XraySyn: Realistic View Synthesis From a Single Radiograph Through CT Priors

Cheng Peng 1*  Haofu Liao 2*  Gina Wong 1  Jiebo Luo 2  Shaohua Kevin Zhou 3,4  Rama Chellappa 1

1 Johns Hopkins University  2 University of Rochester  3 Chinese Academy of Sciences  4 Peng Cheng Laboratory, Shenzhen

Abstract

A radiograph visualizes the internal anatomy of a patient through the use of X-ray, which projects 3D information onto a 2D plane. Hence, radiograph analysis naturally requires physicians to relate the prior about 3D human anatomy to 2D radiographs. Synthesizing novel radiographic views in a small range can assist physicians in interpreting anatomy more reliably; however, radiograph view synthesis is heavily ill-posed, lacking in paired data, and lacking in differentiable operations to leverage learning-based approaches. To address these problems, we use Computed Tomography (CT) for radiograph simulation and design a differentiable projection algorithm, which enables us to achieve geometrically consistent transformations between the radiography and CT domains. Our method, XraySyn, can synthesize novel views on real radiographs through a combination of realistic simulation and finetuning on real radiographs. To the best of our knowledge, this is the first work on radiograph view synthesis. We show that by gaining an understanding of radiography in 3D space, our method can be applied to radiograph bone extraction and suppression without groundtruth bone labels.

Introduction

Radiography, widely used for visualizing the internal human anatomy, applies high-energy radiation, or X-ray, to pass through the body, and measures the remaining radiation energy on a planar detector. Since different organs attenuate X-ray to various degrees, the detected energy is visualized as a 2D image or a radiograph, that reveals the internal structure of the body and provides valuable diagnostic information.

Radiography is fast and economical; however, high energy radiation can cause adverse health effects. Conventionally only a single, frontal view radiograph is acquired per session, e.g. for chest radiography. While physicians can intuitively relate the different organs on a 2D radiograph in 3D space, such intuition is implicit and varies in accuracy. As such, a radiograph view synthesis algorithm can help provide additional information to assist in understanding a patient’s internal structure. The ability to understand radiographs in a 3D context is also essential beyond providing more visual information. For example, since every pixel on a radiograph represents an X-ray traversing in 3D, it is principally ambiguous to label a pixel as a specific anatomical structure, as X-rays inevitably pass through multiple structures. By exploring 3D context, we can disentangle a pixel into values that represent different structures, and lead to improved analysis algorithms. Particularly, bone extraction on real radiographs is difficult, but has been used in fracture analysis (Chen et al. 2020), lesion detection (Li et al. 2020), etc. In this work, we...
tackle radiograph understanding in 3D, specifically through the tasks of novel view synthesis and bone extraction.

Transforming 2D images to 3D objects is by nature ill-posed. In the natural image domain, deep learning-based methods have shown impressive results in addressing such a problem in a data-driven way. Radiograph view synthesis, however, poses several unique challenges. Firstly, there is no multi-view dataset for radiographs due to privacy and radiation concerns, which prohibits the use of supervised learning. Secondly, there lacks a differentiable algorithm that ensures geometrically consistent transformations between radiographs and the 3D space. Lastly, unlike visible light, X-rays can penetrate objects, therefore to inverse the X-ray projection one should take into account both the surface and the internal information in 3D, making the problem even more ill-posed than for natural images and thus poses challenges to unsupervised methods.

While inverse graphics is a daunting task to be solved directly on real radiographs, some attempts have been made to address it through Digitally Reconstructed Radiographs (DRRs) (Ying et al. 2019; Henzler et al. 2018). DRRs (Moore et al. 2011) are simulated radiographs from Computed Tomography (CT) volumes, which are abundantly available. This approach addresses the data scarcity for learning a 2D-3D transformation; however, there remains significant differences between real radiographs and DRRs. The generation of DRRs also cannot be incorporated into a learning-based algorithm, and either is done on-off line (Henzler et al. 2018) or has severe limitations in available view angles (Ying et al. 2019).

In this work, we introduce a novel, two-stage algorithm called XraySyn to estimate the 3D context from a radiograph and use it for novel view synthesis and bone extraction. The first stage of XraySyn, called 3D PriorNet (3DPN), incorporates a pair of differentiable backprojection and forward projection operators to learn the radiograph-to-CT transformation under simulated setting. These operators ensure the transformations between radiograph and CT to be geometrically consistent, therefore significantly reducing the complexity of learning. We further incorporate the differentiable forward projector into a modified DeepDRR (Unberath et al. 2018), which simulates realistic radiographs, to minimize the domain gap between DRRs and real radiographs. The second stage of XraySyn, called 2D RefineNet (2DRN), further enhances the projected radiograph from its estimated 3D CT. By using a Generative Adversarial Network (GAN) and residual connections, 2DRN produces high quality radiographic views and their respective bone structure. In summary, our contributions can be described in four parts:

1. We propose a differentiable forward projection operator and incorporate it within a modified DeepDRR, forming a pipeline called CT2Xray that simulates realistic radiographs from CT, propagates gradients, and runs fast.

2. We propose a 3D PriorNet (3DPN), which incorporates CT2Xray and generates the 3D context from a single radiograph through learning from the paired relationship between simulated radiographs and their CT volumes.

3. We propose a 2D RefineNet (2DRN), which refines on the 2D radiographs projected from 3DPN’s output. By leveraging the availability of CT labels and the CT2Xray pipeline, 2D RefineNet can synthesize not only novel radiograph views, but also the corresponding bone structure.

4. We examine XraySyn, comprising 3DPN and 2DRN, on real radiographs and find the performance of view synthesis and bone extraction visually accurate, despite the lack of direct supervision in the radiograph domain.

Related work

View synthesis from a single image  There is a long history of research in natural image view synthesis. For relevancy and brevity, we focus on recent advances in view synthesis based on a single image and with the use of CNN. One approach to tackling such a task is to generate the new view in an image-to-image fashion. Some methods (Chen et al. 2016; Kulkarni et al. 2015) propose to generate a disentangled space where the image can be projected to and modified from to synthesize new views, while others (Park et al. 2017; Sun et al. 2018; Tatarchenko, Dosovitskiy, and Brox 2016; Zhou et al. 2016) rely on GANs to generate the information that is occluded from the original view. In general, the image-to-image approach is based on sufficient pixel correspondence between views, which provide understanding for recovery in either the image space or latent space. Such pixel correspondence is much weaker between X-ray views. In spirit, our method is more similar to the 3D shape generation approaches (Girdhar et al. 2016; Choy et al. 2016; Liu et al. 2019; Xu et al. 2019; Gkioxari, Johnson, and Malik 2019; Groueix et al. 2018) that concern the generation of 3D surfaces, which are less ill-posed than generating 3D volumes.

Radiograph simulation and transformation to CT  Due to the lack of multi-view radiographs and the difficulties in correctly labelling them, data-driven methods that require large number of radiographs often turn to CT-based radiograph simulations. While Monte-Carlo (MC) methods based on imaging physics (Badal and Badano 2009; Sisniega et al. 2013) can lead to highly realistic radiograph simulations, they are time-consuming and not scalable. Many works (Li et al. 2020; Gozes and Greenspan 2018; Ying et al. 2019; Albarqouni, Fotouhi, and Navab 2017; Campo, Pascau, and Estépar 2018) use DRRs, which are less realistic but computationally inexpensive radiograph simulations, to perform tasks such as bone enhancement, bone suppression, disease identification, CT reconstruction, etc. In particular, Ying et al. 2019 addresses the discrepancy between DRRs and real radiographs by training an additional domain adaptation network. Henzler et al. 2018 uses real cranial X-ray images acquired in a controlled setting to recover the 3D bone structure. As clinical evidence supports that bone suppression on radiographs can improve diagnostic accuracy (Laskey 1996), Li et al. 2020 proposes to achieve bone suppression by learning a bone segmentation network based on DRRs, and apply the network on real radiographs with handcrafted post-processes. Recently, DeepDRR (Unberath et al. 2018) is proposed to model DRR generation more accurately by replicating similar
procedures from the MC simulation counterpart, and shows that CNN models trained on such simulations are able to generalize better on real radiographs.

Method

The main goal of this work is to synthesize novel views from a frontal view radiograph, which requires a degree of 3D knowledge. As multi-view dataset is not readily available for real radiographs, our proposed method, XraySyn, composes of two stages. The first stage learns to estimate 3D knowledge under a simulated setting by using CT volumes. The second stage transfers such learning to generate real radiographs.

The challenge underlying this approach is how to best address the transformation from simulated radiographs to CT volumes, while ensuring the input radiograph can be reproduced from such volume. We first explain the proposed operators that enable learning such a transformation, CT2Xray and Single Image Backprojector. We then introduce XraySyn, which incorporates the two operators for a radiograph-to-CT-to-radiograph algorithm.

CT2Xray

CT2Xray, as shown in Fig. 2, has two parts: (i) the differentiable forward projection; (ii) attenuation-based radiograph simulation, which along with the first part forms CT2Xray and transforms a CT volume into a realistic radiograph. To the best of our knowledge, CT2Xray is the first algorithm that can generate realistic radiographs from CT volumes with gradient propagation along arbitrary viewpoint.

Differentiable forward projector (FP). Let $V_{CT}$ denote a CT image, FP generates a 2D projection of $V_{CT}$ by:

$$I(x) = FP(V_{CT}, T) = \int V_{CT}(T^{-1}p)dl(x) \approx \sum_{p \in I(x)} V_{CT}(T^{-1}p)\Delta p,$$

where $T$ is a homogeneous matrix that controls the rotation and translation of the view point and $I(x)$ is a line segment connecting the simulated x-ray source and detector at $x$. For backpropagation, the gradients of $I$ wrt $V_{CT}$ can be written as

$$\frac{\partial I(x)}{\partial V_{CT}(y)} = \begin{cases} \Delta p, & \text{if } Ty \in I(x), \\ 0, & \text{otherwise}. \end{cases}$$

Equations for (1) and (2) can be implemented through massive parallelism with GPUs, where every line integral over the volume is a standalone operation; as such this implementation can be used in online training.

CT2Xray. Forward projecting CT volumes generates DRRs, which are poor simulations of X-ray images due to the inaccurate assumption that different tissues attenuate the X-ray similarly. DeepDRR [Unberath et al. 2018] produces a better simulation by avoiding this assumption, but it is not differentiable and hence not amenable for end-to-end learning. We contribute a differentiable X-ray simulation pipeline, called CT2Xray, which incorporates the differentiable forward projector and produces more realistic radiographs. In [Unberath et al. 2018], a realistic X-ray attenuation $I_{atten}$ is modeled as

$$I_{atten} = \sum_{E} I_0 e^{-\sum_{m} \mu(m,E)\epsilon_{m} + SE + noise},$$

where $\mu(m,E)$ is the linear attenuation coefficient of material $m$ at energy state $E$ and is measured and known [Hubbell and Seltzer 1995], $\epsilon_{m}$ is the material thickness. SE is the scatter estimation term, $I_0$ is the source X-ray intensity. Along with noise, SE and $I_0$ are omitted for simplicity.

CT2Xray considers only the bone and tissue materials, that is, $m \in \{\text{bone, tissue}\}$. With the aid of a bone mask $V_{mask}$ and using the differentiable forward projector, it calculates the material thickness wrt the projection parameter $T$ as

$$\epsilon_{\text{bone}}(T) = FP(V_{CT} \odot V_{mask}, T),$$

$$\epsilon_{\text{tissue}}(T) = FP(V_{CT} \odot (1 - V_{mask}), T).$$

Radiographs are typically stored and viewed as inverted versions of the measured attenuation; therefore, CT2Xray can be expressed as

$$CT2Xray(V_{CT}, V_{mask}, T) = \max(I_{atten}^{T} - I_{atten}^{0}),$$

with

$$I_{atten}^{T} = \sum_{E} e^{-\sum_{m} \mu(m,E)\epsilon_{m}^{T}(T)}.$$

Single image backprojector (BP). While CT2Xray generates radiographs from CT volumes, an inverse function is needed to transform radiographs back to the respective CT volumes. Such a transformation is clearly ill-posed; however, we can formulate an inverse function of the forward projector to properly place the input X-ray image in 3D. We call such an inverse function the single image backprojector. Following a similar formulation from general backprojection algorithm, which reconstructs CBCT from multi-view radiographs as described in [Kinahan, Defrise, and Clackdoyle 2004], the single image backprojector (BP) is expressed as the following:

$$V_{BP}(y) = BP(I(x), T) = \begin{cases} \frac{I(x)}{|Ty|\Delta p}, & \text{if } Ty \in I(x), \\ 0, & \text{otherwise}. \end{cases}$$

By substituting $V_{BP}$ in (7) into $V_{CT}$ in (1), the same $I$ is recovered under view $T$, hence we denote $V_{BP}$ obtained in this way as the backprojection of $I$ at view $T$. While the same image-wise consistency does not generally apply to CT2Xray, i.e., substituting (7) into (5) does not recover $I$. We show that by using a CNN to complement the single image backprojector, such consistency can be better approximated due to geometric consistency.

XraySyn

XraySyn is trained in two stages, as shown in Fig. 2. The simulation stage trains a radiograph-to-CT transformation network, called 3D PriorNet, and with help of CT2Xray generate radiographic views from the estimated CT. The real-radiograph stage, which trains a 2D RefineNet, then further closes the domain gap between simulated and real radiographs. Due to the need for calculating the material-dependent attenuation, we also gain the ability to transfer
where $CE$ refers to the cross entropy loss and $\lambda$ is defined as:

$$L = \lambda_{CT} L_{CT} + \lambda_{mask} L_{mask} + \lambda_{xray}(L_{xray}^T + L_{xray}^T),$$

where $\lambda_{CT}$, $\lambda_{mask}$, and $\lambda_{xray}$ are weights; $L_{CT}$, $L_{mask}$, and $L_{xray}$ are defined as:

$$L_{CT} = \|V_{CT} - V_{gt}\|_1,$$
$$L_{mask} = CE(V_{mask}, V_{gt}),$$
$$L_{xray}^T = ||I_{CT}^T - I_{ref}^T||_1 + \sum_m ||I_{CT}^m(T) - I_{ref}^m(T)||_1,$$

where $CE$ refers to the cross entropy loss and $L_{xray}^T$ is defined similarly as $L_{CT}$.

$\lambda_{CT}$ is defined as:

$$L_{CT} = \|V_{CT} - V_{mask}\|_1$$

and $\lambda_{xray}$ is defined as:

$$L_{xray}^T = ||I_{CT}^T - I_{ref}^T||_1 + \sum_m ||I_{CT}^m(T) - I_{ref}^m(T)||_1.$$
GAN) is used to ensure $I_{\text{final}}'$ is statistically similar to $I_{\text{in}}^T$ when $T'$ and $T$ are relatively close. The overall loss function for 2DRN is described as:

$$\mathcal{L}_{\text{2DRN}} = \lambda_{\text{recon}}\|I_{\text{final}}' - I_{\text{in}}^T\|_1 + \lambda_{\text{GAN}}\mathcal{L}_{\text{LSGAN}}(I_{\text{final}}', I_{\text{in}}^T).$$

(13)

\(\mathcal{L}_{\text{LSGAN}}\) is defined as:

$$\mathcal{L}_{\text{LSGAN}}(I_{\text{final}}', I_{\text{in}}) = \mathbb{E}(D(I_{\text{in}} - 1)^2) + \mathbb{E}(D(I_{\text{in}}'))^2,$$

$$\mathcal{L}_{\text{LSGAN}}(I_{\text{final}}) = \mathbb{E}(D(I_{\text{final}} - 1)^2),$$

(14)

where \(G\) indicates generator, in this case the composition of \(G, F, \) and \(M, D\) indicates the discriminator.

**Experiments**

**Implementation details** The two stages of XraySyn are trained separately. To train 3DPN, a CT volume \(V_{\text{gt}}\) and its bone mask \(V_{\text{mask}}^{\text{gt}}\) is used to simulate the groundtruth radiographs \(I_{\text{in}}^T\) and \(I_{\text{in}}^{T'}\). \(T\) and \(T'\) are sampled randomly from -18° to 18° in azimuth and elevation angles. During the training of 2DRN, real radiographs are used as \(I_{\text{in}}^T\) in place of simulated radiographs, and the 3DPN’s parameters are frozen. Furthermore, the input is first downsampled through average pooling as real radiographs are of higher resolution. As view angles are not available for real radiographs, we sample \(T\) and \(T'\) randomly in similar fashion as for 3DPN training. For testing on real radiograph, \(T\) is assumed to be the canonical frontal view, \(T'\) are twenty view angles uniformly spaced from -9° to 9° in azimuth. Due to the discretization of the voxel-based representation, the ray tracing process used in forward and backprojection needs to approximate points in space when those points are not on the coordinate grid. We use trilinear interpolation for such approximation. The networks are implemented with Pytorch, and trained using four Nvidia P6000 GPUs for five days. The details of network structure are reported in the supplemental material.

**Dataset** Both CT and radiograph datasets are needed for training XraySyn. To train the 3DPN, we use the LIDC-IDRI dataset (Armato III et al. 2011), which contains 1,018 chest CT volumes. We discard all volumes that have a between-slices resolution higher than 2.5mm. This leads to 780 CT volumes, from which we randomly select 700 for training, 10 for evaluation, and 70 for testing. To preprocess the data, we tightly crop the CT volumes to eliminate excess empty space, and reshape the volumes into resolution of 128 \(\times\) 128 \(\times\) 128. To train 2DRN with real radiographs, we use the TBX11K dataset (Liu et al. 2020). Specifically, to avoid the excessive interference of foreign objects and abnormal anatomy, we manually select 3000 images under the healthy category within TBX11K and crop those images to have similar field-of-view as the CT simulation. The images are then reshaped to resolution of 256 \(\times\) 256. We randomly select 2600 for training, 100 for evaluation, and 300 for testing.

**Evaluation metrics** Evaluation of novel view synthesis on real radiograph can be challenging, as there is no groundtruth. Therefore, we first provide the results of evaluation on simulated radiographs. Peak Signal-to-Noise Ratio (PSNR) and Structured Similarity Index (SSIM) (Wang et al. 2004) are used to measure the quality of synthesized novel views and material decomposition. For real radiographs, we use PSNR and SSIM to measure the quality of reconstruction for the input view \(T\) and Fréchet Inception Distance (FID) (Heusel et al. 2017) to measure the realism of the novel view radiograph in comparison to the input view radiograph. Since radiographs are grayscale images, we do the following to adapt to the ImageNet-trained InceptionV3 model: (i) we repeat the grayscale values across the RGB channels, and (ii) use the middle layer features from InceptionV3, specifically the 768-channel layer before InceptionV3’s last auxiliary classifier, for better generalizability.

**Ablation study** We evaluate the effectiveness of XraySyn against alternative 2D or 3D methods. Firstly, comparisons against 3DPN are made under simulated setting for the following setups:

- **2D Refiner:** An image-to-image method by synthesizing new views from input FP(BP(I_{\text{in}}^T, T), T') through a 2D DenseNet structure. The training is constrained by a L1 loss between the generated X-ray and its groundtruth.

- **X2CT:** Proposed by Ying et al. (Ying et al. 2019) to transform DRRs into CT volumes. We made the following adaptations: (i) instead of DRRs, the inputs are X-rays generated by CT2Xray, and (ii) the training losses are consistent with Eq. (9).

- **3DPN-DRR:** An alternative 3DPN that directly uses forward projection to simulate radiographs, i.e. DRRs. Note that no material decomposition is involved in DRR.

Comparisons against the overall XraySyn method are then made for generating real radiographs. These include:

- **3DPN:** A direct use of 3DPN trained on simulated data.

- **XraySyn_{noM}:** An alternative XraySyn where the 2DRN stage does not have \(M\).

Table 1 summarizes the performance of different implementations over various view angles, and visualization of the results are provided in Fig. 3. Image-to-image approach like 2D Refiner synthesizes mostly a blurry version of the input view. The lack of an explicit 3D loss forces the network to overly smooth the output for the best PSNR. The implementation of X2CT-CNN copies the 2D features along the third axis to achieve the initial upsampling from 2D to 3D, which is geometrically incorrect for arbitrary view angle. While CNN’s strong learning ability still helps X2CT-CNN produce a coarse 3D estimation, it has difficulty in learning the geometric transformation that maps the input image to the correct view. Consequently, the synthesized views are reliant on a poorly estimated 3D structure and lack significant details, specifically over the rib cage area. By using the geometrically consistent backprojection and forward projection, 3DPN-DRR and 3DPN perform much better at preserving the input radiograph during the 2D-to-3D transformation. The
| Method          | View | \(I_{CT}^r\)   | \(t_{bone}^r\) | \(t_{tissue}^r\) | \(I_{CT}^r\)   | \(t_{bone}^r\) | \(t_{tissue}^r\) | Method          | View | \(t_{final}^r\) | \(t_{final}^r\) (FID) |
|----------------|------|----------------|----------------|-----------------|----------------|----------------|----------------|----------------|------|----------------|-----------------|
| 2D Refiner     |      | 20.44/0.755    | 15.99/0.451    | 19.57/0.711     | 20.00/0.750    | 16.12/0.450    | 19.38/0.703    | XraySyn\textsuperscript{noM} |      | 28.42/0.857    | 0.375            |
| X2CT           |      | 21.05/0.925    | 19.79/0.867    | 21.63/0.780     | 24.27/0.860    |                 |               | XraySyn         |      | 30.3/0.865     | 0.319            |
| 3DPN-DRR       |      | 29.49/0.961    | 22.30/0.814    | 24.40/0.866     | 27.22/0.929    | 22.75/0.819    | 0.375          |                 |      |                |                 |
| 3DPN           |      | 22.44/0.817    | 20.07/0.850    |                 | 24.40/0.866    | 21.52/0.741    | 0.375          |                 |      |                |                 |
|                |      | 19.57/0.711    | 16.12/0.450    |                 | 19.38/0.703    | 19.79/0.867    |               |                 |      |                |                 |
| 3DPN-DRR       |      | 27.22/0.929    | 22.75/0.819    | 0.375           |                 |                 |               |                 |      |                |                 |
| 3DPN           |      | 28.42/0.857    | 0.375          |                 |                 |                 |               |                 |      |                |                 |
|                |      | 20.07/0.850    | 21.52/0.741    |                 |                 |                 |               |                 |      |                |                 |

Table 1: Ablation study of our proposed methods against alternative implementations. Note that 3DPN-DRR and 2D Refiner do not use CT2Xray, therefore are without metrics for \(t_{bone}\), \(t_{tissue}\). \(I_{CT}^r\) for 2D Refiner is trivially generated as \(I_{in}^r\). PSNR/SSIM metrics are provided when groundtruth is available, otherwise FID score is reported. The best performing metrics are **bold**, the second best is **underlined**.

Figure 3: Visual comparisons of novel view radiographs generated by different methods based on simulated radiographs as inputs. View angle change is measured azimuthally. For each non-groundtruth image, PSNR/SSIM are provided as metrics. Note that the input view for 2D Refiner is trivially generated as \(I_{in}^r\). Methods that involve the use of CT2Xray are displayed with both CT and 2D Refiner views are much improved in both low-level details and overall realism, along with the corresponding bone map \(t_{bone}\). How-

main issue of DRR is its assumption that the rays attenuate over bone and tissue similarly, when in fact bone attenuates X-Rays much better and therefore has better contrast on real radiographs. The results show that when 3DPN-DRR is used on more realistically simulated radiographs, bone appears much softer. While this impacts PSNR significantly, 3DPN-DRR results have much better SSIM scores compared to X2CT and 2D Refiner due to the superior 3D estimation. Finally, 3DPN, which includes all proposed components, performs much better than other methods (6-7dB more than the second best method in terms of PSNR) at capturing the 3D anatomy from a single radiograph. As 2D-to-3D transformation is still ill-posed, the metrics worsen when the novel view is further from the input view. Please see the supplementary material section for more detailed metrics.

Using 3DPN to perform view synthesis on real radiographs involves additional challenges. Specifically, 3DPN is limited in the available resolution and cannot generate sufficient details, e.g. as shown in Fig. 4, 3DPN-generated bone structure is less visible on the rib cage area, as those bones are small and hard to be accurately estimated in 3D. Furthermore, simulation does not address complex imaging conditions. As a result, views synthesized through 3DPN are blurry and constrained to have a certain image style due to the formulation of CT2Xray. 2DRN is proposed to address these problems by performing refinement on \(t_{CT}^r(T')\) to preserve the material decomposition information.

After refining the 3DPN results in XraySyn\textsuperscript{noM}, the novel views are much improved in both low-level details and overall realism, along with the corresponding bone map \(t_{refbone}\). However, some information is inevitably lost during 3DPN. The input in Fig. 4 is grayscale in the lung area, as opposed to the typical dark color from simulation. This information is not captured in 3DPN, thus the subsequent refinement network is also limited in its recovery performance. \(M\) is designed to shuffle information directly from the input radiograph. To prevent 2DRN from learning an identity transform of the input radiograph, \(M\) generates convolutional filter parameters to constrict the information flow. The complete XraySyn, which includes \(M\) in 2DRN, improves the overall performance by almost 1dB with little additional computational cost. While 2DRN does not explicitly guarantee overall 3D consistency,
Figure 4: Visual comparisons of novel view radiographs generated by different methods based on real radiographs as inputs. View angle change is measured azimuthally. PSNR/SSIM metrics are provided for input radiograph reconstruction, FID is provided for novel view synthesis with respect to input radiographs. The best metrics are **bold**. For more results with animation, please refer to the supplemental material.

Figure 5: Bone-suppressed radiographs obtained by post-processing the results from XraySyn. (top) Input radiographs from TBX11K [Liu et al. 2020] and (bottom) the bone-suppressed results. Note that the arteries, as indicated with blue arrows, are more visible after bone suppression.

Bone suppression A natural application of XraySyn is on radiograph bone suppression, which seeks to reduce bone attenuation and better reveal the underlying tissues. To best preserve the tissue information from input radiograph $I_{in}^T$, we find $t_{bone}^{ref}$ and $t_{tissue}^{recon}$ so that they losslessly reconstruct $I_{in}^T$, through reversing the CT2Xray operation as shown in Eq. [6]. In Fig. [5] we show that our approach suppresses most of the rib cage bones while preserving the tissue content. For a mathematical formulation on reversing CT2Xray, please refer to the supplemental material section.

**Conclusion**

We propose a two-stage radiograph view synthesis method, XraySyn. This method estimates a coarse 3D CT from a 2D radiograph, simulates a novel view from the estimated volume, and finally refines the views to be visually consistent with real radiographs. The learning process of XraySyn is enabled by our proposed differentiable forward projector and backprojector. Furthermore, by incorporating the CT bone labels in CT2Xray that is inspired by DeepDRR and implemented with our differentiable forward projector, we not only achieve realistic simulation for training the radiograph-to-CT transformation, but also gain the ability to transfer bone labels from CT to radiograph. We carefully evaluate our method both on simulated and real radiographs, and find that XraySyn generates highly realistic and consistent novel view radiographs. To the best of our knowledge, this is the first work on radiograph view synthesis, which can help give practitioners a more precise understanding of the patient’s 3D anatomy. XraySyn also opens up possibilities for many downstream processes on radiograph, such as lesion detection, organ segmentation, and sparse-view CT reconstruction. Currently, XraySyn is limited in resolution due to memory constraint of 3DPN. We plan to address this limitation through a more
efficient network design. In addition, we will conduct a study to assess the effect of bone suppression for clinical diagnosis.

References

Albarqouni, S.; Fotouhi, J.; and Navab, N. 2017. X-Ray In-Depth Decomposition: Revealing the Latent Structures. In Descoteaux, M.; Maier-Hein, L.; Franz, A. M.; Jannin, P.; Collins, D. L.; and Duchesne, S., eds., Medical Image Computing and Computer Assisted Intervention - MICCAI 2017 - 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017. Proceedings. Part III, volume 10435 of Lecture Notes in Computer Science, 444–452. Springer. doi:10.1007/978-3-319-66179-7_51. URL https://doi.org/10.1007/978-3-319-66179-7_51

Armato III, S. G.; McLennan, G.; Bidaut, L.; McNitt-Gray, M. F.; Meyer, C. R.; Reeves, A. P.; Zhao, B.; Aberle, D. R.; Henschke, C. I.; Hoffman, E. A.; et al. 2011. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Medical physics 38(2): 915–931.

Badal, A.; and Badano, A. 2009. Accelerating Monte Carlo simulations of photon transport in a voxelized geometry using a massively parallel graphics processing unit. Medical physics 36(11): 4878–4880.

Campo, M. I.; Pascau, J.; and Estépar, R. S. J. 2018. Emphysema quantification on simulated X-rays through deep learning techniques. In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 273–276.

Chen, H.; Wang, Y.; Zheng, K.; Li, W.; Cheng, C.; Harrison, A. P.; Xiao, J.; Hager, G. D.; Lu, L.; Liao, C.; and Miao, S. 2020. Anatomy-Aware Siamese Network: Exploiting Semantic Asymmetry for Accurate Pelvic Fracture Detection in X-ray Images. CoRR abs/2007.01464. URL https://arxiv.org/abs/2007.01464

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In Lee, D. D.; Sugiyama, M.; von Luxburg, U.; Guyon, I.; and Garnett, R., eds., Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, 2172–2180. URL http://papers.nips.cc/paper/6399-infogan-interpretable-representation-learning-by-information-maximizing-generative-adversarial-nets

Choy, C. B.; Xu, D.; Gwak, J.; Chen, K.; and Savarese, S. 2016. 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII, volume 9912 of Lecture Notes in Computer Science, 628–644. Springer. doi:10.1007/978-3-319-46484-8_38. URL https://doi.org/10.1007/978-3-319-46484-8_38

Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; and Ronneberger, O. 2016. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In Ourselin, S.; Joskowicz, L.; Sabuncu, M. R.; Unal, G. B.; and Wells, W., eds., Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016 - 19th International Conference, Athens, Greece, October 17-21, 2016. Proceedings, Part II, volume 9901 of Lecture Notes in Computer Science, 424–432. doi:10.1007/978-3-319-46723-8_49. URL https://doi.org/10.1007/978-3-319-46723-8_49

Girdhar, R.; Fouhey, D. F.; Rodriguez, M.; and Gupta, A. 2016. Learning a Predictable and Generative Vector Representation for Objects. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016. Proceedings, Part VI, volume 9910 of Lecture Notes in Computer Science, 484–499. Springer. doi:10.1007/978-3-319-46466-4_29. URL https://doi.org/10.1007/978-3-319-46466-4_29

Gkioxari, G.; Johnson, J.; and Malik, J. 2019. Mesh R-CNN. In 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, 9784–9794. IEEE. doi:10.1109/ICCV.2019.00988. URL https://doi.org/10.1109/ICCV.2019.00988

Gozes, O.; and Greenspan, H. 2018. Lung Structures Enhancement in Chest Radiographs via CT Based FCNN Training. In Stoyanov, D.; Taylor, Z.; Kainz, B.; Maicas, G.; Beichel, R. R.; Martel, A. L.; Maier-Hein, L.; Bhatia, K. K.; Vercauteren, T.; Oktay, O.; Carneiro, G.; Bradley, A. P.; Nascimento, J. C.; Min, H.; Brown, M. S.; Jacobs, C.; Lassen-Schmidt, B.; Mori, K.; Petersen, J.; Estépar, R. S. J.; Schmidt-Richberg, A.; and Veiga, C., eds., Image Analysis for Moving Organ, Breast, and Thoracic Images - Third International Workshop, RAMBO 2018, Fourth International Workshop, BIA 2018, and First International Workshop, TIA 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16 and 20, 2018, Proceedings, volume 11040 of Lecture Notes in Computer Science, 147–158. Springer. doi:10.1007/978-3-030-00946-4_16. URL https://doi.org/10.1007/978-3-030-00946-4_16

Groueix, T.; Fisher, M.; Kim, V. G.; Russell, B. C.; and Aubry, M. 2018. AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. CoRR abs/1802.05384. URL http://arxiv.org/abs/1802.05384

Henzler, P.; Rasche, V.; Ropinski, T.; and Ritschel, T. 2018. Single-image Tomography: 3D Volumes from 2D Cranial X-Rays. Comput. Graph. Forum 37(2): 377–388. doi:10.1111/cgf.13369. URL https://doi.org/10.1111/cgf.13369

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two-Time-Scale Update Rule Converge to a Local Nash Equilibrium. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, 6626–6637. URL http://papers.nips.cc/paper/7240-gans-trained-by-a-two-time-scale-update-rule-converge-to-a-local-nash-equilibrium

Hubbell, J. H.; and Seltzer, S. M. 1995. Tables of X-ray mass attenuation coefficients and mass energy-absorption coefficients 1 keV to 20 MeV for elements Z= 1 to 48 and additional substances of dosimetric interest. Technical report, National Inst. of Standards and Technology-PL, Gaithersburg, MD (United States). 20 - Analytic Image Reconstruction Methods. In Wernick, M. N.; and Aarsvold, J. N., eds., Emission Tomography: 3D Volumes from 2D Cranial X-Rays. Academic Press. ISBN 978-0-12-744482-6. doi:https://doi.org/10.1016/B978-0-12-744482-6.50023-5. URL http://www.sciencedirect.com/science/article/pii/B9780127444826200235

Kulkarni, T. D.; Whitney, W. F.; Kohli, P.; and Tenenbaum, J. B. 2015. Deep Convolutional Inverse Graphics Network. In Cortes, C.; Lawrence, N. D.; Lee, D. D.; Sugiyama, M.; and Garnett, R., eds., Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, 2539–2547. URL http://papers.nips.cc/paper/5851-deep-convolutional-inverse-graphics-network

Laskey, M. A. 1996. Dual-energy X-ray absorptiometry and body composition. Nutrition 12(1): 45–51.
Network architecture

The proposed XraySyn consists of 3D PriorNet (3DPN) and 2D RefineNet (2DRN). The general model architectures of 3DPN and 2DRN follow 3D-UNet (Cicic et al. 2016) and RDN (Zhang et al. 2018). For 3DPN, the exact network structure is presented in Table 3. For 2DRN, we use RDN with sixteen Residual Dense Blocks (RDBs), eight convolutional layers per RDB, and growth rate of sixty-four. After the Global Feature Fusion (GFF) layer, an additional convolutional layer with kernel size of one is introduced where the filter parameters are generated by a subnetwork $M$. The upsampling layers then upsample spatial dimensions from $128 \times 128$ to $256 \times 256$, the output channel dimension is two, for bone and tissue. The exact network structure of $M$ is presented in Table 2 and is a fully convolutional network. As the spatial dimension decreases four times at every convolutional layer with a stride of two, $M$ is computationally efficient both in memory and speed. In the network tables, $N_c$ denotes the number of output channels, $C'$ denotes the channel dimension of generated features $F^{LR}$. We use ‘K#-C#-S#-P#’ to denote the configuration of the convolution layers, where ‘K’, ‘C’, ‘S’ and ‘P’ stand for the kernel, input channel, stride and padding size, respectively.

| Name   | $N_c$ | Description                        |
|--------|-------|------------------------------------|
| INPUT  | 1     | Input $V_{BP}$                     |
| CONV0  | 96    | K3-C1-S2-P1                        |
| LEAKY RELU | 192 | K3-C96-S2-P1, Negative_slope=0.2 |
| CONV1  | 384   | K3-C192-S2-P1                      |
| BN     |       | BatchNorm3D                        |
| LEAKY RELU | 768 | K3-C384-S2-P2, Negative_slope=0.2 |
| CONV2  | 768   | K3-C768-S2-P2                      |
| BN     |       | BatchNorm3D                        |
| LEAKY RELU | 768 | K3-C768-S2-P2, Negative_slope=0.2 |
| CONV3  | 768   | K3-C768-S2-P2                      |
| BN     |       | BatchNorm3D                        |
| LEAKY RELU | 768 | K3-C768-S2-P2, Negative_slope=0.2 |
| CONV4  | 768   | K3-C768-S2-P2                      |
| BN     |       | BatchNorm3D                        |
| LEAKY RELU | 768 | K3-C768-S2-P2, Negative_slope=0.2 |
| CONV5  | 768   | K3-C768-S2-P2                      |
| BN     |       | BatchNorm3D                        |
| LEAKY RELU | 768 | K3-C768-S2-P2, Negative_slope=0.2 |
| CONV6  | 768   | K3-C768-S2-P2                      |
| BN     |       | BatchNorm3D                        |
| LEAKY RELU | 768 | K3-C768-S2-P2, Negative_slope=0.2 |
| CONV7  | 768   | K3-C1536-S1-P1                     |
| BN     |       | BatchNorm3D                        |
| RELU   | 32    | K3-C1-S2-P1                        |
| RELU   | 64    | K3-C64-S2-P1                       |
| RELU   | 128   | K3-C64-S2-P1                       |
| RELU   | 256   | K3-C128-S2-P1                      |
| RELU   | 512   | K3-C256-S2-P1                      |
| RELU   | 1024  | K3-C512-S2-P1                      |
| RELU   | 2048  | K3-C1024-S2-P1                     |
| RELU   | 4096  | K3-C2048-S2-P1                     |
| CONV0  | 96    | K3-C1-S2-P1                        |
| RELU   | 32    | K3-C1-S2-P1                        |
| RELU   | 64    | K3-C64-S2-P1                       |
| RELU   | 128   | K3-C64-S2-P1                       |
| RELU   | 256   | K3-C128-S2-P1                      |
| RELU   | 512   | K3-C256-S2-P1                      |
| RELU   | 1024  | K3-C512-S2-P1                      |
| RELU   | 2048  | K3-C1024-S2-P1                     |
| RELU   | 4096  | K3-C2048-S2-P1                     |

Table 2: Network architecture of $M$ in 2DRN.

2D Refiner as described in Ablation study also uses the RDN structure for refinement on FP (BP$(I_{in}^{T}, T)$, $T'$), with twenty Residual Dense Blocks (RDBs), eight convolutional layers per RDB, and growth rate of sixty-four. The input for 2D Refiner is a concatenation of FP (BP$(I_{in}^{T}, T)$, $T'$) and $I_{in}$ to ensure both the original information and the geometric transformation are included.

CT2Xray

A more detailed visualization of CT2Xray is shown in Fig. 6. Given the CT volume and its bone label $\{V_{CT}, V_{bone}\}$, we can calculate for $\{V_{bone}, V_{tissue}\}$ and obtain $I_{CT}^{bone}(T)$, $I_{tissue}^{CT}(T)$ through our proposed differential forward projection. The corresponding attenuation is then calculated and summed to produce a realistic radiograph. Compared to the original DeepDRR (Unberath et al. 2018), CT2Xray discarded calculation for air-specific attenuation (air volume is considered as tissue), scatter estimation, and noise. We find that air-specific attenuation calculation as followed by the measurements in (Hubbell and Seltzer 1995) leads to a reduction in contrast on our CT dataset, and that the original CNN-based scatter estimation in DeepDRR is not robust enough on unseen data.

| Name   | $N_c$ | Description                        |
|--------|-------|------------------------------------|
| INPUT  | 1     | Input $V_{BP}$                     |
| CONV0  | 96    | K3-C1-S2-P1                        |
| LEAKY RELU | 192 | K3-C96-S2-P1, Negative_slope=0.2 |
| CONV1  | 384   | K3-C192-S2-P1                      |
| BN     |       | BatchNorm3D                        |
| LEAKY RELU | 768 | K3-C384-S2-P2, Negative_slope=0.2 |
| CONV2  | 768   | K3-C768-S2-P2                      |
| BN     |       | BatchNorm3D                        |
| LEAKY RELU | 768 | K3-C768-S2-P2, Negative_slope=0.2 |
| CONV3  | 768   | K3-C768-S2-P2                      |
| BN     |       | BatchNorm3D                        |
| LEAKY RELU | 768 | K3-C768-S2-P2, Negative_slope=0.2 |
| CONV4  | 768   | K3-C768-S2-P2                      |
| BN     |       | BatchNorm3D                        |
| LEAKY RELU | 768 | K3-C768-S2-P2, Negative_slope=0.2 |
| CONV5  | 768   | K3-C768-S2-P2                      |
| BN     |       | BatchNorm3D                        |
| LEAKY RELU | 768 | K3-C768-S2-P2, Negative_slope=0.2 |
| CONV6  | 768   | K3-C768-S2-P2                      |
| BN     |       | BatchNorm3D                        |
| LEAKY RELU | 768 | K3-C768-S2-P2, Negative_slope=0.2 |
| CONV7  | 768   | K3-C1536-S1-P1                     |
| BN     |       | BatchNorm3D                        |
| RELU   | 32    | K3-C1-S2-P1                        |
| RELU   | 64    | K3-C64-S2-P1                       |
| RELU   | 128   | K3-C64-S2-P1                       |
| RELU   | 256   | K3-C128-S2-P1                      |
| RELU   | 512   | K3-C256-S2-P1                      |
| RELU   | 1024  | K3-C512-S2-P1                      |
| RELU   | 2048  | K3-C1024-S2-P1                     |
| RELU   | 4096  | K3-C2048-S2-P1                     |
| CONV0  | 96    | K3-C1-S2-P1                        |
| RELU   | 32    | K3-C1-S2-P1                        |
| RELU   | 64    | K3-C64-S2-P1                       |
| RELU   | 128   | K3-C64-S2-P1                       |
| RELU   | 256   | K3-C128-S2-P1                      |
| RELU   | 512   | K3-C256-S2-P1                      |
| RELU   | 1024  | K3-C512-S2-P1                      |
| RELU   | 2048  | K3-C1024-S2-P1                     |
| RELU   | 4096  | K3-C2048-S2-P1                     |
| CONV0  | 96    | K3-C1-S2-P1                        |
| RELU   | 32    | K3-C1-S2-P1                        |
| RELU   | 64    | K3-C64-S2-P1                       |
| RELU   | 128   | K3-C64-S2-P1                       |
| RELU   | 256   | K3-C128-S2-P1                      |
| RELU   | 512   | K3-C256-S2-P1                      |
| RELU   | 1024  | K3-C512-S2-P1                      |
| RELU   | 2048  | K3-C1024-S2-P1                     |
| RELU   | 4096  | K3-C2048-S2-P1                     |

Table 3: Network architecture of 3DPN. BN stands for 3D BatchNorm, and UPX2 stands for an upsampling operation by using trilinear interpolation. All convolutional layers are 3D convolutions.
Table 4: Ablation study of 3DPN against alternative implementations on simulated data over synthesized radiographic views are generated from XraySyn and can be used straightfor-wardly to reduce bone attenuation, they do not preserve all the details in input $I_T^{in}$. Therefore, we seek to reverse CT2Xray mathematically to find a $I_{CT}^{recon}$, such that $I_T^{ref}$ and $I_{CT}^{recon}$ reconstruct $I_T^{in}$ losslessly. As shown in Algorithm 1, we first

Algorithm 1: Reverse CT2Xray to achieve bone suppression

**Input:** input radiograph: $I_T^{in}$, estimated bone and tissue: $t_{bone}^{ref}(T)$ and $t_{tissue}^{ref}(T)$, estimated radiograph attenuation: $I_{atten}^{T}$, attenuation coefficient: $\mu(m,E)$.

**Output:** reconstructed tissue: $t_{tissue}^{recon}$, bone suppressed radiograph: $I_{in}^{Suppress}$

\[
I_{in}^{Suppress} = \frac{\max(I_T^{in}) - I_T^{in}}{\max(I_T^{in})}; // invert and normalize $I_T^{in}$
\]

\[
I_{in}^{Attent} = I_{in}^{Attent} * (\max(I_T^{atten}) - \min(I_T^{atten})) + \min(I_T^{atten}); // normalize $I_T^{in}$ to the range of $I_T^{atten}$
\]

Given $I_T^{atten} = \sum E e^{-\sum m \mu(m,E)t_{ref}^{m}(T)} = e^{-\sum m \mu(m,E)t_{ref}^{m}(T)} + ... + e^{-\sum m \mu(m,E)\mu_{ref}^{m,E}(T)} = w_1 * I_{atten}^{T} + ... + w_N * I_{atten}^{T} \]

\[
w_i = e^{-\sum m \mu(m,E)t_{ref}^{m}(T)} // estimate decomposition weights to reverse summation $\sum E$
\]

Given $w_i * I_{atten}^{T} = e^{-\mu(bone,E)I_{bone}^{ref}(T) + \mu(tissue,E)I_{tissue}^{ref}(T)}$

\[
t_{atten}^{recon}(T,i) = \log(w_i * I_{atten}^{T}) = \log\left(\frac{I_{atten}^{T} - I_{atten}^{bone}(T)}{I_{atten}^{bone}(T)}\right) // I_{atten}^{tissue}(T,i) is slightly different based on $w_i$
\]

\[
I_{in}^{Suppress} = \sum E e^{-\mu(tissue,E)(t_{bone}^{ref}(T) + t_{tissue}^{ref}(T))} // replace the attenuation coeff. of bone to tissue
\]

Table 4: Ablation study of 3DPN against alternative implementations on simulated data over synthesized radiographic views from -9° to 9°. PSNR/SSIM metrics are provided. The best performing metrics are bold.

**Reversing CT2Xray for bone suppression**

Bone suppression seeks to reduce bone attenuation such that the overlapped tissue is more visible. While $I_T^{ref}$ and $I_T^{atten}$ are generated from XraySyn and can be used straightfor-
transform $I_{\text{in}}^T$ from the range of image pixels to physical attenuation by using $I_{\text{atten}}^T$ (described in Eq. [6]) as the reference, and obtain $I_{\text{inAtten}}^T$. We then use $\mu_{\text{tissue}}^\text{ref}$ as a reference to calculate the weight of each term in the summation over energy level $E$ that produced $I_{\text{atten}}^T$, and apply these weights on $I_{\text{inAtten}}^T$. $\mu_{\text{tissue}}^\text{recon}$ can then be calculated straightforwardly and used in the forward attenuation calculations to obtain bone-suppressed radiograph $I_{\text{Supress}}^T$, where the attenuation coefficient of $\mu_{\text{bone}}^\text{ref}$ is reduced. A comparison of bone suppression by using $\{\mu_{\text{bone}}^\text{ref}, \mu_{\text{recon}}\}$ and $\{\mu_{\text{bone}}^\text{ref}, \mu_{\text{tissue}}^\text{ref}\}$ are shown in Fig. 7.

Figure 7: Visual comparison of bone suppression generated from $\{\mu_{\text{bone}}^\text{ref}, \mu_{\text{recon}}\}$ and $\{\mu_{\text{bone}}^\text{ref}, \mu_{\text{tissue}}^\text{ref}\}$. Note that details are much clearer in the $\{\mu_{\text{bone}}^\text{ref}, \mu_{\text{tissue}}^\text{ref}\}$ generated image.

More detailed metrics on simulated data

A more detailed view synthesis evaluation is shown in Table 4, where PSNR/SSIM is measured over ten different synthesized views. As we can see, view synthesis becomes less accurate with respect to the ground truth as the view angle deviates further away from the input view. This is consistent with intuition as a single view contains very limited 3D information.

Additional visual results

For additional visual results of XraySyn on radiograph view synthesis and bone suppression, please refer to the attached multi-media file.