Computerized Medical Imaging and Graphics 98 (2022) 102067

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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

XctNet: Reconstruction network of volumetric images from a single X-ray image

Zhiqiang Tan[a,b], Jun Li[a,b], Huiren Tao[c], Shibo Li[a,*], Ying Hu[a,*]

[a] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen Key Laboratory of Minimally Invasive Surgical Robotics and System, Shenzhen 518055, China
[b] University of Chinese Academy of Sciences, CAS, Beijing 100049, China
[c] Department of Orthopaedics, Shenzhen University General Hospital , Shenzhen University Clinical Medical Academy, Shenzhen 518055, China

ARTICLE INFO

Keywords:
3D reconstruction
Self-attention module
Deep learning
X-ray
Volumetric images

ABSTRACT

Conventional Computed Tomography (CT) produces volumetric images by computing inverse Radon transformation using X-ray projections from different angles, which results in high dose radiation, long reconstruction time and artifacts. Biologically, prior knowledge or experience can be utilized to identify volumetric information from 2D images to certain extents. a deep learning network, XctNet, is proposed to gain this prior knowledge from 2D pixels and produce volumetric data. In the proposed framework, self-attention mechanism is used for feature adaptive optimization; multiscale feature fusion is used to further improve the reconstruction accuracy; a 3D branch generation module is proposed to generate the details of different generation fields. Comparisons are made with the state-of-arts methods using public dataset and XctNet shows significantly higher image quality as well as better accuracy (SSIM and PSNR values of XctNet are 0.8681 and 29.2823 respectively).

1. Introduction

Image based 3D reconstruction, which is to infer 3D shapes from single or multiple 2D images, has been explored in the field of computer vision for decades (Han et al., 2021) and has become the basis of many fields, such as robot navigation, 3D modeling and animation, object recognition, scene understanding, medical diagnosis, etc. However, it is not straightforward to extract volumetric information from digital images without disparity knowledge from stereo correspondence, due to the lack of approaches to derive depth information from pixels. Projectional radiography, a conventional way to observe the inside of objects or bodies, is no different than normal photography, except that the pixels carry rich observations of transparent volumetric structure other than opaque surface. Specifically, in X-ray radiographs, each pixel is the line integral of attenuation data following Radon transformation in 2D space, so that the inversion of radiographs is not 2D–3D point conversion but the conversion from 2D pixel to spatial line distribution. Therefore, 3D reconstruction from radiographs is an even more challenging task.

Clinically, there are only limited number of available approaches for 3D reconstruction. Computed tomographic (CT), a well-developed way of radiograph-based 3D reconstruction, is the commonly used way to obtain patients' volumetric information and has many variations such as CBCT, PET-CT, etc. CT is inherently an inverse Radon transformation process, in which the spatial distribution function of the X-ray attenuation is solved by Inverse Fourier Transformation (IFT), so that the angular integral of projection views from all directions would be calculated. In practice, projections from a large number of different angular positions are requisite in order to maintain acceptable resolution and mitigate physics-based artifacts of the tomographs. The reconstruction process intrinsically determines the inevitable limitations of CT, such as high radiation, long reconstruction time and patient-movement-based artifacts. Other than CT, the novel EOS imaging system offers a better alternative for full-body biplanar X-ray scan and 3D reconstruction of the whole skeleton. The new technology is extremely helpful in diagnosis of orthopedic diseases, such as adolescent idiopathic scoliosis (AIS) and adult degenerative knee arthritis (Lenke et al., 2001; Ovadia, 2013). However, the reconstruction process of EOS imaging is based on statistical shape models (SSMs), so that the obtained model is not the exactly same reflection of the patient but a semantically similar virtual one instead.

Biologically, although our eye-brain vision system does not make 3D reconstruction from plain pixels, we can still partially obtain the hidden spatial information from subtle evidence like: shadow, occlusion, light/

[*] Corresponding authors.
E-mail addresses: zq.tan@siat.ac.cn (Z. Tan), jun.li@siat.ac.cn (J. Li), huiren_tao@163.com (H. Tao), sb.li@siat.ac.cn (S. Li), ying.hu@siat.ac.cn (Y. Hu).

https://doi.org/10.1016/j.compmedimag.2022.102067
Received 18 December 2021; Received in revised form 15 March 2022; Accepted 8 April 2022
Available online 15 April 2022

shade, relative size, etc. The knowledge we use in this evidence-based stereo reconstruction process can be defined as prior knowledge, which plays an essential part in human vision-based judgment and identification system. When we look at photographs or images by bare eyes, the relative spatial relation of the objects or bodies could always be deduced by combining pixel information with prior knowledge. Similarly, radiologists are able to tell the spatial information of human bodies from radiographs by applying prior knowledge from anatomy and everyday practice. Therefore, from the aspect of biomimetic, prior knowledge has potential to reconstruct 3D information at least partially from radiographs theoretically.

Deep learning, which shows great advantages over traditional methods in fitting complex nonlinear mathematical relations, has brought an evolution to numerous medical fields such as medical image segmentation, lesion area recognition, medical image registration, etc. (Feng et al., 2020; Schwartz et al., 2019; Singh et al., 2020). The possibilities of 3D reconstruction from 2D radiographs have not been observed for long, but only in recent years does the attempts of the inverse mapping emerge. Henzler et al. (2018) first, to our knowledge, applied a deep Convolutional Neural Network (CNN) to single-radiograph tomography and reconstructed 3D cranial volumes from 2D X-rays. Kasten et al. (2020) used an end-to-end CNN for 3D reconstruction of knee bones from bi-planar X-ray images. Shen et al. (2019) developed a deep network system with representation, transformation, generation modules to generate volumetric tomography images from single or multiple 2D X-rays. Through the current literature research, it can be found that the current CNN-based reconstruction methods use the end-to-end network structure, which will cause a certain loss of image resolution due to the network sampling process; Secondly, the task of CT volumetric images reconstruction based on X-ray image is quite computationally expensive. Thus, this paper constructs a lightweight CNN-based reconstruction network, XctNet, which can not only improve the information loss caused by the sampling process, but also greatly reduce the required computing resources. To summarize, the contribution can be seen as follow:

- This paper constructs a lightweight CNN based reconstruction network, XctNet, which can also ensure the reconstruction accuracy of the network on the premise of greatly reducing the required computing resources.
- We attempt to add attention mechanism and multi-scale feature fusion module into the feature extraction process to redundant features on the reconstructed image and further improve the pixel loss in the reconstruction process.
- We propose a 3D branch generation module, namely New Inception module, which can better generate the details of different generation fields by using different sizes of convolution kernels.

2. Related work

2.1. 2D–3D reconstruction via deep learning

Various deep learning algorithms have been proposed in 3D reconstruction of natural images, including supervised learning, unsupervised learning and semi supervised learning etc. (Han et al., 2019). Wu et al. (2015) proposed a convolutional deep belief network to represent a geometric 3D shape (3D ShapeNet) as a probability distribution of binary variables on a 3D voxel grid and also constructed ModelNet in order to train 3D deep learning model. Wu et al. (2016) used generative adversarial network to generate 3D objects from a probabilistic space by leveraging recent advances in volumetric convolutional networks and generative adversarial nets (3D-GAN). Wang et al. (2017) introduced a hybrid framework, which combined a 3D Encoder-Decoder Generative Adversarial Network (3D-ED-GAN) and a Long-term Recurrent Convolutional Network (LRCN), and their model was fit into GPU memory compared with other 3D CNN methods. Li et al. (2017) introduced a Generative Recursive Autoencoder for Shape Structures (GRASS) and proved that without supervision, their network can learn meaningful structural hierarchies. Yan et al. (2016) formulated an encoder-decoder network for predicting 3D models from a single-view 2D image. Choy et al. (2016) designed a recurrent network to reconstruct 3D models from a sequence of multi-view images.

2.2. Reconstruction of volumetric images from X-rays

CT reconstruction is an inverse mapping mathematical process, which generates tomographic images from X-ray projection data acquired at many different angles around the patient (Stierstorfer et al., 2004). The quality of reconstruction has a fundamental impact on the radiation dose used and the researchers are trying to find better reconstruction algorithm to ensure both the accuracy and resolution of the reconstructed image while minimizing radiation dose (Kak and Slaney, 1987; Hsieh, 2003).

A multi-detector spiral CT reconstruction method is proposed based on cone beam geometry (Taguchi and Aradate, 1998). Hu (1999) studied the scanning and reconstruction principles of multi-slice spiral CT, especially the scanning and reconstruction principles of 4-slice spiral CT, and concluded that the volume coverage speed of 4-slice spiral CT is 2–3 times that of single-slice spiral CT, which can provide the same image quality. Schaller et al. (2001) introduced a high-quality image reconstruction approach for helical CBCT and Flohr et al. (2003) proved its effectiveness in a 16-slice CT scanner.

The EOS system, originated from the Nobel prize-winning invention MWPC (the Multiwire Proportional Chamber) particle detector by Dr. Georges Charpak, is able to produce full-body stereo images of patients using biplanar low-dose X-ray scan and is regarded as the most advanced image acquisition equipment in orthopedics at present (Melhem et al., 2016; Song et al., 2020). Rehm et al. (2017) compared EOS imaging equipment with CT imaging equipment and showed that the EOS system can obtain high-quality images with less doses. Post et al. (2018) proposed a three-dimensional spine classification method based on the EOS system. However, the 3D reconstruction of EOS depends on parametric models and statistical inferences from collected biplanar X-ray scans, so that the generated skeleton model is only a parametric virtual substitute and is limited in circumstances of severe skeletal malformations or abnormalities like congenital scoliosis (CS) and ankylosing spondylitis (AS).

In recent years, deep learning has been widely adopted in the field of medical imaging. Meng et al. (2020) used semi-supervised learning to reconstruct high-dose volumetric images from low-dose volumetric images. Henzler et al. (2018) proposed a deep convolution to generate 3D images from a single X-ray animal skull image. Its network architecture adopts an end-to-end structure, and compared with some previously proposed network structures, it proved that their network can achieve better reconstruction results. Shen et al. (2019) proposed a 2D to 3D network model architecture and brought the idea of converting the 2D feature information into a spatial tensor in order to perform a 3D deconvolution. However, Shen's reconstruction network includes enormous amount of parameters to update, which leads to computational inefficiency. Multiple studies on machine-learning-based 3D reconstruction have also been carried out in the field of dentistry, spine, chest, etc (Ying et al., 2019; Bayat et al., 2020; Čavojská et al., 2020). The mentioned works also have limitations in generalization and tend to underperform on different datasets in practice. In this paper, a more lightweight CNN-based network is constructed to improve the reconstruction accuracy and reduce the computational cost.

3. Methodology

The architecture of XctNet reconstruction network, shown in Fig. 1, includes the two major parts:The X-ray feature extraction module, Multi-scale feature fusion module and the volumetric image generation

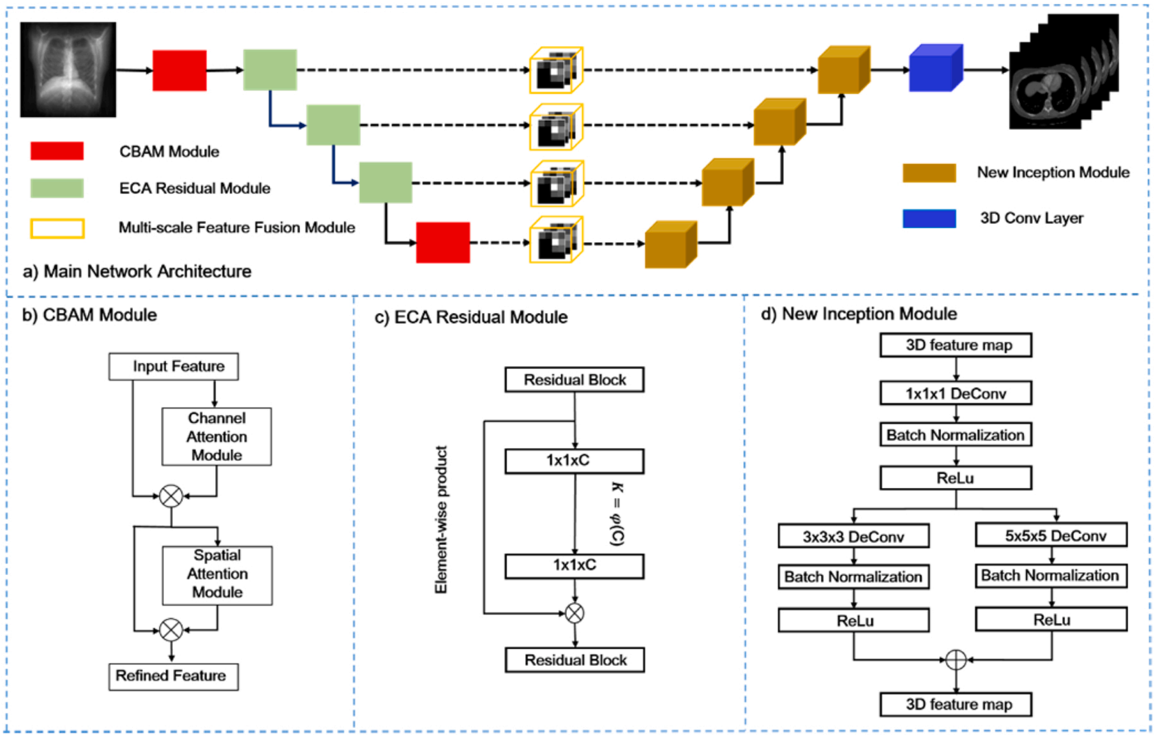


Fig. 1. Architecture of XctNet. The model contains X-ray feature extraction module, multi-scale feature fusion module and the volumetric images generation module. The input of the model is a single 2D projection image. The X-ray feature extraction module extract feature information from the input X-ray image. The multi-scale feature extraction module converts 2D features into 3D features and performs feature fusion with the corresponding 3D generation module. The volumetric images generation module, which consists of a series of New Inception module, uses the extracted feature data to generate the corresponding volumetric image.

module. The details of each module and the loss function will be explained in the following sections.

3.1. XctNet reconstruction network architecture

X-ray feature extraction module: The introduction of deep residual network is to solve the problem of gradient disappearance caused by too many network layers in the training process. The most representative structure of the residual network is ResNet (He et al., 2016), which directly connects the input terminal to the following layer through the shortcut structure, thereby protecting the integrity of the transmitted data. The feature extraction module is built based on ResNet34. The input is a single X-ray image which size is 128×128, the first layer of the model is composed of a convolutional layer with a kernel size of 7×7 and stride 2, and the second to fifth layers are composed of four residual blocks which contains two 3×3 convolutional layers. The number of channels of the convolutional layer in each residual block is kept the same to ensure that the shortcut path and the residual path can maintain the same size during the element-wise addition operation. The size of the feature representation output is 4×4. In addition, through experiments, we find that when using the encoder/decoder structure for pixel level vision tasks, the convolution layer can only use local information to calculate the target pixel value. Therefore, the lack of global information will undoubtedly lead to deviation. The error caused by the convolution layer can be described by the covariance between the pixel values shown in Eq. (1). Each pixel value x_i in the feature map obtained by the convolution layer can be used as a random variable, and $\overline{x}$ is the mean value of the feature map. The similarity between the two variables can be evaluated by calculating the covariance of the two random variables. The attention mechanism is to use the similarity between pixels to improve the performance of convolution layer.

$$Cov(x,y) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) \tag{1}$$

In order to reduce the error caused by the convolution process, this paper introduces two attention mechanisms, CBAM (Convolutional Block Attention module) (Woo et al., 2018) and ECA (Efficient channel Attention module) module (Wang et al., 2020), to adaptively improve the feature extraction ability and reduce the error. Specifically, CBAM derives the attention graph from the 2D information of space and channel, then, the attention graph with the input features will be multiplied to adaptively optimize the eigenvalues. The module structure is shown in Fig. 1(b). In the 2D feature extraction module, CBAM is mainly used for convolution feature extraction of the first layer and the last layer, so as to improve the ability of feature adaptive extraction on the premise of ensuring that the overall network structure is not affected. For the intermediate convolution layer, ECA module is used to improve its feature extraction performance. As a local cross-channel interaction module that does not reduce the feature dimension, ECA module obtains local cross-channel interaction information by combining each channel and its adjacent k channels. ECA module can be realized by one-dimensional convolution layer with the size of k. It is worth noting that ECA, as a lightweight module, does not add a large number of additional parameters. The network structure diagram is shown in Fig. 1(c). In this paper, the feature extraction capability of the middle layer is improved by combining the residual module and ECA module. The feature map obtained by the residual module will be input into the ECA module for adaptive optimization. In addition, the feature map obtained by the residual module will also be combined with the adaptively optimized feature map through element-wise product, so as to get a refined feature map.

Multi-scale feature fusion module: CNN based reconstruction network structure extracts 2D features layer by layer through down sampling and then reconstructs 3D information through up sampling operation. In the down sampling process, the shallow network has strong semantic information representation ability, but lacks spatial geometric details; the deep network has strong representation ability of geometric detail information, but it lacks semantic representation ability. The traditional encoder/decoder structure directly inputs the features to the decoder for up sampling after the down sampling process. For the volume image reconstruction task, volume images usually have aplenty of detail information. While the image generated based on the traditional encoder/decoder structure will lack a lot of detail information. In order to solve the loss of detail in volumetric images and improve the fine-grained features of network generated images, a multi-scale feature fusion method is proposed in this paper.

The process of multi-scale convolution mainly includes two factors: feature propagation and cross-scale communication (Feng et al., 2020). In the multi-scale feature extraction structure proposed in this paper, the input feature will be divided into high-scale feature $\mathbf{X}_{\mathbf{high}}$ and low-scale feature $\mathbf{X}_{\mathbf{low}}$ to obtain the corresponding high-scale feature output $\mathbf{Y}_{\mathbf{high}}$ and low-scale feature output $\mathbf{Y}_{\mathbf{low}}$ respectively. The multi-scale feature transform process can be seen as follow:

$$\begin{bmatrix} \mathbf{Y}_{\mathbf{high}} \\ \mathbf{Y}_{\mathbf{low}} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\omega} \end{bmatrix} \begin{bmatrix} \mathbf{X}_{\mathbf{high}} \\ \mathbf{X}_{\mathbf{low}} \end{bmatrix} \tag{2}$$

Eq. (2) can be encapsulated as the aggregating transformation performed on the input feature maps. Where $\mathbf{I}$ refers to the information mapping, $\boldsymbol{\omega}$ represents the transformation in the same scale. In other words, high-scale features will be connected with the corresponding generation module through layer skipping connection, while low-scale features will be down sampled through a series of convolution layers.

The overall network structure is shown in Fig. 1(a). Feature maps of different layers are extracted from the original network structure and converted into corresponding 3D feature maps through transform module and then the multi-scale 3D feature maps are combined with corresponding 3D feature generation layers, so as to improve the fine granularity of generation results and reduce the loss of information in the reconstruction process. For details, to convert 2D projection data to volumetric images data requires data conversion. A transform module is added to bridge the 2D feature extraction module and the 3D generation module. the multi-scale 2D features, which size is$(\mathrm{C}, \mathrm{H}, \mathrm{W})$, are converted to $(\mathrm{C}, 1, \mathrm{H}, \mathrm{W})$ by the dimension conversion function, after that, the converted 3D feature map, which size is $(\mathrm{C}, \mathrm{D}, \mathrm{H}, \mathrm{W})$ can be obtained through a deconvolution operation with a kernel size of $\mathrm{D} \times 1 \times 1$. In addition, the ReLU activation function and the batch normalization function are also included to better learn the transformation relationship in the transform process.

Volumetric image generation module: Inspired by the Inception structure of GoogleNet (Szegedy et al., 2015), which can solve overfitting and gradient disappearance problems, a 3D deconvolution form of the Inception structure (New Inception) is added to the 3D generation network, which is composed of a 3D point-wise convolution with a kernel size of $1 \times 1 \times 1$ and two 3D deconvolution with kernel size of $3 \times 3 \times 3$ as well as $5 \times 5 \times 5$. As can be seen in the Fig. 1(d), the 3D point-wise convolution layer for the deconvolution module can superimpose more deconvolutions in the generation field of the same size, thereby more details could be obtained in the generated images. In addition, 3D point-wise convolution also plays a fundamental role in dimensionality reduction. Performing 3D deconvolution operations will generate a huge amount of calculation. The number of input features can be effectively reduced by adding 3D point-wise convolution, so as to increase the computational efficiency. The New Inception structure is composed of two branches, each of which uses filters of different sizes for deconvolution. The branches can generate information of different scales and generate richer results. The New Inception structure uses the principle of decomposing a sparse matrix into a dense matrix for calculation. The feature dimension is decomposed into multiple densely distributed sub-feature sets. The highly correlated features are clustered together and the unrelated features will be weakened. Finally, they will be spliced together in the feature dimension and consistent with the input dimension. This approach reduces the calculation cost and ensures that the final training results will not be affected.

3.2. Data pre-processing

The original data needs to be preprocessed before fed into the network model for training. First of all, all input data need to be resized to the same size. The 2D images and the corresponding 3D CT images used for training are resized to 128 × 128 and 128 × 128 × 128 separately. In practice, 2D–3D data pairs should be composed of X-ray and CT images, owing to the lack of corresponding paired images, this paper uses digitally reconstructed radio algorithm (DRR) to generate an approximate single 2D projection image to obtain the corresponding 2D–3D data pairs. As shown in the Fig. 2, this article uses point source vision based DRR projection algorithm to generated 2D projection (Moturu and Chang, 2018). The advantage of this method is that the point source can be randomly selected to obtain X-rays, which makes the data change slightly. Specifically, after the light source point is selected and the projection distance is fixed (centered on the front of the CT volumetric image), 2D projections are generated according to Beer's law (Feeman, 2010), in which the intensity loss measurement of X-rays passing through the body is modeled by Beer's law. The information (spacing, size, direction) of the CT image is obtained and the image is used as the input of the DRR algorithm, and then the image is resampled by coordinate transformation. Moreover, we set the distance from the light source to the projection plane to 400 mm, and the default pixel spacing of the projection pixel plane is 0.8 × 0.8 and set the threshold to − 80, and the bilinear interpolation is used to integrate each voxel plane traversed, so as to obtain the anterior-posterior positions of the 2D projected image. In order to enrich the sample size of training data, data augmentation, which includes scale change, rotation change, mirror image and translation change, brightness change, chroma change, contrast change and sharpness change, is performed before training. Moreover, the pixel-wise input data is normalized to the interval [0,1].

3.3. Evaluation metrics

In order to evaluate the performance of the model, we tested the trained model on the test set and used different evaluation metrics to evaluate the predicted reconstruction results. Four evaluation functions is used in this paper for model evaluation, namely: MSE (mean squared error), MAE (mean absolute error), SSIM (structural similarity) and PSNR (peak signal noise ratio). MSE and MAE are used to evaluate the deviation between the predicted reconstruction result and the target value. The smaller MSE/MAE value, the closer the reconstruction result is to the real situation. The image evaluation metric SSIM, incorporating the information of luminance, contrast and structures, is used to evaluate the degree of similarity between images. The commonly used PSNR is applied to evaluate the quality of our reconstructed volumetric images. Generally, the resultant images with better structural and higher resolution will have higher SSIM and PSNR values. Each metric value is averaged for all test samples and different methods are compared as

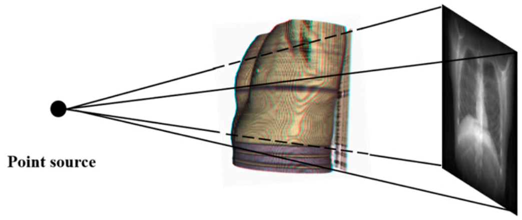


Fig. 2. DRR projection algorithm based on point source vision.

shown in Table 2.

4. Reconstruction experiments results

4.1. Datasets

The input sample consists of a single 2D projection image X and volumetric CT images Y. In the training process, with the single X-ray image as input $X \in \mathbb{R}^{H \times W}$, the model output is volumetric images $Y_{pre} \in \mathbb{R}^{C \times H \times W}$, while $Y_{GT} \in \mathbb{R}^{C \times H \times W}$is the ground truth which is the reference standards for model training.

In order to verify the effectiveness of the model, we used the public dataset, the Lung Image Database Consortium image collection (LIDC-IDRI) (Armato et al., 2011), which contains 1081 CT volumetric image cases. The original data will be divided into training set, test set and verification set separately, on this basis, the original data is expanded to 59,708 cases through data augmentation. Among them, the training set is 35,825 cases, the verification set is 11,941 cases, and the test set is 11, 942 cases. At the same time, the corresponding input X of each case is generated by DRR projection.

4.2. Training details

The size of the input X is 128 × 128 and the size of the ground truth Y_{GT} is 128 × 128 × 128. The network is trained on a device with three NVIDIA Tesla V100 graphic processing units, and the training platform is Pytorch. Training epoch is 64. As an important parameter in deep learning training, it is particularly important to select the appropriate learning rate. This paper constructs an adaptive learning rate adjustment strategy, which can modify the learning rate according to the specific situation in the training process, so as to ensure the best effect of training. As shown in Eq. (3), the loss function used in all three trained networks was MSE. The training results were shown in Fig. 3.

$$L_{MSE}(Y, Y_{GT}) = \frac{1}{N}\sum_{n=1}^{N}(Y_{GT_n} - Y_n)^2 \tag{3}$$

In addition, to verify the effectiveness of the multi-scale feature module and attention mechanism feature extraction module proposed in this paper, we construct three network models with different structures (ResXct, CBAM/ECAXct, XctNet) and reproduce the network structure proposed by Shen et al. (2019), namely ReconNet, The specific structure is shown in the Table 1:

Table 1
Structural details between different networks.

	Representation network (Layer)	Volumetric image generation module (Layer)	Attention module	Multi-scale feature fusion module
ReconNet	10	10	/	/
ResXct	5	5	/	/
CBAM/ ECAXct	10	5	√	/
XctNet	10	5	√	√

4.3. Training result analysis

We show the reconstruction results of ResXct, CBAM/ECAXct, XctNet as well as ReconNet on the LIDC-IDRI data set. These abnormalities can further prove the effectiveness of the XctNet model.

As shown in Fig. 3, a test sample is randomly selected to show the generation results of the different model and the numbers of the selected slices are 3, 15, 35, 65 and 85. The results shown in Fig. 3(a)–(d) are slice images which are randomly selected from the test sample; Fig. 3(e) is the corresponding ground truth. From the overall result of reconstruction, the result generated by our XctNet is closer to the ground truth. In terms of details, the content of the slice image generated by the original version model is relatively vague; compared with the ReconNet model, the overall contour of the intermediate version model is clearer, and some internal details, such as rib areas, can be reconstructed. On the other hand, from the chest slice data at different positions, the best reconstruction detail is the bony area, while the reconstruction accuracy of internal organs in the chest is blurred to varying degrees. Besides, by randomly selecting multiple test data for analysis, it can be found that in the reconstructed volumetric data, the reconstruction performance of the middle of the volume is generally better than that of the front of the volume and the end of the volume. The main reasons for this phenomenon are as follows: First of all, owing to this paper is to reconstruct the

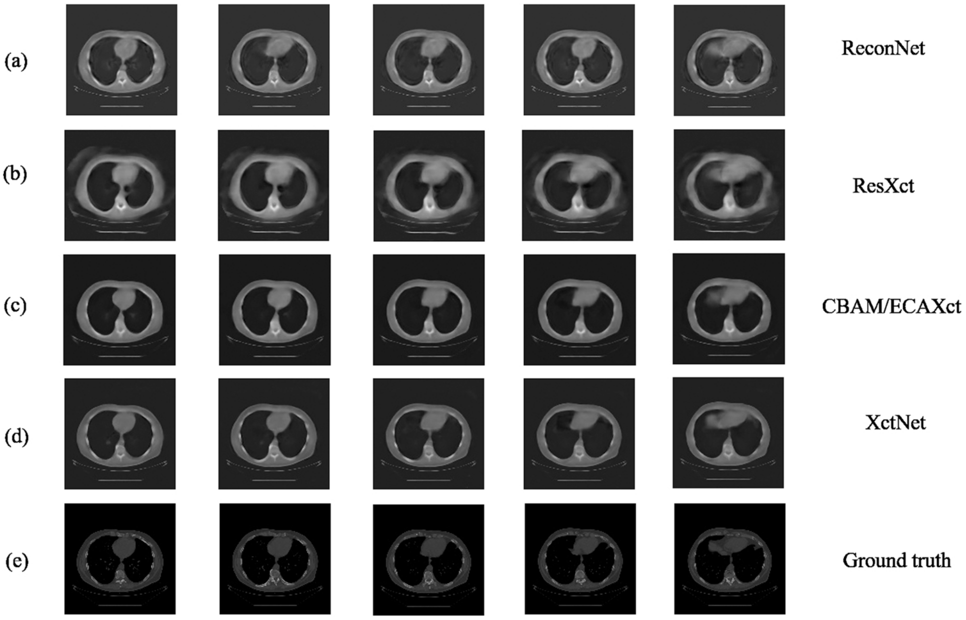


Fig. 3. Volumetric image examples from the test set. (a)–(d) represent the results generated by the ReconNet, ResXct, CBAM/ECAXct and XctNet respectively; (e) is the ground truth. The results shown in the figure comes from different slice graphs in a volumetric image randomly selected.

whole thoracic cavity, we do not preprocess according to the HU values of different tissues and organs, but process the CT data of the whole thoracic cavity. Therefore, there will be some deviation for the details such as internal organs in the reconstruction process. Secondly, due to the different sources of the data sets, the quality of the thoracic cavity area can not be guaranteed to be completely consistent, resulting in the quality of the front and end volume data of the reconstructed CT volumetric image will be relatively worse than that of the middle volume. It can be seen that the main reason for this phenomenon is due to the complexity of data, but this does not mean that our model has limitations. As can be seen, compared with the above different models, XctNet has the best reconstruction performance in terms of overall contour and internal details.

To show the details of the differences in results between different models, as can be seen in Fig. 4, gray values indicate areas with insignificant differences, while white and black values represent areas with large differences. It can be seen from the figure that XctNet has the smallest difference with ground truth compared with other models.

4.4. Comparison with state-of-the-art

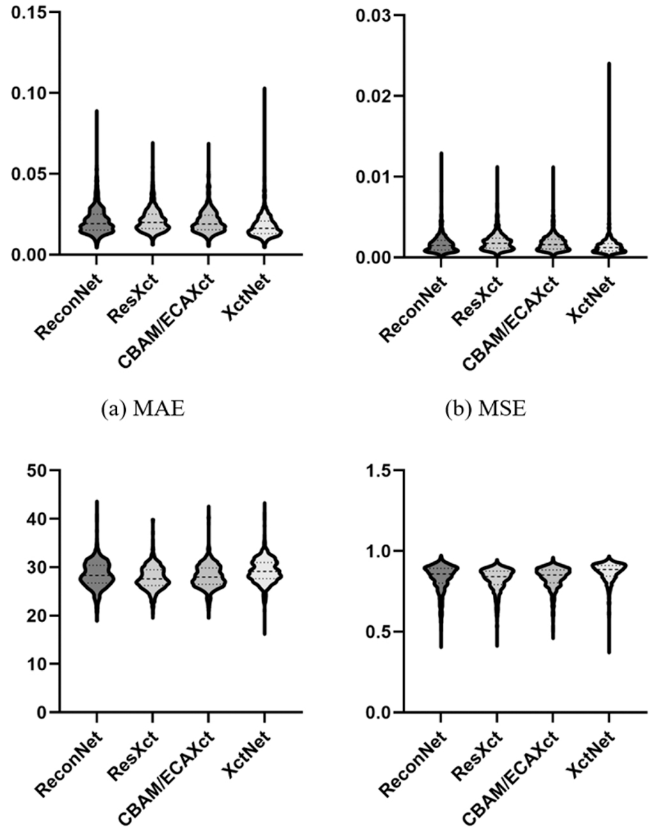
In order to make a quantitative analysis of XctNet and the proposed contrast network, four evaluation metric functions are used to analyze the difference between the predicted reconstructed image and the ground truth. In addition, by comparing the evaluation metric differences between the models, the effectiveness of attention mechanism and multi-scale fusion module can be further illustrated.

It is worth noting that the volumetric data used in this paper is composed of 128 slices of data. It is worth noting that the volumetric data used in this paper are composed of 128 slices of data. The 128 slices of volumetric data is evaluated separately by using different evaluation metric functions, and then obtain the final evaluation result by taking the average of all slices.As shown in Table 2, our XctNet can achieve the best evaluation results, and its PSNR and SSIM can reach 29.2823 and 0.8681 respectively. Incidentally, all the evaluation metric values in Table 2 are the mean values of the test samples. On the other hand, the evaluation results obtained by ReconNet perform better than our baseline model, RexXct, which shows that increasing the network depth is effective for the reconstruction results. In addition, the evaluation results of CBAM/ECAXct are similar to ReconNet, that is, adding lightweight attention mechanism is also an effective method to enhance the performance of the model without increasing the network depth.

From the overall distribution of the evaluation results of the test set, as shown in Fig. 5, the four violin graphs represent the results of different evaluation functions. Overall, XctNet achieves the better results in all evaluation metrics. On the other hand, as shown in Fig. 5(c), the distribution of PSNR values of all models are mostly concentrated near the inferior quartile, that is mainly because PSNR evaluates the gray difference between images and due to the data set used in this paper is complicated, the prediction results are usually different. From the comparison results, the interquartile range (IQR) of XctNet is smaller than the other three models, which shows that XctNet model is more stable. Through the above data analysis, we can get the conclusions that self-attention mechanism and multi-scale feature fusion module can

Table 2
Evaluation on reconstruction results of lung CT cases.

	MAE	MSE	PSNR	SSIM
ReconNet	0.0211	0.0017	28.4891	0.8349
ResXct	0.0213	0.0019	27.8489	0.8250
CBAM/ECAXct	0.0205	0.0017	28.1859	0.8351
XctNet	**0.01764**	**0.0013**	**29.2823**	**0.8681**



(a) MAE

(b) MSE

(c) PSNR

(d) SSIM

Fig. 5. Distribution of evaluation results of different models. (a)–(d) represents the distribution of evaluation results of MAE, MSE, PSNR and SSIM on LIDC-IDRI data set respectively. It can be seen that the result distribution in (a)–(c) approach to the inferior quartile and (d) approach to the superior quartile.

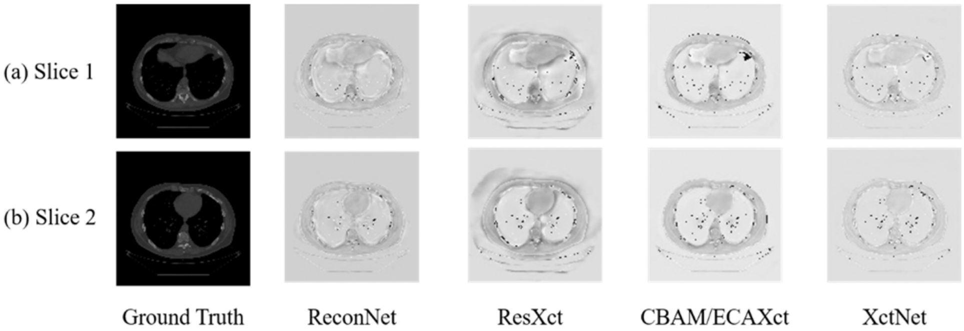


Fig. 4. Comparison of deviation with respect to ground truth. The first column correspond to the ground truth. The second column shows the difference between ReconNet and ground truth. Other columns represent the difference between the corresponding model and the ground truth.

greatly improve the output accuracy of the reconstructed model.

By summarizing and analyzing the training results of the four groups of models, the evaluation metrics obtained by our XctNet exceeds the ReconNet. This result confirms that without additional network depth, attention mechanism and multi-scale feature fusion module can also improve the accuracy of the model.

5. Discussion

The advantages of the XctNet model is further illustrated by analyzing the semantic representation of the model. For the 2D projection image based reconstruction task, only when the feature extraction module extracts the key and useful feature information can the volumetric images be correctly reconstructed. The three models constructed in this paper contain 512 feature maps with a size of 4×4. For better visualization, 16 feature maps are randomly selected to illustrate the feature representation between different models. As shown in Fig. 6(a)–(b), it can be seen that the feature map generated by the feature extraction module with self-attention mechanism is more concise than the feature map without self-attention mechanism. Comparing (b)–(c), it can be seen that CBAM/ECA attention mechanism can further remove redundant features. On the other hand, from the perspective of the generated volumetric image, the image quality generated by the model with self-attention mechanism is significantly better. In other words, the feature extraction module without self-attention mechanism learns a lot of redundant information, which leads to the phenomenon of fuzzy generation results. Therefore, the conclusion can be got that the self-attention mechanism can help the model to better learn the feature information.

In order to further verify the performance of XctNet model, the current state of art CNN models are compared. As shown in Table 3, ReconNet model has the highest model complexity, and its FLOPs (floating point operations) reaches 1.304×10^{12}. ResXct as our baseline model, which complexity is lower than ReconNet, and the error rate is almost the same. Our XctNet has greatly improved its error rate with only a little increase in complexity. This phenomenon shows the following two aspects. Firstly, the lightweight attention mechanism can improve the performance of the model without increasing the complexity of the model. Secondly, the New Inception module proposed in this paper can greatly reduce the amount of model calculation and generate volumetric images with richer content.

According to the definition of information entropy, it represents the overall characteristics of an information source in an average sense. For image information, we can describe the amount of information contained in the image according to image entropy. As shown in Eq. (4), x represents each pixel in the image, the image entropy reflects the average information of an image and the image entropy obtained for specific image information is unique. Therefore, the generation quality of volumetric images can be evaluated from the perspective of image entropy.

Table 3
Comparison with different CNNs models on the LIDC-IDRI data set.

	#. Param. (× 10^7)	FLOPs (× 10^9)	Error rate (%)
ReconNet	58.85914	1304	2.125
ResXct	**3.02648**	**144.359**	2.126
CBAMXct	3.03315	144.360	2.059
ECAXct	**3.02648**	**144.359**	2.004
CBAM/ECAXct	3.03316	144.360	2.001
XctNet	3.45363	211.109	**1.791**

$$H_{Entropy}(X) = -\sum_{n=1}^{m} p_i(x) log p_i(x) \quad (4)$$

As shown in Fig. 7, two test samples are randomly selected to illustrate the distribution of image entropy comes from different model, which shows the ground truth and the entropy map of the different models. It can be seen that, the area in which the entropy map tends to be cold indicates that it contains less information and the area in which the entropy map tends to be warm indicates that it contains more information. By comparing the results of the three models with the ground truth, we can find that XctNet can get a distribution map more inclined to the ground truth by adding self-attention mechanism and multi-scale fusion module. By comparing the ReconNet and ResXct with the other two models, it further shows that the traditional end-to-end network will bring deviation in the convolution process, and also verifies the effectiveness of the improved method proposed in this paper.

To verify the performance of the model on clinical X-ray images, we obtained 10 original X-ray chest images through the spine surgery of the General Hospital of Shenzhen University. As shown in Fig. 8, the reconstruction results of two groups of clinical X-ray images are shown. It can be seen that the accuracy of the reconstruction result are worse than that of the 2D projection used in this paper. The main reason for this phenomenon is that there are some differences between clinical X-ray images and 2D projections. Therefore, in the future work, more in-depth research from clinical X-rays need to conducted. However, from the generation results of ReconNet and XctNet proposed in this paper, XctNet performs better in clinical X-rays data, which also confirms that the network we constructed has considerable superiority.

6. Conclusion

In this paper, we focus on the reconstruction quality of volumetric image. In order to obtain more accurate reconstruction results, a lightweight reconstruction network, XctNet, is constructed. The network structure has the following three innovations:

Firstly, self-attention mechanism is be added to the original residual feature extraction module to remove redundant features; Secondly, a multi-scale feature fusion module is proposed in this paper to improve

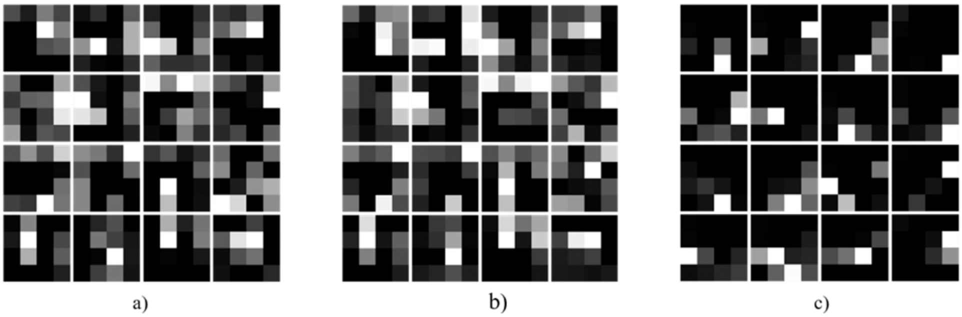


Fig. 6. Feature extraction and network structure analysis. a) Feature map learned from 2D feature extraction module without attention mechanism; b) Feature map learned from 2D feature extraction module only with CBAM attention mechanism; c) Feature map learned from 2D feature extraction module with CBAM/ECA attention mechanism.

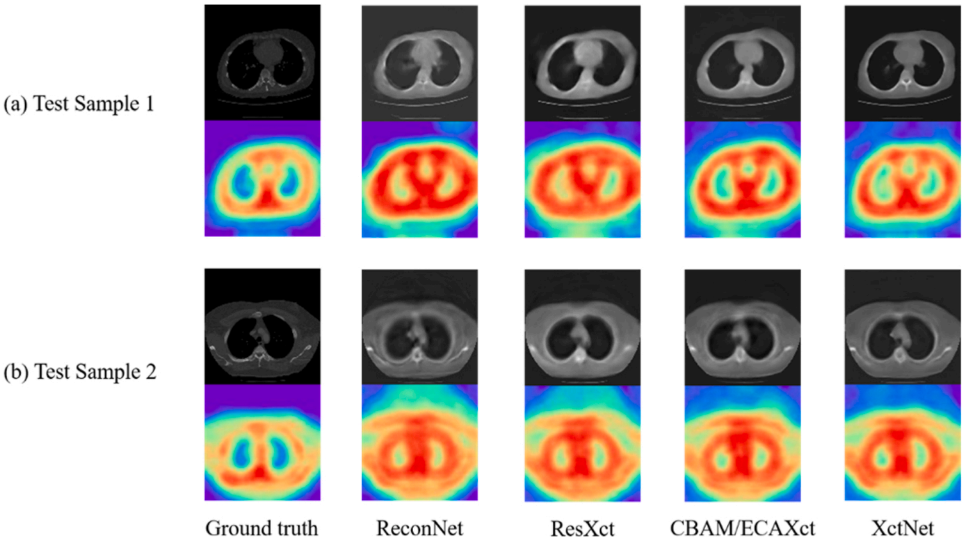


Fig. 7. The entropy map generated by different network structures. a–b) represent different test samples and their corresponding entry information. The variation of image entropy is closely related to the content contained in the image. As can be seen, the less content the image contains, the lower the image information entropy, that is, the color of the entropy map tends to be cold.

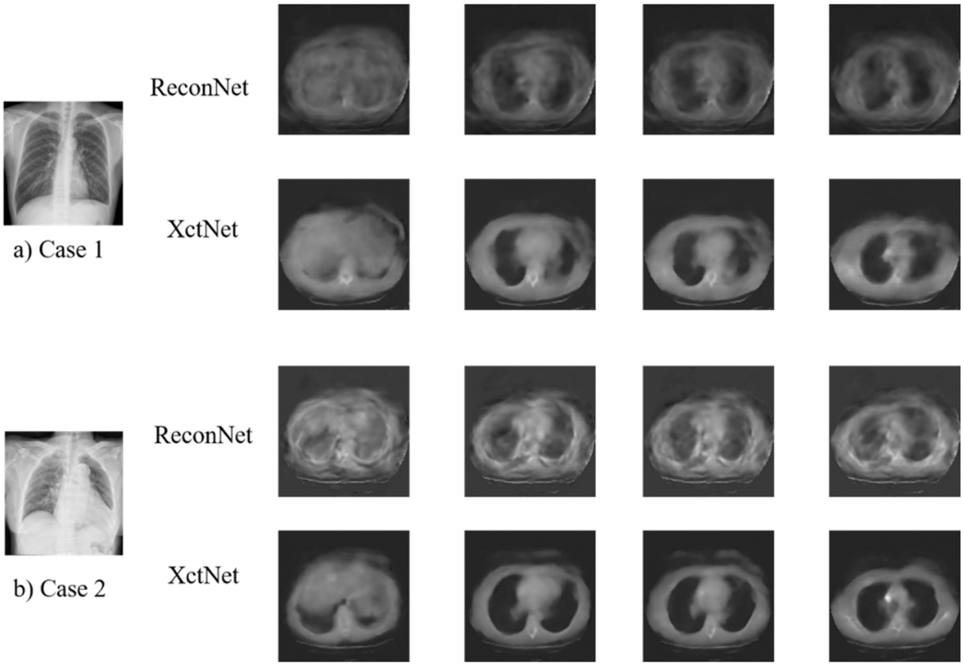


Fig. 8. Clinical X-ray reconstruction results of difference cases. a–b) represent volumetric images reconstructed from different X-ray images. The number of slices shown in the figure are 30, 60, 70, and 90.

the quality of reconstructed image and other details. Finally, a feature generation module called New Inception module is constructed to obtain richer feature information and more accurate reconstruction results. At the same time, there are still some problems to be solved in this paper. In the actual application scenario, the corresponding 2D projection image should be an X-ray image, but the 2D projection used in this article is projected by the DRR algorithm. To solve this problem, using style transfer algorithm may considered to solve the difference between clinical X-rays and DRR projection. In conclusion, XctNet as a lightweight framework can further improve the results of volumetric image reconstruction.

CRediT authorship contribution statement

Zhiqiang Tan: Conceptualization, Methodology, Software, Investigation. **Jun Li:** Data curation, Software **Huiren Tao:** Resources, Validation. **Shibo Li:** Visualization, Writing – review & editing. **Ying Hu:** Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by Key-Area Research and Development Program of Guangdong Province (No.2020B0909020002), National Natural Science Foundation of China (Grant No. 62003330), Shenzhen Fundamental Research Funds (Grant No. JCYJ20200109114233670 , No.JCYJ20190807170407391, JCYJ20180507182415428), Natural Science Foundation of Guangdong Province (Grant No. 2019A1515011699) and Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology.

References

Armato, Samuel G., McLennan, Geoffrey, Bidaut, Luc M., McNitt-Gray, Michael F., Meyer, C.R., Reeves, Anthony P., Zhao, Binsheng, Aberle, Denise R., Henschke, Claudia I., Hoffman, Eric A., Kazerooni, Ella A., MacMahon, Heber, Van Beeke, Edwin J.R., Yankelevitz, David F., Biancardi, Alberto M., Bland, Peyton H., Brown, Matthew S., Engelmann, Roger M., Laderach, G.E., Max, Daniel, Pais, Richard C., Qing, D.P., Roberts, Rachael Y., Smith, Amanda R., Starkey, Adam, Batrah, Poonam, Caligiuri, Philip, Farooqi, Ali O., Gladish, Gregory W., Jude, Cecilia Matilda, Munden, Reginald F., Petkovska, Iva, Quint, Leslie E., Schwartz, Lawrence H., Sundaram, Baskaran, Dodd, Lori E., Fenimore, Charles, Gur, David, Petrick, Nicholas A., Freymann, John B., Kirby, Justin S., Hughes, Brian, Casteele, Alessi Vande, Gupte, Sangeeta, Sallamm, Maha, Heath,Michael, Kuhn, M., Dharaiya, Ekta, Burns, Richard, Fryd, David, Salganicoff, Marcos, Anand, V., Shreter, Uri, Vastagh, Stephen, Croft, Barbara Y., 2011. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med. Phys. vol. 38, issue 2, pp. 915–31.

Bayat, Amirhossein, Kumar Sekuboyina, Anjany, Paetzold, Johannes C., Payer, Christian, Štern, Darko, Urschler, Martin, Kirschke, Jan S., Menze, Bjoern H., 2020. Inferring the 3D standing spine posture from 2D radiographs. MICCAI.

Čavojská, Jana, Petrasch, Julian, Mattern, Denny, Lehmann, Nicolas J., Voisard, Agnès, Böttcher, Peter, 2020. Estimating and abstracting the 3D structure of feline bones using neural networks on X-ray (2D) images. Commun. Biol. 3 (n. pag).

Choy, Christopher Bongsoo, Xu, Danfei, Gwak, JunYoung, Chen, Kevin, Savarese, Silvio, 2016. 3D-R2N2: a unified approach for single and multi-view 3D object reconstruction. ECCV 628–644.

Feeman, Timothy G., 2010. The Mathematics of Medical Imaging.

Feng, Ruicheng, Guan, Weipeng, Qiao, Yu, Dong, Chao, 2020. Exploring Multi-Scale Feature Propagation and Communication for Image Super Resolution. ArXiv abs/2008.00239, n. pag.

Flohr, T., Stierstorfer, K., Bruder, H., Simon, J., Polacin, A., Schaller, S., 2003. Image reconstruction and image quality evaluation for a 16-slice CT scanner. Med. Phys. 832–845.

Han, Xian-Feng, Laga, Hamid, Bennamoun, Mohammed, 2019. Image-based 3D object reconstruction: state-of-the-art and trends in the deep learning era. IEEE Trans. Pattern Anal. Mach. Intell. 43 (5), 1578–1604.

Han, Xian-Feng, Laga, Hamid, Bennamoun, Mohammed, 2021. Image-based 3D object reconstruction: state-of-the-art and trends in the deep learning era. IEEE Trans. Pattern Anal. Mach. Intell. 43, 1578–1604.

He, Kaiming, Zhang, X., Ren, Shaoqing, Sun, Jian, 2016. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–8.

Henzler, Philipp, Rasche, Volker, Ropinski, Timo, Ritschel, Tobias, 2018. Single-image tomography: 3D volumes from 2D cranial X-rays. Comput. Graph. Forum 37 (n. pag).

Hsieh, Jiang, 2003. Computed Tomography: Principles, Design, Artifacts, and Recent Advances.

Hu, H., 1999. Multi-slice helical CT: scan and reconstruction. Med Phys. 26, 5–18.

Kak, A.C., Slaney, M., 1987. Principles of Computed Tomographic Imaging. SIAM, Philadelphia, PA.

Kasten, Yoni, Doktofsky, Daniel, Kovler I., 2020. End-To-End Convolutional Neural Network for 3D Reconstruction of Knee Bones from Bi-Planar X-Ray Images. ArXiv abs/2004.00871, n. pag.

Lenke, Lawrence G., Betz, Randal R., Harms, Jürgen, Bridwell, Keith H., Clements, David H., Lowe, Thomas G., Blanke, Kathy M., 2001. Adolescent idiopathic scoliosis: a new classification to determine extent of spinal arthrodesis. J. Bone Jt. Surg. 83, 1169–1181.

Li, Jun, Xu, Kai, Chaudhuri, Siddhartha, Yumer, Ersin, Zhang, Hao, Guibas, Leonidas J., 2017. GRASS: Generative Recursive Autoencoders for Shape Structures. ArXiv abs/1705.02090, n. pag.

Melhem, E., Assi, A., El Rachkidi, R., Ghanem, I., 2016. EOS® biplanar X-ray imaging: concept, developments, benefits, and limitations. J. Child.'s Orthop. 1–4.

Meng, M., Li, S., Yao, L., Li, D., Zhu, M., Gao, Q., Xie, Q., Zhao, Q., Bian, Z., Huang, J., Meng, D., 2020. Semi-supervised learned sinogram restoration network for low-dose CT image reconstruction. International Society for Optics and Photonics. Phys. Med. Imaging 11312, 113120B.

Moturu, Abhishek, Chang, Alex, 2018. Creation of Synthetic X-Rays to Train a Neural Network to Detect Lung Cancer.

Ovadia, Dror, 2013. Classification of adolescent idiopathic scoliosis (AIS). J. Child.'s Orthop. 7, 25–28.

Post, Mareille, Verdun, Stéphane, Roussouly, Pierre, Abelin-Genevois, Kariman, 2018. New sagittal classification of AIS: validation by 3D characterization. Eur. Spine J. 28, 551–558.

Rehm, Johannes, Germann, Thomas, Akbar, Michael, Pepke, Wojciech, Kauczor, Hans Ulrich, Weber, Marc-André, Spira, Daniel, 2017. 3D-modeling of the spine using EOS imaging system: inter-reader reproducibility and reliability. PLoS One 12 (n. pag).

Schaller, Stefan, Stierstorfer, Karl, Bruder, Herbert, Kachelriess, Marc, Flohr, Thomas G., 2001. Novel approximate approach for high-quality image reconstruction in helical cone-beam CT at arbitrary pitch. SPIE Med. Imaging.

Schwartz, John T., Gao, Michae C., Geng, Eric, Mody, Kush S., Mikhail, Christopher M., Cho, Samuel K., 2019. Applications of machine learning using electronic medical records in spine surgery. Neurospine 16, 643–653.

Shen, Liyue, Zhao, Wei, Xing, Lei, 2019. Patient-specific reconstruction of volumetric computed tomography images from a single projection view via deep learning. Nat. Biomed. Eng. 3, 880–888.

Singh, Amitojdeep, Sengupta, Sourya, Lakshminarayanan, Vasudevan, 2020. Explainable deep learning models in medical image analysis. J. Imaging 6 (n. pag).

Song, Weinan, Liang, Yuan, Wang, Kun, He, Lei, 2020. Oral-3D: reconstructing the 3D bone structure of oral cavity from 2D panoramic X-ray. ArXiv abs/2003.08413, n. pag.

Stierstorfer, Karl, Rauscher, Annabella, Boese, Jan, Bruder, Herbert, Schaller, Stefan, Flohr, Thomas G., 2004. Weighted FBP–a simple approximate 3D FBP algorithm for multislice spiral CT with good dose usage for arbitrary pitch. Phys. Med. Biol. 49 (11), 2209–2218.

Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott E., Anguelov, Dragomir, Erhan, D., Vanhoucke, Vincent, Rabinovich, Andrew, 2015. Going deeper with convolutions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9.

Taguchi, Katsuyuki, Aradate, Hiroshi, 1998. Algorithm for image reconstruction in multi-slice helical CT. Med. Phys. 25 (4), 550–561.

Wang, Qilong, Wu, Banggu, Zhu, Pengfei, Li, P., Zuo, Wangmeng, Hu, Qinghua, 2020. ECA-net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11531–9.

Wang, Weiyue, Huang, Qiangui, You, Suya, Yang, Chao, Neumann, Ulrich, 2017. Shape Inpainting using 3D generative adversarial network and recurrent convolutional networks. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2317–25.

Woo, Sanghyun, Park, Jongchan, Lee, Joon-Young, Kweon, In-So, 2018. CBAM: convolutional block attention module. ECCV.

Wu, Jiajun, Zhang, Chengkai, Xue, Tianfan, Freeman, Bill, Tenenbaum, Joshua B., 2016. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. NIPS.

Wu, Zhirong, Song, Shuran, Khosla, Aditya, Yu, Fisher, Zhang, Linguang, Tang, Xiaoou, Xiao, Jianxiong, 2015. 3D shapenets: a deep representation for volumetric shapes. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1912–20.

Yan, Xinchen, Yang, Jimei, Yumer, Ersin, Guo, Yijie, Lee, Honglak, 2016. Perspective transformer nets: learning single-view 3D object reconstruction without 3D supervision. NIPS 1696–1704.

Ying, Xingde, Guo, Heng, Ma, Kai, Wu, Jian, Weng, Zhengxin, Zheng, Yefeng, 2019. X2CT-GAN: reconstructing CT from biplanar X-rays with generative adversarial networks. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10611–20.