pix2pix MTdG*. up) Dice Scores, Down) False Negative Rates.

achieved results, highlighted in blue, demonstrate that the *multitask pix2pix* with an embedded dilation in the generator (*MTdG*) surpasses the *u-net* method. For further statistical investigation, figure 2 shows the box plots of the segmentation scores evaluated using the Dice and false-negative rate of the heart, left and right lung for the *u-net* and *pix2pix MTdG*. The inner line, the bottom and top edges of the box indicate median, the 25th, and 75th percentiles, respectively. While the whiskers are extended to the most extreme not outlier data points, while the outliers are plotted individually using the '+' symbol. As an example, the segmentation result of the *pix2pix MTdG* for the best and worst achieved Dice scores are shown in Figure 3.

To the best of our knowledge, no multitask framework has been found in the literature to benchmark the proposed multitask network for our tasks. A comparison between *u-net* (implemented by us) and *pix2pix MTdG* along with the p-values of student's t-test is provided in Table I which clearly shows that the *pix2pix MTdG* method yields better results with

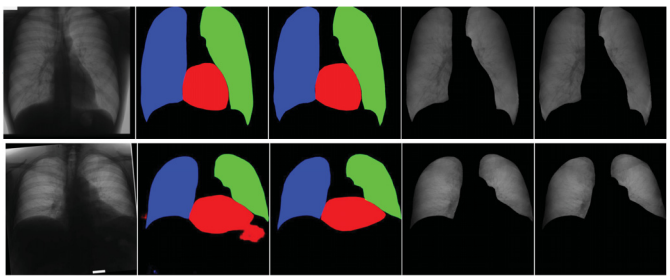


Fig. 3. Results by proposed method *pix2pix MTdG* regarding to the best (top) and worst (bottom) Dice scores. Columns left to right are, input image, segmentation result, segmentation target, bone suppression result, and bone suppression target. Top) Best Dice: average of Dice, Jaccard, FPR and FNR are 0.99, 0.99, 0.01 and 0.01. Bottom) Worst Dice: average of Dice, Jaccard, FPR and FNR are 0.94, 0.90, 0.12 and 0.01.

a significant difference ($p < 0.001$). However, to contrast these results with other methods in the literature, due to the variations in utilizing different folding schemes and image sizes, a fair comparison of the results is not a straightforward process.

The results of different state-of-the-art algorithms on the JSRT dataset are summarized in Table II. The settings reported from each method are also provided in Table II with '−' to mean that the value is not reported. For a fair comparison, the proposed method has also been tested using 256 × 256 image size, 3-fold cross-validation and without any augmentation. As presented in the Table, while the scores are really close to each other, the best-achieved results are the ones provided by *pix2pix MTdG* in 512 × 512 image resolution. In all these other settings, the *pix2pix MTdG* performance is still reasonable and is comparable to the performance of other techniques.

B. Task 2: Bone Suppression

The second task of the *pix2pix MTdG* network is bone suppression. The results of this task are evaluated via the structural similarity index (SSIM) metric for similarity estimation, and the root mean squared error (RMSE) metric to measure the difference between predicted and actual values [77]. The RMSE measure between two X and Y images is expressed by Equation (9), where N in the total number of pixels in image and i is the pixel index. SSIM is a reference-based quality assessment metric, which compares the local patterns of pixel intensities between the reference and output images. The maximum value of 1 implies that the two images are structurally similar, while a zero value indicates that there is no structural similarity between them. Usually, the SSIM index is calculated via windowing on the images with 8 × 8 window size and 1 pixel striding. At the end, the mean of the computed values (M-SSIM) would be reported. The M-SSIM measure between two images X and Y and the default SSIM measure between two windows W_i^X and W_i^Y are as defined by Equations (10) and (11) where $\mu_{W_i^X}$, $\mu_{W_i^Y}$, $\sigma^2_{W_i^Y}$, $\sigma^2_{W_i^X}$, and $\sigma_{W_i^X W_i^Y}$ show the average of W_i^X, the average of W_i^Y, the variance of W_i^X, the variance of W_i^Y and the Covariance of W_i^X and W_i^Y, respectively. The default settings

TABLE II
COMPARISON THE SEGMENTATION RESULTS OF DIFFERENT METHODS ON JSRT DATASET

Method	Image Size	Augmentation	Evaluation scheme	Dice	Jaccard	Dice	Jaccard
				Lungs		Heart	
Human observer [62]	2048×2048	No	-	-	0.946	-	0.887
InvertedNet [15]	256×256	No	3-fold CV	0.974	0.950	0.937	0.882
u-net by [24]	256×256	No	5-fold CV	0.976	0.962	-	-
u-net by [23]	256×256	No	5-fold CV	-	0.959	-	0.899
u-net by us	512×512	Yes	5-fold CV *	0.981	0.964	0.965	0.933
MTdG (proposed)	512×512	Yes	5-fold CV *	0.989	0.978	0.977	0.956
u-net by [25]	512×512	Yes	train/test split (80%/20%)**	0.986	-	-	-
SegNet by [24]	256×256	No	5-fold CV	0.979	0.955	0.944	0.896
SCAN [26]	400×400	No	train/test split (209/38)	0.973	0.947	0.927	0.866
FCN by [24]	256×256	No	5-fold CV	0.974	0.950	0.942	0.892
MTdG (proposed)	256×256	No	5-fold CV	0.974	0.962	0.934	0.928
MTdG (proposed)	256×256	No	3-fold CV	0.962	0.953	0.921	0.916
ASM tuned [62]	256×256	No	2-fold CV	-	0.927	-	0.814
Hybrid voting [62]	256×256	No	2-fold CV	-	0.949	-	0.86
Seghers et. al [76]	256×256	No	train/test split (50/44)	-	0.951	-	-
Ibragimov et. al [28]	-	No	-	-	0.953	-	-

*: p-values from significance test are reported in Table I.
**: JSRT dataset is not used for this work.

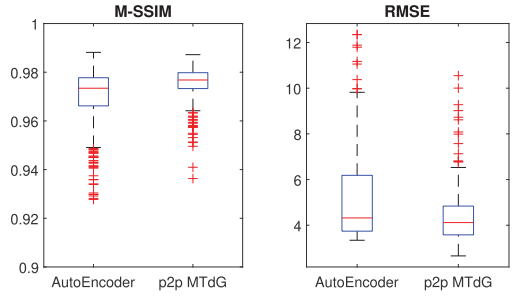


Fig. 4. Bone suppression task showing boxplots of the results. Left) M-SSIM similarity score. Right) RMSE difference.

of $c_1 = (0.01\ L)^2$ and $c_2 = (0.03\ L)^2$ are considered with L being the dynamic range of the pixel-values (i.e. 255 in our experiment).

$$RMSE(X,Y) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2} \tag{9}$$

$$M\text{-}SSIM = \frac{1}{N}\sum_{i=1}^{N} SSIM(W_i^X, W_i^Y) \tag{10}$$

$$SSIM(W_i^X, W_i^Y) = \frac{(2\mu_{W_i^X}\mu_{W_i^Y} + c_1)(2\sigma_{W_i^X W_i^Y} + c_2)}{(\mu^2_{W_i^X} + \mu^2_{W_i^Y} + c_1)(\sigma^2_{W_i^X} + \sigma^2_{W_i^Y} + c_2)} \tag{11}$$

While there is much literature about bone suppression in CXR images and even with the availability of proprietary software, there is no shared dataset available nor is there any open source codes or models that are shared with the research community. There is just one competitor study for bone suppression with shared source code in which the authors used an interesting *AutoEncoder* network architecture [37]. The box plots of the achieved M-SSIMs and RMSEs for the *pix2pix MTdG* and *AutoEncoder* are shown in Figure 4 and Table III, proving the good performance of the proposed *pix2pix MTdG*. The best and worst results with regards to the RMSE measurement are shown in Figure 5.

TABLE III
RESULTS OF BONE SUPPRESSION TASK VIA DIFFERENT METHODS

	AutoEncoder [37]	pix2pix MTdG	p-value
MSSIM	0.970 ± 0.011	0.976 ± 0.006	8.6 e-24
RMSE	5.096 ± 1.812	4.297 ± 1.046	5.3 e-45

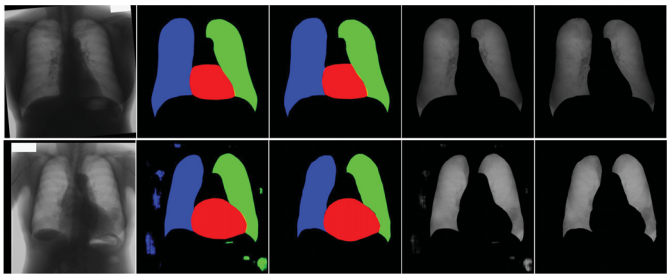


Fig. 5. Results by proposed method *pix2pix MTdG* regarding to the best (top) and worst RMSEs (bottom). Columns left to right: input image, segmentation result, segmentation target, bone suppression result and bone suppression target. Top) Best RMSE: SSIM: 0.99, RMSE: 2.66. Bottom) Worst RMSE: SSIM: 0.94, RMSE: 10.55.

C. Discussion

As discussed earlier, the proposed *pix2pix MTdG* method provides promising results in accomplishing both segmentation and bone suppression tasks simultaneously. In this subsection, other characteristics of the proposed method are discussed and a summary of these results is provided in Table IV.

1) Performance Analysis: The performance of the proposed method is assessed here with respect to the results obtained and the network parameters that were considered. The following remarks can be made:

a) As shown in Tables I and IV, the proposed multitask pix2pix with dilation (*pix2pix MTdG*) achieves the best results for both organs segmentation and bone suppression tasks.

b) Using multitask pix2pix without dilation was not as effective for improving the results.

c) Another advantage of *pix2pix MTdG* is in the number of required parameters of the network, which is an intrinsic requirement for the multitask pix2pix scheme. The trainable parameters of pix2pix, u-net, and multitask pix2pix are 57, 190, 084; 31, 084, 008 and 57, 199, 303, respectively.

TABLE IV
SUMMARY OF THE PROPERTIES OF DIFFERENT METHODS

Task		pix2pix ST	pix2pix STdG	U-net	AutoEncoder	pix2pix MT	pix2pix MTdG
Segmentation	Average Dice	0.977 ± 0.008	0.984 + 0.008	0.976 ± 0.007	-	0.978 ± 0.008	0.985 ± 0.007
	Average FNR	0.018 ± 0.008	0.014 + 0.009	0.017 ± 0.006	-	0.018 ± 0.008	0.012 ± 0.007
Rib Suppression	Average MSSIM	0.969 ± 0.009	0.974 ± 0.007	-	0.970 ± 0.011	0.968 ± 0.007	0.976 ± 0.006
	Average RMSE	5.296 ± 1.688	4.697 ± 1.587	-	5.096 ± 1.812	5.382 ± 1.470	4.297 ± 1.046
	No. Parameters	2 × 57,190,084		31,084,008	64,400	57,199,303	
	Minimum Epochs	250		48	150	300	
	Training time	786 = 2 × 393 min		248 min	852 min	662 min	

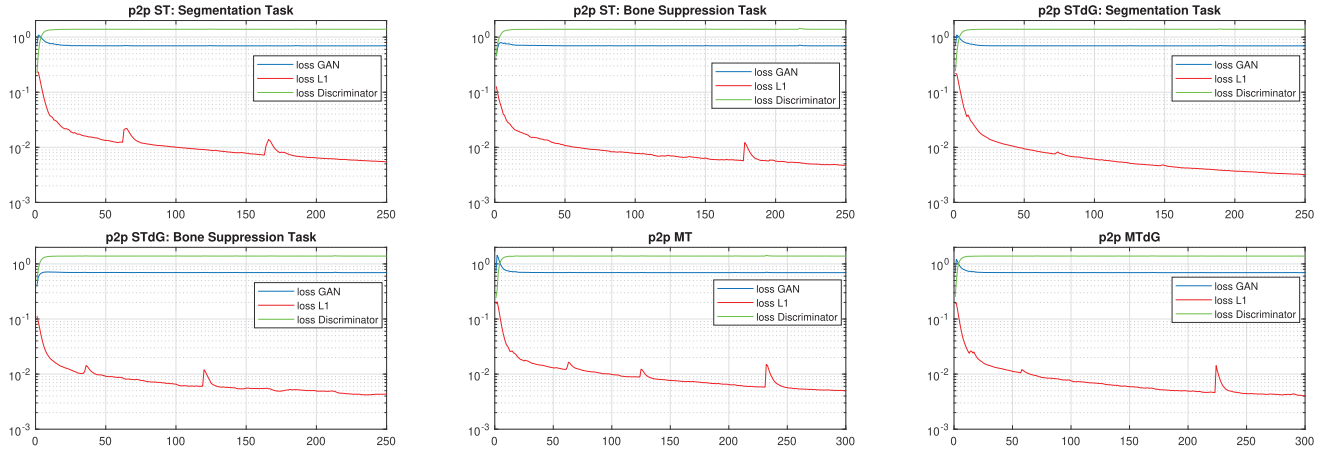


Fig. 6. The loss curves of the different schemes for a training session as a function of epochs. p2p: *pix2pix*, ST: single task, STdG: single task with dilation in generator, MT: multitask, MTdG: Multitask with dilation in generator.

Note that the u-net architecture can only perform the segmentation task. To employ two separate pix2pix networks for the two tasks, a large number of training parameters $2 \times 57,190,084$ would be needed. In other words, the performance of *multitask pix2pix* is reasonable in the number of parameters used while it maintains comparable good results to the state-of-the-art techniques.

d) Nonetheless, the *multitask pix2pix* framework has one drawback, which is the number of required iteration/epochs for the training phase. In the experiments conducted in this study, *pix2pix MTdG* required almost 300 epochs while u-net and single task *pix2pix* converged in only 50 and 250 epochs, respectively, meaning that the multitask framework requires more training time in comparison with other methods. It is worth noting that the above concern is just for training and since a pre-trained network is used to generate the output in the testing phase, both methods perform similarly fast requiring only 1.2 seconds.

2) Ablation Study: In order to investigate the effectiveness of dilation and multitasking, the ablation study is considered. For this reason, six processing schemes are addressed: two *single-task pix2pix* (*p2p ST*) for two tasks, two *single-task pix2pix with dilation* (*p2p STdG*), *a multitask pix2pix* (*p2p MT*) and *a multitask pix2pix with dilation* (*p2p MTdG*). Figure 6 shows the loss curves for the training session of the schemes with respect to epochs. The loss GAN, loss L1 and loss discriminator expressed as $\mathbb{E}[-log(F + \epsilon)]$, $\mathbb{E}[| Y - \hat{Y} |_1]$ and $\mathbb{E}[-(log(R + \epsilon) + log(1 - F + \epsilon))]$, respectively, were included in the loss functions (3) and (4). The training is stopped when the L1 loss reaches and stabilizes at almost 0.005. As shown in Fig. 6, all of the schemes are able to converge and make the generator the declared winner. Furthermore, while the multitask schemes need more epochs for reaching the desired L1 loss value, there is no significant difference between the achieved final loss values.

In contrast to the almost similar final training loss values of the different schemes, as shown in Table IV, there are differences in the test sessions and in the outputs of the schemes' models. The best results are achieved by *pix2pix MTdG* while the multitasking without dilation was not helpful. This could be inferred as a hyper-parameter selection criterion, as the dilation makes the receptive fields bigger to become more suitable to the nature of the CXR input images and the organs' shapes and sizes. On the other hand, when dilation is exploited in the network, multitasking (*p2p MTdG*) assumes the benefits and generates better results than those obtained from a single task with dilation (*p2p STdG*); this outcome is achieved despite the fact that the multitask network has half of the weights in comparison to two single-task networks. Figures 7 and 8 demonstrate the results of segmentation and bone suppression tasks for different schemes as can be observed from the box-plots of Dice score, false negative rates, RMSE and M-SSIM score in concurrence to the results shown in table IV, especially when comparing for outliers, percentile edges and median line.

3) Qualitative Analysis: Figure 9 presents the segmentation results of the different methods for various subjects. From left to right, the images shown are the input image, the target image, single task *pix2pix* output, *u-net* output, *pix2pix MT* output, and *pix2pix MTdG* output. In retrospect, the following assessments can be made:

a) As presented in the first row, the *pix2pix MTdG* framework delivered the best outcomes in comparison with the other methods.

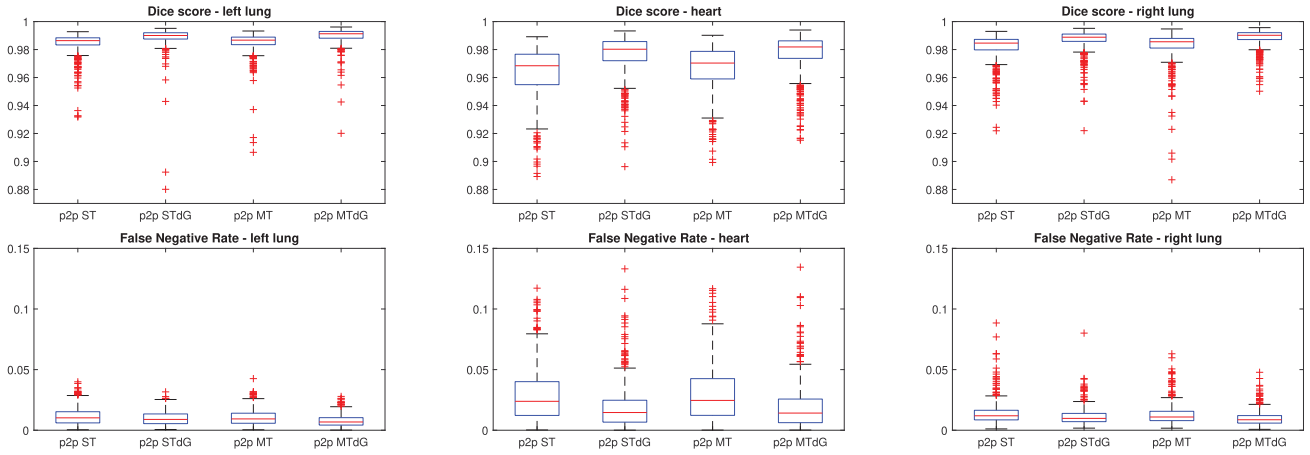


Fig. 7. Segmentation task evaluation showing box-plots for different schemes. (Left) left lung, (Middle) heart, (Right) right lung. (*pix2pix ST*, *pix2pix STdG*, *pix2pix MT* and *pix2pix MTdG*).

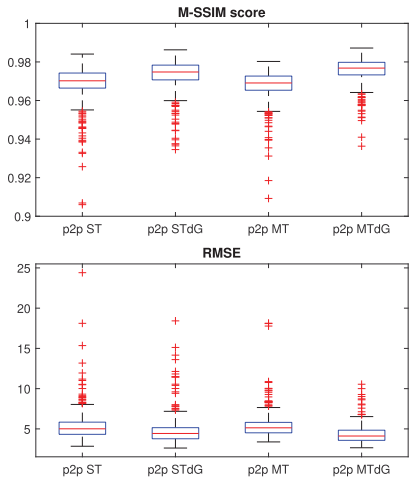


Fig. 8. Bone suppression task evaluation showing the box-plots for the different schemes. (*pix2pix ST*, *pix2pix STdG*, *pix2pix MT* and *pix2pix MTdG*).

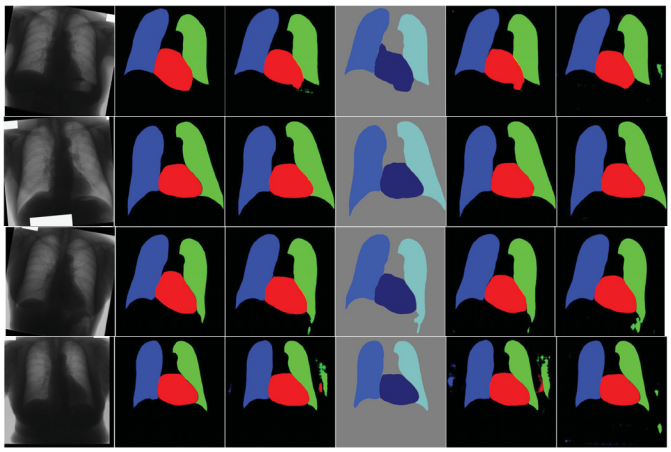


Fig. 9. Segmentation results of different schemes for different subjects. Top to bottom are subjects and left to right are input, target, *pix2pix ST*, *u-net*, *pix2pix MT* and *pix2pix MTdG*.

b) However, the results of all the other methods in the second row are almost the same.

c) The third row shows that the *multitask pix2pix* without dilation achieved better results in contrast to the other methods. This could be related to the fact that this dilation process may not be as effective in some specific cases. We intend to combine the layers with and without dilation in future work to see if the accuracy could be improved further.

d) Moreover, as it is illustrated in the fourth row of Figure 9, the *pix2pix* based methods suffer from false-positive segments such as isolated islands. Although this is the case for the u-net method as well, u-net demonstrates better results in this case. This could be associated with the fact that the loss function of u-net is addressing the segmentation constraints while the loss function of *pix2pix* is constructed to perform the pixel-wise comparison. This drawback could simply be removed by employing post-processing techniques such as connected components and considering the island area, which is not the aim of this paper at this juncture, but could be exploited in future work. Another approach to deal with this drawback is to implement segmentation related loss functions such as dice score to *pix2pix* loss function. While this technique was discussed by some authors in the literature, our investigation did not prove its efficiency in delivering any improvement to our case.

4) Generalization of Method and Future Work: The size of the input CXR images considered in this study are 512×512. This resolution chosen for the proposed method is also exploited in the Stanford's *CheXNext* network [11] with promising results. Nonetheless, manufacturers of X-ray radiography continue to improve its resolution and hence 512×512 could be limiting in obtaining similar optimal results sought with a higher resolution. Since more pixels mean more revealing information that could enhance the prospects for more accurate diagnosis, it is thus reasonable to extend the proposed network for dealing with higher resolution images such as 2048×2048. It should be noted that since the generator of the proposed method is a holistic network (not locally and patch- or block-based), it is expected to yield more promising results with higher resolution images, but at the expense of more convolutional layers and with a higher demand for more variables to contend with, and hence more taxing computational requirements.

With these contending challenges in mind, and in order to show the potential benefits of the *image-to-images* translation and *multitask pix2pix* with respect to the generalization aspect to other domains of application, even under the 512×512 resolution, this section provides preliminary results on two

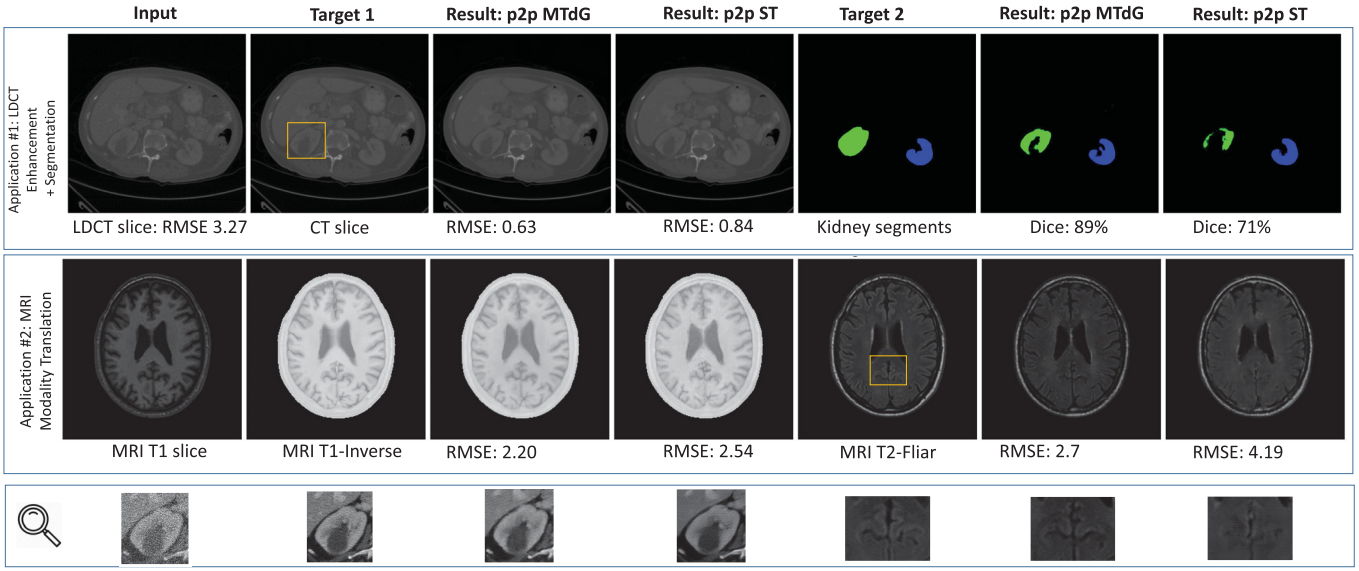


Fig. 10. Preliminary results of two more applications of *multitask pix2pix* for joint tasks. Top row) Low dose CT: image enhancement and segmentation. Mid row) MRI Neuroimage translation: T1 to T1-inverse and T2-flair. Bottom row) Magnification of the selected region (area inside yellow rectangle). p2p: *pix2pix*, ST: single task, MTdG: Multitask with dilation in generator.

more experiments that were conducted. The first experiment involves low dose CT image enhancement and segmentation of kidneys. The second is a neuroimaging translation for cross-modality generation of T2-flair and T1-inverse from the T1 input image.

In the first experiment, Close attention is given in the literature to LDCT imaging for of its use of a lower dose of radiation, for being affordable, and for its faster scanning time, making it most suitable for screening, diagnosis and follow up visits [78]. While segmentation is an intrinsic problem in imaging, likewise in LDCT, the image enhancement aspect that could lead to better image quality is also a state-of-the-art issue which is being addressed in the literature [79].

In the second experiment, while *Cross-modality generation* can serve as an auxiliary method in clinical diagnosis [80], it also has great potential for multimodal registration [81]–[83], segmentation [81], [83], [84], super-resolution [85] and structural information improvement [86] (e.g. MRI 3T to 7T). For this second experiment, an MRI T1 volume of the brain is translated into an MRI T1-inverse and a T2-flair, in a slice by slice fashion.

Figure 10 shows the qualitative and quantitative results (as seen in one slice) of single task and *multitask pix2pix* using leave one subject out (LOSO) as a test evaluation scheme. The network is trained using all of the 2D axial slices from the training subjects and is tested on all of the 2D axial slices of the test subject not seen in the training phase. For these studies, the hyper-parameters and networks, which are the same as the aforementioned for CXR analysis, are not optimized for these applications. For both applications, the multitask method outperforms the single task methods while using only half of the network weights. It is worth mentioning that there is a significant difference in segmentation task for LDCT application and T2-flair task for the neuroimaging application. The magnification area is shown to emphasize the differences. These results clearly show the potential benefits of using *image-to-images* translation to other domains of application involving different imaging modalities even when the proposed model as used for these additional applications are implemented on the model solely based on CXR images.

V. Conclusion

Of all the many existing medical imaging modalities, X-ray imaging remains the most widely used modality as it is the most cost effective and one of the easiest to administer. Chest X-ray remains an essential imaging modality for the diagnosis and follow up treatment of many diseases affecting the lungs, heart and bone structure within the chest area. In this study, a new deep learning based *image-to-images* approach was proposed that simultaneously suppresses the bones that hinder the visibility and scrutiny of organs and nodules and segments the organs within the chest area. Essentially, and for the first time, the architectural design of this deep learning-based model exploits in the most effective way the interplay of parameters in between the two tasks to optimize the outcome for both tasks at once. In order to perform these two essential tasks of bone suppression and organs segmentation, the well-established pix2pix network is extended to generate two output images simultaneously (an image with bones suppressed, and a second image showing the segmented organs), yielding the new *image-to-images* with an automated end-to-end framework instead of the traditional *image-to-image* approach that deals with each task separately. The proposed method was trained via an end-to-end process and is evaluated by cross validation and significance testing with several standard metrics, resulting in highly accurate results for both tasks. Through two additional empirical evaluations involving low-dose CT images and neuroimaging, the proposed architectural design of the model is shown to be amenable for generalization to other domains of application, although developed around CXR imaging. In summary, the contributions of this work can be summarized as follow:

- This work is the first to try to extend the application of *image-to-image* network to *image-to-images* network, while optimizing the use of parameters and securing computational efficiency.
- The network is improved through the inclusion of dilated convolutions in some specific layers, which are shown to improve the accuracy of the results significantly.
- The *image-to-images* network is used to accomplish simultaneously the two common and most needed tasks of bone suppression and organ segmentation in CXR images while optimizing the outcomes for both.
- All of the developed code is shared publicly online for validation purposes and for use by the research community interested in the automated diagnosis and in treatment follow up using chest X rays.

REFERENCES

[1] S. Raoof, D. Feigin, A. Sung, S. Raoof, L. Irugulpati, and E. C. Rosenow, "Interpretation of plain chest roentgenogram," *Chest*, vol. 141, no. 2, pp. 545–558, Feb. 2012.

[2] R. D. Neal *et al.*, "Immediate chest X-ray for patients at risk of lung cancer presenting in primary care: Randomised controlled feasibility trial," *Brit. J. Cancer*, vol. 116, no. 3, pp. 293–302, Jan. 2017.

[3] O. Ruuskanen *et al.*, "Viral pneumonia," *Lancet*, vol. 377, no. 9773, pp. 1264–1275, 2011.

[4] K. C. Santosh, S. Vajda, S. Antani, and G. R. Thoma, "Edge map analysis in chest X-rays for automatic pulmonary abnormality screening," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 11, no. 9, pp. 1637–1646, 2016.

[5] S. Vajda *et al.*, "Feature selection for automatic tuberculosis screening in frontal chest radiographs," *J. Med. Syst.*, vol. 42, no. 8, p. 146, Aug. 2018.

[6] O. Gozes and H. Greenspan, "Lung structures enhancement in chest radiographs via CT based FCNN training," in *Image Analysis for Moving Organ, Breast, and Thoracic Images*. Granada, Spain: Springer, 2018, pp. 147–158.

[7] I. Diamant *et al.*, "Chest radiograph pathology categorization via transfer learning," in *Deep Learning for Medical Image Analysis*. Amsterdam, The Netherlands: Elsevier, 2017, pp. 299–320.

[8] C. Qin, D. Yao, Y. Shi, and Z. Song, "Computer-aided detection in chest radiography based on artificial intelligence: A survey," *BioMed. Eng. OnLine*, vol. 17, no. 1, Dec. 2018.

[9] K. C. Santosh and S. Antani, "Automated chest X-ray screening: Can lung region symmetry help detect pulmonary abnormalities?" *IEEE Trans. Med. Imag.*, vol. 37, no. 5, pp. 1168–1177, May 2018.

[10] A. Karargyris *et al.*, "Combination of texture and shape features to detect pulmonary abnormalities in digital chest X-rays," *Int. J. Comput. Assist. Radiol. surgery*, vol. 11, no. 1, pp. 99–106, Jan. 2016.

[11] P. Rajpurkar *et al.*, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLoS Med.*, vol. 15, no. 11, Nov. 2018, Art. no. e1002686.

[12] A. Majkowska *et al.*, "Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation," *Radiology*, vol. 294, no. 2, pp. 421–431, Feb. 2020.

[13] H. Wang and Y. Xia, "ChestNet: A deep neural network for classification of thoracic diseases on chest radiography," 2018, *arXiv:1807.03058*. [Online]. Available: http://arxiv.org/abs/1807.03058

[14] X. Li *et al.*, "Multi-resolution convolutional networks for chest X-ray radiograph based lung nodule detection," *Artif. Intell. Med.*, to be published.

[15] A. A. Novikov, D. Lenis, D. Major, J. Hladuvka, M. Wimmer, and K. Buhler, "Fully convolutional architectures for multiclass segmentation in chest radiographs," *IEEE Trans. Med. Imag.*, vol. 37, no. 8, pp. 1865–1876, Aug. 2018.

[16] D. Lee, H. Kim, B. Choi, and H.-J. Kim, "Development of a deep neural network for generating synthetic dual-energy chest X-ray images with single X-ray exposure," *Phys. Med. Biol.*, vol. 64, no. 11, Apr. 2019, Art. no. 115017.

[17] S. Tabarestani *et al.*, "A distributed multitask multimodal approach for the prediction of Alzheimer's disease in a longitudinal study," *NeuroImage*, vol. 206, Feb. 2020, Art. no. 116317.

[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[19] P. Wang *et al.*, "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1451–1460.

[20] M. Mafi, H. Martin, M. Cabrerizo, J. Andrian, A. Barreto, and M. Adjouadi, "A comprehensive survey on impulse and Gaussian denoising filters for digital images," *Signal Process.*, vol. 157, pp. 236–260, Apr. 2019.

[21] M. Goryawala, S. Gulec, R. Bhatt, A. J. McGoron, and M. Adjouadi, "A low-interaction automatic 3D liver segmentation method using computed tomography for selective internal radiation therapy," *BioMed. Res. Int.*, vol. 2014, pp. 1–12, Jul. 2014.

[22] A. Mansoor *et al.*, "Segmentation and image analysis of abnormal lungs at CT: Current approaches, challenges, and future trends," *RadioGraphics*, vol. 35, no. 4, pp. 1056–1076, Jul. 2015.

[23] C. Wang, "Segmentation of multiple structures in chest radiographs using multi-task fully convolutional networks," in *Proc. Scandin. Conf. Image Anal.* Tromsø, Norway: Springer, 2017, pp. 282–289.

[24] H. Oliveira and J. dos Santos, "Deep transfer learning for segmentation of anatomical structures in chest radiographs," in *Proc. 31st SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2018, pp. 204–211.

[25] J. Islam and Y. Zhang, "Towards robust lung segmentation in chest radiographs with deep learning," 2018, *arXiv:1811.12638*. [Online]. Available: http://arxiv.org/abs/1811.12638

[26] W. Dai, N. Dong, Z. Wang, X. Liang, H. Zhang, and E. P. Xing, "SCAN: Structure correcting adversarial network for organ segmentation in chest X-rays," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Granada, Spain: Springer, 2018, pp. 263–273.

[27] J. C. Souza, J. O. B. Diniz, J. L. Ferreira, G. L. F. da Silva, A. C. Silva, and A. C. de Paiva, "An automatic method for lung segmentation and reconstruction in chest X-ray using deep neural networks," *Comput. Methods Programs Biomed.*, vol. 177, pp. 285–296, Aug. 2019.

[28] B. Ibragimov, B. Likar, F. Pernus, and T. Vrtovec, "Accurate landmark-based segmentation by incorporating landmark misdetections," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 1072–1075

[29] Y. Shao, Y. Gao, Y. Guo, Y. Shi, X. Yang, and D. Shen, "Hierarchical lung field segmentation with joint shape and appearance sparse learning," *IEEE Trans. Med. Imag.*, vol. 33, no. 9, pp. 1761–1780, Sep. 2014.

[30] S. Candemir *et al.*, "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE Trans. Med. Imag.*, vol. 33, no. 2, pp. 577–590, Feb. 2014.

[31] P. Vock and Z. Szucs-Farkas, "Dual energy subtraction: Principles and clinical applications," *Eur. J. Radiol.*, vol. 72, no. 2, pp. 231–237, Nov. 2009.

[32] K. Martini, M. Baessler, S. Baumueller, and T. Frauenfelder, "Diagnostic accuracy and added value of dual-energy subtraction radiography compared to standard conventional radiography using computed tomography as standard of reference," *PLoS ONE*, vol. 12, no. 3, Mar. 2017, Art. no. e0174285.

[33] S. Albarqouni, J. Fotouhi, and N. Navab, "X-ray in-depth decomposition: Revealing the latent structures," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2017, pp. 444–452.

[34] W. Yang *et al.*, "Cascade of multi-scale convolutional neural networks for bone suppression of chest radiographs in gradient domain," *Med. Image Anal.*, vol. 35, pp. 421–433, Jan. 2017.

[35] N. Matsubara, A. Teramoto, K. Saito, and H. Fujita, "Bone suppression for chest X-ray image using a convolutional neural filter," *Australas Phys. Eng. Sci. Med.*, vol. 43, pp. 1–12, Nov. 2019.

[36] S. Chen and K. Suzuki, "Separation of bones from chest radiographs by means of anatomically specific multiple massive-training ANNs combined with total variation minimization smoothing," *IEEE Trans. Med. Imag.*, vol. 33, no. 2, pp. 246–257, Feb. 2014.

[37] M. Gusarev, R. Kuleev, A. Khan, A. Ramirez Rivera, and A. M. Khattak, "Deep learning models for bone suppression in chest radiographs," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Aug. 2017, pp. 1–7.

[38] D. Yul Oh and I. Dong Yun, "Learning bone suppression from dual energy chest X-rays using adversarial networks," 2018, *arXiv:1811.02628*. [Online]. Available: http://arxiv.org/abs/1811.02628

[39] M.-C. Huynh, T.-H. Nguyen, and M.-T. Tran, "Context learning for bone shadow exclusion in CheXNet accuracy improvement," in *Proc. 10th Int. Conf. Knowl. Syst. Eng. (KSE)*, Nov. 2018, pp. 135–140.

[40] I. M. Baltruschat *et al.*, "When does bone suppression and lung field segmentation improve chest X-ray disease classification?" in *Proc. IEEE 16th Int. Symp. Biomed. Imaging (ISBI)*, Apr. 2019, pp. 1362–1366.

[41] Y. Gordienko *et al.*, "Deep learning with lung segmentation and bone shadow exclusion techniques for chest X-ray analysis of lung cancer," in *Proc. Int. Conf. Theory Appl. Fuzzy Syst. Soft Comput.* Kiev, Ukraine: Springer, 2018, pp. 638–647.

[42] *Phillips Bone Suppression*. Accessed: Nov. 30, 2019. [Online]. Available: https://www.philips.com.au/healthcare/resources/landing/bone-supression

[43] *Clearread-Xray From Riveraintech*. Accessed: Nov. 30, 2019. [Online]. Available: https://www.riveraintech.com/clearread-xray/xray-bone-suppress/

[44] *Carestream Bone Suppression*. Accessed: Nov. 30, 2019. [Online]. Available: https://www.carestream.com/blog/2014/08/20/uncompromised-quality-bone-suppression-chest-x-ray-images/

[45] Y. Zhang and Q. Yang, "A survey on multi-task learning," 2017, *arXiv:1707.08114*. [Online]. Available: http://arxiv.org/abs/1707.08114

[46] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.

[47] V. Sze, Y.-H. Chen, J. Emer, A. Suleiman, and Z. Zhang, "Hardware for machine learning: Challenges and opportunities," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Apr. 2018, pp. 1–8.

[48] S. Kaji and S. Kida, "Overview of image-to-image translation by use of deep neural networks: Denoising, super-resolution, modality conversion, and reconstruction in medical imaging," *Radiol. Phys. Technol.*, vol. 12, no. 3, pp. 235–248, Sep. 2019.

[49] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1511–1520.

[50] J.-Y. Zhu *et al.*, "Toward multimodal image-to-image translation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 465–476.

[51] X. Qi, Q. Chen, J. Jia, and V. Koltun, "Semi-parametric image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8808–8816.

[52] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. 2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2337–2346.

[53] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[54] M. Eslami, C. Neuschaefer-Rube, and A. Serrurier, "Automatic vocal tract landmark localization from midsagittal MRI data," *Sci. Rep.*, vol. 10, no. 1, pp. 1–13, Dec. 2020.

[55] M. Eslami, C. Neuschaefer-Rube, and A. Serrurier, "Automatic vocal tract segmentation based on conditional generative adversarial neural network," in *Studientexte Zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*. Stockholm, Sweden: ESSV, 2019, pp. 263–270.

[56] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," 2018, *arXiv:1809.07294*. [Online]. Available: http://arxiv.org/abs/1809.07294

[57] X. Wang, H. Yan, C. Huo, J. Yu, and C. Pant, "Enhancing Pix2Pix for remote sensing image classification," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2332–2336.

[58] M. Sato, K. Hotta, A. Imanishi, M. Matsuda, and K. Terai, "Segmentation of cell membrane and nucleus by improving Pix2Pix," in *Proc. 11th Int. Joint Conf. Biomed. Eng. Syst. Technol.* Funchal, Portugal: SciTePress, 2018, pp. 216–220.

[59] V. V. Kniaz, "Conditional GANs for semantic segmentation of multispectral satellite images," *Proc. SPIE, Image Signal Process. Remote Sens. XXIV*, vol. 10789, Oct. 2018, Art. no. 107890R.

[60] J. Shiraishi *et al.*, "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," *Amer. J. Roentgenol.*, vol. 174, no. 1, pp. 71–74, Jan. 2000.

[61] *JSRT Dataset*. Accessed: Jan. 30, 2019. [Online]. Available: http://db.jsrt.or.jp/eng.php

[62] B. van Ginneken, M. B. Stegmann, and M. Loog, "Segmentation of anatomical structures in chest radiographs using supervised methods: A comparative study on a public database," *Med. Image Anal.*, vol. 10, no. 1, pp. 19–40, Feb. 2006.

[63] *JSRT Segmentation Dataset*. Accessed: Jan. 30, 2019. [Online]. Available: https://www.isi.uu.nl/Research/Databases/SCR/

[64] S. Juhász, Á. Horváth, L. Nikházy, G. Horváth, and Á. Horváth, "Segmentation of anatomical structures on chest radiographs," in *Proc. 12th Medit. Conf. Med. Biol. Eng. Comput.* Chalkidiki, Greece: Springer, 2010, pp. 359–362.

[65] *Bone Suppressed JSRT Dataset*. Accessed: Jan. 30, 2019. [Online]. Available: https://www.mit.bme.hu/eng/events/2013/04/18/bone-shadow-eliminated-images-jsrt-database and https://www.kaggle.com/hmchuong/xray-bone-shadow-supression

[66] K. Qu, X. Chai, T. Liu, Y. Zhang, B. Leng, and Z. Xiong, "Computer-aided diagnosis in chest radiography with deep multi-instance learning," in *Proc. Int. Conf. Neural Inf. Process.* Guangzhou, China: Springer, 2017, pp. 723–731.

[67] M. Ding *et al.*, "Local-global classifier fusion for screening chest radiographs," *Proc. SPIE, Med. Imag., Imag. Inf. Healthcare, Res., Appl.*, vol. 10138, Mar. 2017, Art. no. 101380A.

[68] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017, *arXiv:1712.04621*. [Online]. Available: http://arxiv.org/abs/1712.04621

[69] H. Shi, L. Wang, G. Ding, F. Yang, and X. Li, "Data augmentation with improved generative adversarial networks," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 73–78.

[70] *Multi-Atlas Labeling Beyond the Cranial Vault—Workshop and Challenge at MICCAI 2015*. Accessed: 2019-12-30. [Online]. Available: https://www.synapse.org/#!Synapse:syn3193805/wiki/217752

[71] H. Chen *et al.*, "Low-dose CT with a residual encoder-decoder convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 36, no. 12, pp. 2524–2535, Dec. 2017.

[72] H. J. Kuijf. *Grand Challenge on MR Brain Segmentation at MICCAI 2018*. Accessed: Dec. 30, 2019. [Online]. Available: https://mrbrains18.isi.uu.nl/

[73] *Pix2Pix Code Via Tensorflow*. Accessed: Jan. 20, 2019. [Online]. Available: https://phillipi.github.io/pix2pix/

[74] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Munich, Germany: Springer, 2015, pp. 234–241.

[75] T. Falk *et al.*, "U-Net: Deep learning for cell counting, detection, and morphometry," *Nature Methods*, vol. 16, no. 1, pp. 67–70, Jan. 2019.

[76] D. Seghers, D. Loeckx, F. Maes, D. Vandermeulen, and P. Suetens, "Minimal shape and intensity cost path segmentation," *IEEE Trans. Med. Imag.*, vol. 26, no. 8, pp. 1115–1129, Aug. 2007.

[77] P. Jagalingam and A. V. Hegde, "A review of quality metrics for fused image," *Aquatic Procedia*, vol. 4, pp. 133–142, Jan. 2015.

[78] D. Ardila *et al.*, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature Med.*, vol. 25, no. 6, pp. 954–961, Jun. 2019.

[79] H. Shan *et al.*, "Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction," *Nature Mach. Intell.*, vol. 1, no. 6, pp. 269–276, Jun. 2019.

[80] L. Xiang, Y. Li, W. Lin, Q. Wang, and D. Shen, "Unpaired deep cross-modality synthesis with fast training," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Granada, Spain: Springer, 2018, pp. 155–164.

[81] S. Roy, A. Carass, and J. L. Prince, "Magnetic resonance image example-based contrast synthesis," *IEEE Trans. Med. Imag.*, vol. 32, no. 12, pp. 2348–2363, Dec. 2013.

[82] Q. Yang, N. Li, Z. Zhao, X. Fan, E. I-Chao Chang, and Y. Xu, "MRI cross-modality NeuroImage-to-NeuroImage translation," 2018, *arXiv:1801.06940*. [Online]. Available: http://arxiv.org/abs/1801.06940

[83] Q. Yang, N. Li, Z. Zhao, X. Fan, E. I-Chao Chang, and Y. Xu, "MRI cross-modality NeuroImage-to-NeuroImage translation," 2018, *arXiv:1801.06940*. [Online]. Available: http://arxiv.org/abs/1801.06940

[84] Y. Huo, Z. Xu, S. Bao, A. Assad, R. G. Abramson, and B. A. Landman, "Adversarial synthesis learning enables segmentation without target modality ground truth," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 1217–1220.

[85] Y. Huang, L. Shao, and A. F. Frangi, "Simultaneous super-resolution and cross-modality synthesis in magnetic resonance imaging," in *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics*. Springer, 2019, pp. 437–457.

[86] Y. Zhang, P.-T. Yap, L. Qu, J.-Z. Cheng, and D. Shen, "Dual-domain convolutional neural networks for improving structural information in 3 T MRI," *Magn. Reson. Imag.*, vol. 64, pp. 90–100, Dec. 2019.