MFCT-GAN: multi-information network to reconstruct CT volumes for security screening

Yixiang Jiang
Polytechnic Institute, Zhejiang University, Hangzhou, China

Received 18 October 2021
Revised 6 January 2022
Accepted 20 February 2022

Abstract

Purpose – At airport security checkpoints, baggage screening is aimed to prevent transportation of prohibited and potentially dangerous items. Observing the projection images generated by X-rays scanner is a critical method. However, when multiple objects are stacked on top of each other, distinguishing objects only by a two-dimensional picture is difficult, which prompts the demand for more precise imaging technology to be investigated for use. Reconstructing from 2D X-ray images to 3D-computed tomography (CT) volumes is a reliable solution.

Design/methodology/approach – To more accurately distinguish the specific contour shape of items when stacked, multi-information fusion network (MFCT-GAN) based on generative adversarial network (GAN) and U-like network (U-NET) is proposed to reconstruct from two biplanar orthogonal X-ray projections into 3D CT volumes. The authors use three modules to enhance the reconstruction qualitative and quantitative effects, compared with the original network. The skip connection modification (SCM) and multi-channels residual dense block (MRDB) enable the network to extract more feature information and learn deeper with high efficiency; the introduction of subjective loss enables the network to focus on the structural similarity (SSIM) of images during training.

Findings – On account of the fusion of multiple information, MFCT-GAN can significantly improve the value of quantitative indexes and distinguish contour explicitly between different targets. In particular, SCM enables features more reasonable and accurate when expanded into three dimensions. The appliance of MRDB can alleviate problem of slow optimization during the late training period, as well as reduce the computational cost. The introduction of subjective loss guides network to retain more high-frequency information, which makes the rendered CT volumes clearer in details.

Originality/value – The authors' proposed MFCT-GAN is able to restore the 3D shapes of different objects greatly based on biplanar projections. This is helpful in security check places, where X-ray images of stacked objects need to be distinguished from the presence of prohibited objects. The authors adopt three new modules, SCM, MRDB and subjective loss, as well as analyze the role the modules play in 3D reconstruction. Results show a significant improvement on the reconstruction both in objective and subjective effects.

Keywords Security screening, 3D reconstruction, GAN, Subjective visual enhancement

Paper type Research paper

1. Introduction

Baggage screening is a vital procedure in security domain. It's common to use X-rays in two or three orthogonal views for security checking on the metro station or airport. Normally, the security needs to verify whether prohibited items included in luggage or container (Benjamin, 1995), according to X-Rays pictures. The explicit contours and textures of items are important to their judgment. However, X-rays projections are inaccurate to reflect the three-dimensional information of the object, which requires security personnel with sufficient experience to distinguish the true shape of the object. CT is capable of generating

Journal of Intelligent Manufacturing and Special Equipment
Vol. 3 No. 1, 2022
pp. 17-30
Emerald Publishing Limited
e-ISSN: 2633-660X
p-ISSN: 2633-6596
DOI 10.1108/JIMSE-10-2021-0035

a set of 2D pictures that accurately reflect their 3D information, but it is expensive and not suitable for use in security screening occasion (World Health Organization, 2011). Thereby, we propose a 3D reconstruction method based on GAN, which can restore two orthogonal X-ray projections to CT volumes as realistically as possible.

Studies have shown that some aerospace accidents are caused by smuggling of explosives into checked baggage, as well as subway accidents. Therefore, a set of accurate college baggage check-in security system design, which can effectively prevent the recurrence of unsafe events (Shi *et al.*, 2021). Based on 3D X-ray CT images, it's a wildly used method to detect the thread items with X-ray scanners in aviation security (Wang *et al.*, 2020). For some items under potential danger, such as firearms, sharp objects, sharp edges, etc., they have a significant physical appearance (such as shape, volume, texture, etc.) (Mouton *et al.*, 2014). Hence, with the two-dimensional images obtained by the X-ray scanner and further reconstruction, security personnel are able to identify potentially prohibited items from the passenger's baggage, without need to manual search (Mouton and Breckon, 2015), which greatly improves the operational efficiency.

Currently, item-screening through 2D X-ray images is still a manual inspection process. This workflow is cumbersome and requires security personnel with mature relevant experience and training (Shanks and Bradley, 2004), which inevitably raises the operational threshold and tends to produce inaccurate judgments when the operator is immaturely experienced, thus reducing screening efficiency. In particular, in some cases where objects overlap exist, image interpretation becomes a challenging task because the projection map generated by X-rays does not reflect whether the objects produce overlap and simple visual inspection cannot detect whether the prohibited items are obscured by other objects (Megherbi *et al.*, 2013). Therefore, if the image can be reconstructed into a three-dimensional form, the image of the object into the real shape so that the operator can obtain the information masked in the two-dimensional X-ray, so as to obtain a clearer observation and judgment, which will greatly improve the efficiency of the security screening task.

Among most cases in the aviation security infrastructure, explosive detection systems (EDS) are now the only one CT's application approved (Singh, 2003). Based on EDS, the dual-energy CT (DECT) is the technique to distinguish different materials. The principle consists in using two different X-ray spectra to deduce the chemical composition of the investigated material based on its reaction under these spectral conditions (Jin, 2011). The DECT is divided in three categories: post-reconstruction techniques (Graser *et al.*, 2009), pre-reconstruction techniques (Alvarez and Macovski, 1976) and iterative reconstruction techniques (Semerci and Miller, 2012). Although the technique has improved the reconstruction performance, increasing computational demand becomes a significant problem. It's necessary to rotate swiftly to collect enough X-ray apparatus around the items, which is a high-cost and time-wasting process. Hence, we need to find a low computation cost method to reconstruct from 2D to 3D with less data acquisition.

For most CT reconstruction algorithms, numbers of X-ray images are required for input, which requires a certain amount of machine computational performance. Some typical principles, such as maximum likelihood (Shepp and Vardi, 1982) and sparsity (Lustig *et al.*, 2007), are used to improve the quality of tomographic reconstruction. Those methods are very time consuming, which is not suitable for the needs of fast inspection in security scenarios. In fact, the vast majority of security machines only obtain 1–3 projection images of objects mutually orthogonal to each other for screening by security personnel, so the question of how to reconstruct 3D information using as few images as possible is significant.

Traditional CT reconstruction methods, which are based on mathematical and theoretical knowledge, often require the creation of fairly accurate models. For instance, filtered back projection and iterative reconstruction (Herman, 2009), which is the one-dimensional Fourier transform of the projection is equivalent to the two-dimensional Fourier transform of the

original image. The introduction of *a priori* knowledge is also a typical method, such as statistical shape models (Lamecker *et al.*, 2006) or anatomical structures knowledge (Serradell *et al.*, 2011). However, these reconstruction methods based on mathematical knowledge require the construction of corresponding mathematical models for different objects, which means that the generalizability of the methods is not enough. Deep learning has a natural advantage in some scenarios, such as modeling of invisible parts, where traditional algorithms have difficulty estimating the depth of objects with *a priori* knowledge. Eigen *et al.* use a two-staged convolutional neural networks (CNN) to generate a 2D depth map from a 2D image (Eigen *et al.*, 2014). Philipp Henzler *et al.* apply U-NET network to get a better performance (Henzler *et al.*, 2017). Ying *et al.* (2019) propose X2CT-GAN, which perform better than traditional CNNs in terms of subjective effect of reconstruction. In addition, the format of the input data is also an important issue. Wurfl *et al.* (2016) work on X-ray sinogram as input, which is not readable for human. Magnor *et al.* reconstruct the 3D model with single X-ray image. Because a single two-dimensional picture lacks much three-dimensional information (Magnor *et al.*, 2004), the effect of three-dimensional reconstruction with only one picture is very blurred. Therefore, if two or more images are used as input, the output reconstruction will be better. Thus, inspired by previous work, we apply GAN to reconstruct CT from two X-ray images.

To sum up, our contributions include the following four main points.

(1) We propose SCM module, which introduces the second image as weight map for correction when expanding from 2D to 3D after single-channel feature extraction. The numerical and physical information are combined to enhance the reconstruction effect.

(2) We apply MRDB connection for feature extraction, which reduces the number of model parameters while alleviating the problem of model instability.

(3) We propose subjective loss function for training to improve the generated subjective effect.

(4) Compared with other reconstruction algorithms, our method improves both quantitative and qualitative indicators; especially for the restoration of internal details, the effect is significantly improved.

2. Network framework

In general, similar to X2CT-GAN (Peng *et al.*, 2020), the overall framework of our network combines GAN and U-like network (U-NET). The input is two 2D X-ray projection images and the output is 3D CT volumes. After encoder-decoder, the features of two networks are fused together and put into a new upsampling decoder to generate the final result. An overview of our network is shown in Figure 1. Here are the details of the MFCT-GAN.

2.1 Generator

The role of the generator is to produce a set of 3D CT volumes from two mutually orthogonal 2D X-ray images (vertical plane, horizontal plane or width plane). The network consists of three main components: feature extraction using MRDB connectivity, 3D decoder with upscaling module and features fusion with SCM.

Since it's a dual input, there are two parallel coder-decoder networks in generator. Features fusion component aims to integrate the double channels' 3D features, generating the final reconstructed 3D CT volumes. Given that input is dual-view X-ray images, how to extract the two images features independently and fuse them properly will directly affect the quality of reconstruction. Thereby, some modifications will be applied to raw network.

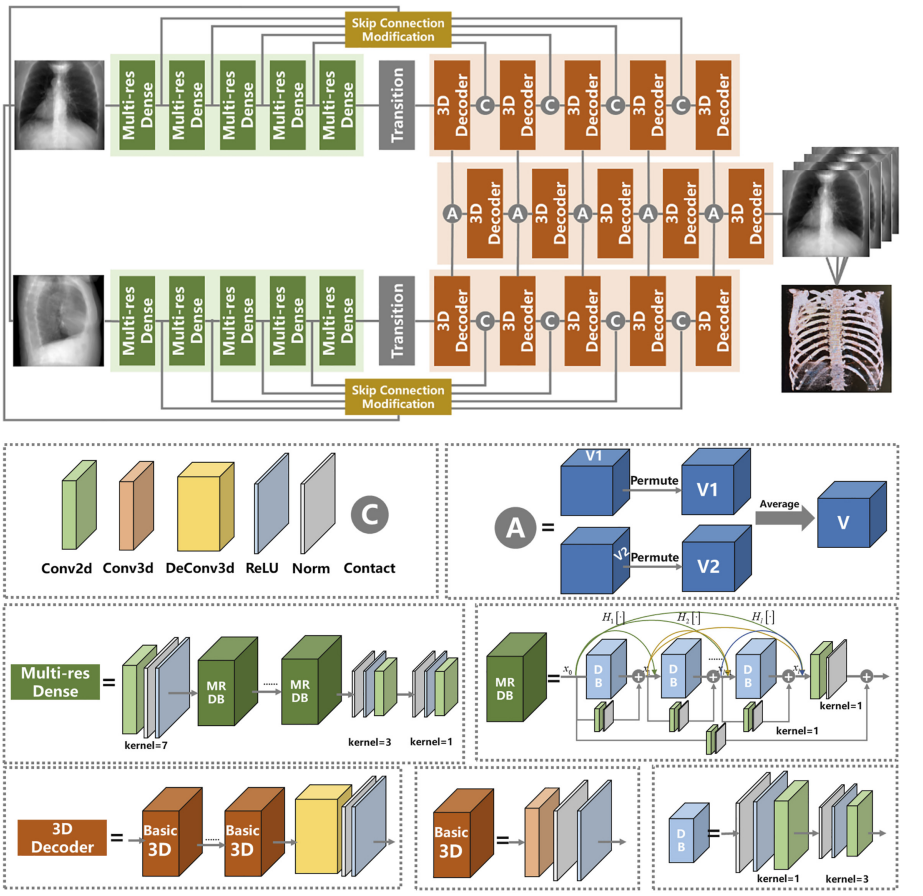


Figure 1.
The architecture of the MFCT-GAN contains two parallel encoder-decoder network

Note(s): Two X-ray image are required to input with posterior-anterior and lateral views. Our proposed modules are included here besides subjective loss function. The transition block contains fully connected layer to flatten the features, and then reshaped to three dimensional shape with batch and channels

(1) *Features extraction*: In order to extract sufficient feature information from input and reduce image pixel size, feature extraction is usually performed with dense block (DB) at beginning.

MRDB Dense Net reduces gradient to a certain extent by directly linking features among layers, which makes great use of global features (Huang *et al.*, 2017). However, as the number of DBs increases and network layers deepen, Dense Net suffers from the problem of gradient disappearance. The introduction of residual learning can alleviate this problem to some extent, which is called residual dense block (RDB) (Zhang *et al.*, 2018). However, Peng *et al.* (2020) found that multiple residual connections can sufficiently enhance the flow of information, as well as reduce the number of model parameters. Taking the network depth and efficiency of training into consideration, we propose a modified multi-residual dense

network in downsampling network, which is shown in Equation (1). The relationship between input and output can be expressed as follows:

$$x_l = H_l[x_0, x_1, \ldots, x_{l-1}] + x_{l-1}, \tag{1}$$

where the $x_0, x_1, \ldots, x_{l-1}$ denotes the $l-$ th layers' DB output and $H_l[\,]$ denotes DB, which produces growth rate of feature maps. Different layers' output is converted as $H_l[\,]$ input. For the last layer, we introduce the first input as residual learning.

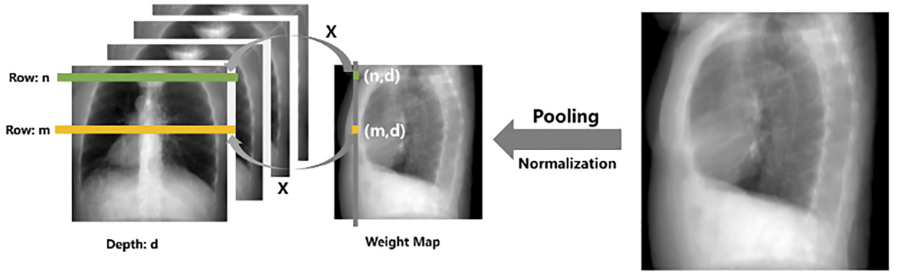
As for other details of MRDB, due to the connection of multiple residuals, the input and output share the same number of channels and image size. This is because each DB is followed by a 1×1 kernel filter to increase the number of channels and thus connecting the residuals with the previous layer. Each MRDB is followed by a transition block to change the number of input channels for the next MRDB. Hence, the feature extraction capability of the network will be increased effectively.

(2) *3D decoder*: When the process of feature extraction for the biplanar input image is completed, one dimension needs to be augmented for subsequent decoding. Inspired by previous work (Isola *et al.*, 2017), we add a depth channel to the input data with the same number of width and length channels, in other words, expanding the two-dimensional to three-dimensional by duplicating the feature maps.

After bridging the 2D encoder and 3D decoder, we apply the classical upsample method to decode, which consists of two main modules: one is Conv3d-Norm-ReLU for generating more details of reconstruction and the other is Deconv3d-Norm-ReLU for restoring the size of 3D CT volumes.

SCM Combining the long- and short-skip connections is beneficial for deep neural networks (Drozdzal *et al.*, 2016). It is very common to employ skip connections to link encoders and decoders. The situation becomes slightly different when cope with 2D-to-3D task. Since it is necessary to expand the features in encoder from two dimensions to three dimensions and then deliver them to following decoder, how to ascend dimension becomes a critical problem. Usually, duplicating the feature map in the depth channel is a common operation, which is inaccurate and rough (Ying *et al.*, 2019; Peng *et al.*, 2020; Ratul *et al.*, 2021). To better utilize the biplanar information, we propose a novel skip connection module (SCM), shown in Figure 2, to transmit low-level features to high-level features.

In summary, when the first image is encoded and need to expand one dimension, the second image is introduced as weight map and the 3D features of the first input are corrected. Finally, the rectified features are fed into decoder. Specifically, the value pixels in the second image are



Note(s): Through the pooling operation, the second image shares the same size as the first set of feature maps correspondingly. After normalization, the second picture turn into a weight map to modify the feature maps, thus making the operation of ascending dimension more reasonable

Figure 2.
The details of SCM

normalized to a weight map with the shape (H, W). The shape of the 3D feature of the first image is (D, H, W). In particular, the values of D, H and W are equal. After that, 3D features are corrected by multiplying the factor in weight map. Specifically, all the pixels in row H of the feature maps are productized by the weight factor in column H of D row in the weight map.

In order to make the weight map consistent with the shape of feature maps, a similar averaging pooling operation is introduced to reduce the size of the weight map. The modification from weight map makes the feature maps work better when expanding from two-dimensional to three-dimensional. The dual-view input information are also fully utilized.

(3) *Features fusion*: The shortcoming of using a single 2D image to reconstruct into 3D CT volumes is the weak generalization of model. It is rarely enough to learn useful 3D features only by relying on the recurrent training of the deep network, especially for the application where there are many different kinds of objects. That is the reason why we use two mutually orthogonal X-ray images as model input. The complementary information enables the network to generate more accurate reconstruction results. Naturally, after parallel dual-channel encoder-decoder network, we need to fuse those features information.

We apply the third decoder network with the same structure as the 3D decoder mentioned before. Given that in reality, two X-ray images are captured at almost the same time, which means there is no motion occlusion in images. Therefore, we consider that outputs of both decoders share the same information importance. On account of this, the third decoder's input is the average of the outputs from the dual-parallel decoders. As result, the output of third decoder network is final reconstructed 3D CT volumes.

2.2 Discriminator

Based on PatchGANs (Ledig *et al.*, 2017; Zhu *et al.*, 2017), which perform great generalization property and is frequently applied in generating images, we use the modified PatchDiscriminator (Isola *et al.*, 2017) to work on our network. The vanilla discriminator of PatchGAN is a matrix of $N * N$ rather than a scalar value. By discriminating each patch, local image features can be extracted and characterized, which is more conducive to high-resolution-image-generation task.

We replace the original conv2d module with conv3d module. Conv3d-Norm-ReLU with kernel size = 3 are used three times, followed by same architecture with kernel size = 1 and end with a conv3d layer.

3. Loss functions

In order to balance the quantitative metrics values and qualitative subjective evaluation after 3D reconstruction, we apply four loss functions to constrain the generative model.

3.1 Adversarial loss

GAN is a significant architecture to generate photorealistic images, which is well studied in recent research (Goodfellow *et al.*, 2014). Typically, the classical GAN use sigmoid cross-entropy as objective function, which is usually suitable for logical classification. Gradient dispersion inevitably becomes a potential problem. Least squares GAN (LSGAN) replace the original loss function with least squares loss function (Mao *et al.*, 2017). The new object function penalizes samples which are in discriminative truth away from the decision boundary and drags the false samples back into the boundary. In the end, the problem of disappearing gradients is alleviated, which result in improvement on the generated images. The LSGAN loss is defined as follows:

$$\ell_{\mathrm{LSGAN}}(D) = \frac{1}{2}\left[E_{x,y}(D(y|x)-1)^2 + E_x(D(G(x)|x)-0)^2\right],$$
$$\ell_{\mathrm{LSGAN}}(G) = \frac{1}{2}\left[E_x(D(G(x)|x)-1)^2\right], \tag{2}$$

where G denotes the generator, D denotes the discriminator, x denotes the input of two orthogonal X-ray projection images and the y denotes the corresponding CT volumes ground truth.

3.2 Projection loss

Since the loss function of LSGAN put the same importance on each pixel point, the generated image effect is close to the true value in the whole aspect. However, it cannot keep similar to the true value in the structure. And in real life, due to a prior knowledge, even if there is only a two-dimensional picture, people can easily imagine its original three-dimensional appearance (Jiang *et al.*, 2018). Thus, we use projection loss as prior knowledge to constrain the geometric shape in network, which is defined as (Ying *et al.*, 2019) follows:

$$\begin{aligned}\ell_{\mathrm{pro}} = \frac{1}{2}\Big[&E_{x,y}\left\|P_v(y)-P_v(G(x))\right\|_2^2 \\ &+E_{x,y}\left\|P_h(y)-P_h(G(x))\right\|_2^2 \\ &+E_{x,y}\left\|P_w(y)-P_w(G(x))\right\|_2^2\Big],\end{aligned} \tag{3}$$

where P_v denotes the projection in the vertical plane, P_h denotes the projection in the horizontal plane and P_w denotes the projection in the width plane.

3.3 Reconstruction loss

The binarization calculation is done in least squares function for the generated image and the ground truth, which can make it difficult for model to focus on the regions with larger pixel values on images during training. Therefore, in the final generated CT volumes after rendered, the blurring will occur in reconstruction and the information is seriously lost. Therefore, another pixel-level loss function needs to be introduced. Inspired by previous work (Johnson *et al.*, 2016), we apply volume reconstruction loss to constrain the model in pixel, which is defined as follows:

$$\ell_{\mathrm{rec}} = E_{x,y}\|y-G(x)\|_2^2. \tag{4}$$

3.4 Subjective loss

Both projection loss and reconstruction loss are biased toward pixel-level operation, which lead to high peak signal-to-noise radio (PSNR) calculation. However, the high PSNR metric does not have a completely positive correlation with the subjective effect seen by the human eyes. Rouditchenko *et al.* (2019) proposed a novel, differentiable error function, combined with l1-norm and SSIM, showing great improvement on image restoration ( ). SSIM is the function to compute similarity between two images, which is related to subjective evaluation of images. Based on previous study, we propose subjective loss, which is defined as follows:

$$\ell_{subj} = E_{x,y}\omega\cdot\ell_{\mathrm{SSIM}} + E_{x,y}(1-\omega)\cdot smooth_{\ell_1}, \tag{5}$$

where ℓ_{SSIM} represents the SSIM loss function and $smooth_{\ell_1}$ represents the smooth ℓ_1 loss, which aims to avoid the gradient no longer changing when the learning rate is too small in the late training period (Yu *et al.*, 2016).

3.5 Total loss

After introducing four loss functions mentioned above, our final objective function is as follows:

$$\begin{aligned} D &= \arg\min_{D} \alpha_1 \cdot \ell_{\mathrm{LSGAN}}(D), \\ G &= \arg\min_{G} [\alpha_1 \cdot \ell_{\mathrm{LSGAN}}(G) + \alpha_2 \cdot \ell_{\mathrm{pro}} + \alpha_3 \cdot \ell_{\mathrm{rec}} + \alpha_4 \cdot \ell_{subj}], \end{aligned} \tag{6}$$

where the α is the weight of different loss function, representing the importance of four loss terms. Given that in a realistic security check place, we put more attention on subjective similarity. Therefore, we will appropriately increase the weight of subjective loss. The final weight is set as follows: $\alpha_1 = 0.1, \alpha_2 = 8, \alpha_3 = 8, \alpha_4 = 2, \omega = 9$.

4. Experiment details

4.1 Datasets

In order to better train the model, we need the X-ray projection maps obtained from the security scanner and the corresponding CT volumes. However, due to the high cost, and the fact that the corresponding available datasets do not exist online, therefore, we use the available chest CT scan dataset on public: the lung image database consortium (LIDC-IDRI) (Armato *et al.*, 2011). To obtain the corresponding 2D orthogonal projection images, the corresponding X-rays are synthesized by using the digitally reconstructed radiographs (DRR) (Milickovic *et al.*, 2000) technique with CycleGAN (Zhu *et al.*, 2017) in reference to the work of Ying *et al.* (2019). In summary, there are 920 paired datasets for training and 98 paired datasets for testing. Each paired dataset contains two X-rays images with resized shape of 128×128 and a CT volume set with resized shape of $128 \times 128 \times 128$.

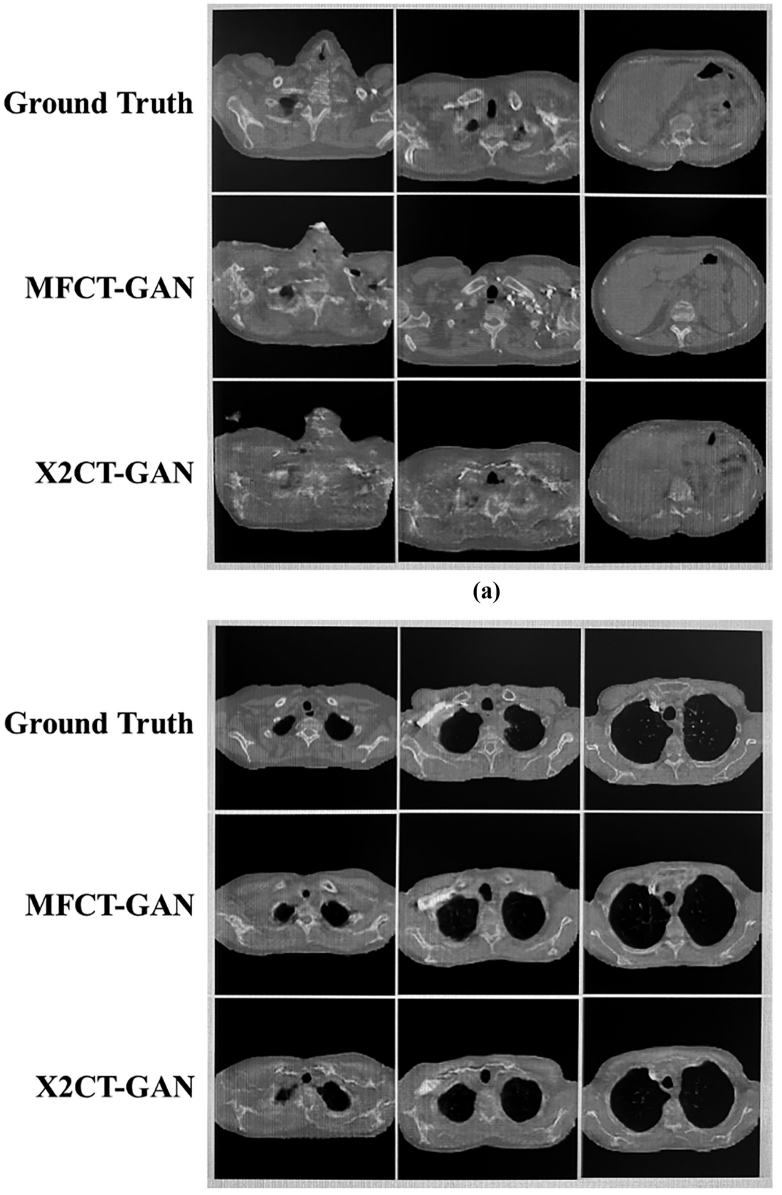
4.2 Metrics

We use two typical metrics as our quantitate results: PSNR and SSIM. PSNR is calculated based on the mean square error and reflects the relationship between the maximum signal and the background noise (Horé and Ziou, 2010). In a word, it's an objective index for evaluating images. SSIM is calculated based on the brightness and contrast of local patterns (Wang *et al.*, 2004). This index is close to the real-human perception situation, so SSIM is an image-quality-evaluation standard in line with human intuition.

5. Results

5.1 Qualitative analysis

As shown in Figure 3, the first row shows the CT ground truth and the second row shows the generated effect of the baseline X2CT-GAN+B while the third row shows our proposed model MFCT-GAN reconstruction effect. We compare the subjective quality of them. It can be seen that our model produces higher-quality reconstruction compared to the baseline. In particular, (1) our model can produce more explicit contour boundaries, which can clearly distinguish between cavities and solids; (2) for different internal organs, we can see clearer anatomical structures, such as the shape of the spine and spinal cord and (3) for consecutive CT images, our model can capture structural changes of organs fast, so as to adjust the generation effect of the next CT image in time.



Note(s): The first row is ground truth. The second row is MFCT-GAN(ours). The third row is X2CT-GAN (baseline). It can be seen that our proposed method has a clearer reconstruction of the edges of different internal organ contours

Figure 3.
Comparison of the reconstruction results of the two methods

We visualize the CT sequence by volume rendering (Brian *et al.*, 1996) as shown in Figure 4. From left to right, this is the reconstruction effect, including ground truth, X2CT-GAN+B and MFCT-GAN (ours). As we can see, the baseline method is prone to useless

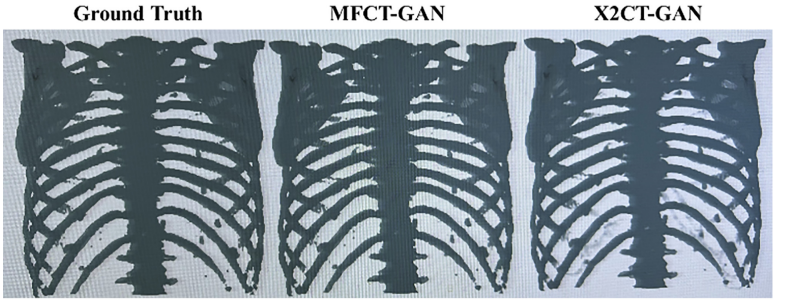


Figure 4.
Volume rendering of bones from CT volumes

Note(s): From left to right, the first is GroundTruth, the second one is reconstructed from MFCT-GAN (ours) and the last is reconstructed from X2CT-GAN (baseline)

Table 1.
Quantitative results of 2DCNN, X2CT-GAN and MFCT-GAN

Model	PSNR (dB)	SSIM
2DCNN (Henzler *et al.*, 2017)	23.1	0.461
X2CT-GAN-S (Ying *et al.*, 2019)	22.3	0.525
X2CT-GAN-B (Ying *et al.*, 2019)	26.19	0.656
MFCT-GAN-B (ours)	*31.01*	*0.676*

Note(s): 2DCNN generate CT volumes from single input. "-S" indicates the single-view X-ray input and "-B" represents the biplanar X-ray input. The best results are shown in italics

Table 2.
Comparison of computational performance results of different feature extraction networks

Method	Parameters (M)	Iteration time per step (s)
DB	61.74	4.02
RDB (Zhang *et al.*, 2018)	51.63	3.99
MRDB (ours)	*48.12*	*2.99*

Note(s): RDB represents residual dense block, which is well applied in neural network's backbone. The best results are highlighted in italics

Table 3.
Evaluation of different proposed modules

SCM		√ (Layer 1)	√ (Layer 1–2)	√ (Layer 1–3)	√ (Layer 1–4)	√ (Layer 1–4)	√ (Layer 1–4)
MRDB					√	√	√
Subjective loss						√ ($\omega = 0.8$)	√ ($\omega = 0.9$)
PSNR (dB)	27.15	29.3772	30.1684	30.2057	30.5508	30.7102	*31.0051*
SSIM	0.638	0.6107	0.6541	0.6302	0.6491	0.6588	*0.6761*

Note(s): "Layer 1" represents modification on the first multi-res dense, "Layer 1–2" denotes modification on the first and second multi-res dense and "Layer 1–4" denotes modification on the overall multi-res dense. "ω" is the weight of SSIM loss in subjective loss. The best results are bold for viewing

redundant data in the thoracic body and the internal real vascular restoration is not as detailed and accurate as our proposed method.

In the realistic security scenario, the 3D restoration of the object overlay needs to be solved at first. Therefore, the accurate distinction between internal objects becomes the main indicator of the quality of our reconstruction. According to Figure 4, we are confident that the MFCT-GAN model can achieve it greatly.

5.2 Quantitate analysis

In this section, we discuss the metric enhancement of our proposed method. 2DCNN is a CT reconstruction method that appeared very early, only for single-view input (Henzler *et al.*, 2017); X2CT-GAN is the baseline method, where "S" denotes single-view X-ray input and "B" denotes biplanar X-rays input.

We use PSNR and SSIM as evaluation metrics, and the results are shown in Table 1. It can be clearly seen that the 3D reconstruction using GAN network works better than the traditional CNN. The dual-view input can contain more 3D information, so the reconstruction accuracy is higher. And specifically comparing the baseline with MFCT-GAN, our proposed method has a significant improvement in PSNR up to 4.82 dB (18.4%). Meanwhile, the SSIM metric improvement increases slightly by 0.02 (3.05%). When the PSNR value exceeds 30, we can consider the image quality as good. On account of introduction of subjective loss function, although the index increase is limited, the subjective effect does improve greatly.

Analysis of the calculated performance changes is shown in Table 2. It is easy to conclude that the use of residual learning can effectively reduce the number of model parameters and training speed; the MRDB applied in our method can further reduce the number of parameters and achieve faster computational speed, which effectively saves training time.

5.3 Ablation study

To investigate the effectiveness of the three improved modules, an ablation study was conducted and the results are shown in Table 3.

(1) The SCM part has the most obvious improvement on PSNR, which is due to the correction of another orthogonal view picture when the dimensionality expansion is performed before the feature jump connection. And the introduction of subjective loss function has the most obvious improvement on SSIM; this is because the subjective loss includes calculation of SSIM, so the training process on the network switches the importance on pixel-level alignment to the SSIM instead.

(2) SCM gradually improves the reconstruction effect as the encoder network deepens. It is reasonable to assume that the smaller the input size to the decoder, the more obvious the alignment effect will be.

(3) ω in subjective loss indicates the weight of SSIM, and SSIM can retain high-frequency information better. However, smooth L1 will pay more attention to low-frequency information. Since we concentrate on the accuracy of object internal reconstruction, the reconstruction effect can be improved by appropriately increasing the weight of w. In addition, the optimal value of w can be further investigated.

6. Conclusion

In this paper, we propose a multi-information fusion network, named MFCT GAN, to reconstruct 3D CT volumes from biplanar X-ray projection images. In order to cope with the security check scenario that requires fast restoration of object 3D information, we propose two modules and a loss function, for SCM, MRDB and subjective loss function, which are

used to improve the reconstruction quality of the vanilla network. Through qualitative and quantitative results analysis, it can be proved that our proposed network can restore the contours of different parts inside the object well and the model training speed is faster. Due to the limited dataset, we will use the actual X-ray images generated by security scanner with corresponding CT volumes for training in the future. Also, we want to design a better volume rendering method to achieve end-to-end reconstruction, which aims to improve the screening efficiency of security personnel to serve more scenarios.

References

Alvarez, R.E. and Macovski, A. (1976), "Energy-selective reconstructions in X-ray computerised tomography", *Physics in Medicine and Biology*, Vol. 21 No. 5, pp. 733-744.

Armato, S.G., III, McLennan, G., Bidaut, L., *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*, Vol. 38 No. 2, pp. 915-931.

Benjamin, R. (1995), "Object-based 3D X-ray imaging for second-line security screening", *European Convention on Security and Detection*, pp. 310-313.

Brian, S.K., David, G.H., Donald, F.B. and Fishman, E.K. (1996), "Skeletal 3-D CT: advantages of volume rendering over surface rendering", *Skeletal Radiology*, Vol. 25, pp. 207-214.

Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S. and Pal, C. (2016), "The importance of skip connections in biomedical image segmentation", *Deep Learning and Data Labeling for Medical Applications*, pp. 179-187.

Eigen, D., Puhrsch, C. and Fergus, R. (2014), "Depth map prediction from a single image using a multi-scale deep network", arXiv preprint arXiv:1406.2283.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014), "Generative adversarial nets", *Advances in Neural Information Processing Systems*, pp. 2672-2680.

Graser, A., Johnson, T.R.C., Chandarana, H. and Macari, M. (2009), "Dual energy CT: preliminary observations and potential clinical applications in the abdomen", *European Radiology*, Vol. 19 No. 1, pp. 13-23.

Henzler, P., Rasche, V., Ropinski, T. and Ritschel, T. (2017), "Single image tomography: 3D volumes from 2D X-rays", arXiv preprint arXiv:1710.04867.

Herman, G.T. (2009), *Fundamentals of Computerized tomography:Image Reconstruction from Rojection*, Springer-Verlag, London.

Horé, A. and Ziou, D. (2010), "Image quality metrics: PSNR vs. SSIM", in *2010 20th International Conference on Pattern Recognition*, pp. 2366-2369.

Huang, G., Liu, Z., Maaten, L.V.D. and Weinberger, K.Q. (2017), "Densely connected convolutional networks", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700-4708.

Isola, P., Zhu, J.Y., Zhou, T. and Efros, A.A. (2017), "Image-to-image translation with conditional adversarial networks", in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 5967-5976.

Jiang, L., Shi, S., Qi, X. and Jia, J. (2018), "GAL: geometric adversarial loss for single-view 3D-object reconstruction", in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 820-834.

Jin, Y.N. (2011), "Implementation and optimization of dual energy computed tomography", [PhD Thesis], University of Erlangen-Nuremberg.

Johnson, J., Alahi, A. and Li, F.F. (2016), "Perceptual losses for real-time style transfer and super-resolution", in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 694-711.

Lamecker, H., Wenckebach, T.H. and Hege, H.C. (2006), "Atlas based 3D-shape reconstruction from X-ray images", in *18th International Conference on Pattern Recognition*, pp. 371-374.

Ledig, C., Theis, L., Huszar, F., *et al.* (2017), "Photo-realistic single image super-resolution using a generative adversarial network", *Proc. IEEE Conf. Compute Vision and Pattern Recognition*, pp. 4681-4690.

Lustig, M., Donoho, D. and Pauly, J.M. (2007), "Sparse MRI: the application of compressed sensing for rapid MR imaging", *Magnetic Resonance in Medicine*, Vol. 58 No. 6, pp. 1182-1195.

Magnor, M., Kindlmann, G. and Hansen, C. (2004), "Constrained inverse volume rendering for planetary nebulae", *IEEE Visualization*, Vol. 2004, pp. 83-90.

Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z. and Smolley, S.P. (2017), "Least squares generative adversarial networks", in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2813-2821.

Megherbi, N., Breckon, T.P. and Flitton, G.T. (2013), "Investigating existing medical CT segmentation techniques within automated baggage and package inspection", in *Proc. SPIE 8901, Optics and Photonics for Counterterrorism, Crime Fighting and Defence IX; and Optical Materials and Biomaterials in Security and Defence Systems Technology X, 89010L (16 October 2013)*.

Milickovic, N., Baltas, D., Giannouli, S., Lahanas, M. and Zamboglou, N. (2000), "CT imaging based digitally reconstructed radiographs and their application in brachytherapy", *Physics in Medicine & Biology*, Vol. 45 No. 10, pp. 2787-7800.

Mouton, A. and Breckon, T.P. (2015), "A review of automated image understanding within 3d baggage computed tomography security screening", *Journal of X-Ray Science and Technology*, Vol. 23 No. 5, pp. 531-555.

Mouton, A., Breckon, T.P., Flitton, G.T. and Megherbi, N. (2014), "3D object classification in baggage computed tomography imagery using randomised clustering forests", in *IEEE International Conference on Image Processing*, pp. 5202-5206.

Peng, J.S., Fu, K., Wei, Q.J., Qin, Y. and He, Q.W. (2020), "Improved multiview decomposition for single-image high- resolution 3D object reconstruction", *Wireless Communications and Mobile Computing*, Vol. 2020, pp. 2-4.

Ratul, M.A.R., Yuan, K. and Lee, W.S. (2021), "CCX-rayNet: a class conditioned convolutional neural network for biplanar X-rays to CT volume", in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1655-1659.

Rouditchenko, A., Zhao, H., Gan, C., McDermott, J. and Torralba, A. (2019), "Self-supervised audio-visual Co-segmentation", in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2357-2361.

Semerci, O. and Miller, E.L. (2012), "A parametric level-set approach to simultaneous object identication and background reconstruction for dual-energy computed tomography", *IEEE Transactions on Image Processing*, Vol. 21 No. 5, pp. 2719-2734.

Serradell, E., Romero, A., Leta, R., Gatta, C. and Morenonoguer, F. (2011), "Simultaneous correspondence and non-rigid 3d reconstruction of the coronary tree from single x-ray images", in *2011 International Conference on Computer Vision*, pp. 850-857.

Shanks, N. and Bradley, A. (2004), *Handbook of Checked Baggage Screening: Advanced Airport Security Operation*, WileyBlackwell, NJ.

Shepp, L.A. and Vardi, Y. (1982), "Maximum likelihood reconstruction for emission tomography", *IEEE Transactions on Medical Imaging*, Vol. 1 No. 2, pp. 113-122.

Shi, Y., Zhou, X.D., Cheng, J., *et al.* (2021), ""One-Time face recognition system" drives changes in civil aviation smart security screening mode", *China's e-Science Blue Book 2020*, pp. 399-425.

Singh, S. (2003), "Explosives detection systems (EDS) for aviation security", *Signal Processing*, Vol. 83 No. 1, pp. 31-55.

Wang, Z., Bovik, A.C., Sheikh, H.R. and Simoncelli, E.P. (2004), "Image quality assessment: from error visibility to structuralsimilarity", *IEEE Trans. Image Processing*, Vol. 13 No. 4, pp. 600-612.

Wang, Q., Ismail, K.N. and Breckon, T. (2020), "An approach for adaptive automatic threat recognition within 3D computed tomography images for baggage security screening", *Journal of X-Ray Science and Technology*, Vol. 28 No. 1, pp. 35-58.

World Health Organization. (2011), *Baseline Country Survey on Medical Devices 2010*, pp. 145-146.

Wurfl, T., Ghesu, F.C., Christlein, V. and Maier, A. (2016), "Deep learning computed tomograph", *Medical Image Computing and Computer Assisted Intervention*, pp. 432-440.

Ying, X.D., Guo, H., Ma, K., Wu, J., Weng, Z.X. and Zheng, Y.F. (2019), "X2CT-GAN: reconstructing CT from biplanar X-rays with generative adversarial networks", in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10619-10628.

Yu, J.H., Jiang, Y.N., Wang, Z.Y., Cao, Z.M. and Huang, T. (2016), "UnitBox: an advanced object detection network", in *Proceedings of the 24th ACM international conference on Multimedia*, pp. 516-520.

Zhang, Y., Tian, Y., Kong, Y., Zhong, B. and Fu, Y. (2018), "Residual dense network for image super-resolution", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2472-2481.

Zhu, J.Y., Park, T., Isola, P. and Efros, A.A. (2017), "Unpaired image to-image translation using cycle-consistent adversarial networks", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2242-2251.

Further reading

Zhao, H., Gallo, O., Frosio, I. and Kautz, J. (2017), "Loss functions for image restoration with neural networks", *IEEE Transactions on Computational Imaging*, pp. 47-57.

Corresponding author

Yixiang Jiang can be contacted at: 1405576787@qq.com