3DSRNet: 3-D Spine Reconstruction Network Using 2-D Orthogonal X-Ray Images Based on Deep Learning

Yuan Gao, Hui Tang, Rongjun Ge, Jin Liu, Xin Chen, Yan Xi, Xu Ji, Huazhong Shu, Jian Zhu, Gouenou Coatrieux, et al.

▶ **To cite this version:**

Yuan Gao, Hui Tang, Rongjun Ge, Jin Liu, Xin Chen, et al.. 3DSRNet: 3-D Spine Reconstruction Network Using 2-D Orthogonal X-Ray Images Based on Deep Learning. IEEE Transactions on Instrumentation and Measurement, 2023, 72, pp.4506214. 10.1109/TIM.2023.3296838 . hal-04244429

HAL Id: hal-04244429
https://univ-rennes.hal.science/hal-04244429v1

Submitted on 9 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3296838

3DSRNet: 3D Spine Reconstruction Network Using 2D Orthogonal X-ray Images Based on Deep Learning

Yuan Gao, Hui Tang, Rongjun Ge, Jin Liu, Xin Chen, Yan Xi, Xu Ji, Huazhong Shu, Jian Zhu, Gouenou Coatrieux, *Senior Member, IEEE*, Jean-Louis Coatrieux, *Fellow, IEEE*, Yang Chen, *Senior Member, IEEE*

***Abstract*—Orthopedic spine disease is one of the most common diseases in the clinic. The diagnosis of spinal orthopedic injury is an important basis for the treatment of spinal orthopedic diseases. Due to the complexity of the spine structure, doctors usually need to rely on orthopedic CT image data for accurate diagnosis. In some cases, such as poor areas or in emergency situations, it is difficult for doctors to make accurate diagnoses using only 2D x-ray images due to lack of 3D imaging equipment or time crunch. Therefore, an approach based on 2D x-ray images is needed to solve this problem. In this paper, a novel 3D spine reconstruction technique based on 2D orthogonal x-ray images (3DSRNet) is designed. 3DSRNet uses a generative adversarial network architecture and novel modules to make 3D spine reconstruction more accurate and efficient. Spine reconstruction CNN-transformer framework (SRCT) is employed to effectively integrate local bone surface information and long-range relation spinal structure information. Spine reconstruction texture framework (SRTE) is used to extract spine texture features to enhance the effect of pixel-level reconstruction. Experiments show that 3DSRNet achieves excellent 3D spine reconstruction results on multiple metrics including PSNR (45.4666 dB), SSIM (0.8850), CS (0.7662), MAE (23.6696), MSE (9016.1044), and LPIPS (0.0768).**

Manuscript created October, 2022; This work was supported in part by the State Key Project of Research and Development Plan under Grants 2022YFC2408500 and 2022YFC2401600, in part by the National Key Research and Development Program of China under Grant 2022YFE0116700, in part by the National Natural Science Foundation of China under Grant T2225025, in part by the Key Research and Development Programs in Jiangsu Province of China under Grants BE2021703 and BE2022768, in part by Jiangsu Province Science Foundation for Youths under Grant BK20220825.

Yuan Gao, Hui Tang, Xin Chen, Xu Ji, Huazhong Shu, and Yang Chen are with the Lab of Image Science and Technology, Key Laboratory of Computer Network and Information Integration (Ministry of Education), Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, School of Computer Science and Engineering, Southeast University, Nanjing 211189, China (e-mail: yuangao@seu.edu.cn, corinna@seu.edu.cn, xinchen@seu.edu.cn, xuji@seu.edu.cn, shu.list@seu.edu.cn, chenyang.list@seu.edu.cn). Rongjun Ge is with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China (e-mail: rongjun.ge@nuaa.edu.cn). Jin Liu is with the College of Computer and Information, Anhui Polytechnic University, Wuhu 241000, China (e-mail: liujin@ahpu.edu.cn). Yan Xi is with the Jiangsu First-imaging Medical Equipment Co., Ltd., Nantong 226100, China (e-mail: yanxi@first-imaging.com). Jian Zhu is with the Cancer Hospital Affiliated to Shandong First Medical University (Shandong Cancer Institute, Shandong Cancer Hospital), Jinan 250117, China (e-mail: zhujian.cn@163.com). Gouenou Coatrieux is with the Information and Image Processing Department, Institut Mines-Telecom, Telecom Bretagne, Brest 29238, France (e-mail: gouenou.coatrieux@telecom-bretagne.eu). Jean-Louis Coatrieux is with the Centre de Recherche en Information Biomédicale Sino-Francais, Inserm, University of Rennes 1, Rennes 35042, France (e-mail: jean-louis.coatrieux@univ-rennes1.fr).

First author: Yuan Gao, Co-first author: Hui Tang, Corresponding author: Yang Chen, Co-corresponding author: Jian Zhu.

The opinions expressed here are entirely that of the author. No warranty is expressed or implied. User assumes all risk.

***Index Terms*—3D reconstruction, deep learning, spine, CT, x-ray.**

I. Introduction

SPINE is the axial skeleton of the human body and the pillar of the body, with functions such as load-bearing, shock absorption, protection, and movement. Spinal disease is among the most common conditions in the clinic [1]. When the spine is damaged due to accidents, diseases, aging, etc., the diagnosis and treatment of spinal orthopaedic injuries should be carried out in time. Specific treatment indications are as follows: 1. Spinal fractures damage the spinal cord and peripheral spinal nerves. 2. Burst fractures of the lumbar, thoracic, and cervical spine cause vertebral or spinal instability. 3. The fracture is accompanied by dislocation or space-occupying lesion [2]. Orthopedic spine treatment is a typical and challenging bone tissue treatment plan, with difficulties such as high trauma risk and long treatment time [3]. It requires high doctors' clinical experience, especially the planning ability of orthopaedic treatment programs [4]. Doctors usually make treatment programs based on the spine imaging data, because imaging scans can reveal detailed information of the spine.

Human skeletal tissues have big density differences and natural contrasts, which are very suitable for disease inspection using x-ray imaging [5]. X-ray imaging technology has the advantages of simple operation, fast imaging and low cost [6]. However, due to the overlapping of tissue structures in x-ray images, it is difficult for doctors to accurately complete the planning of orthopaedic treatment programs [7]. CT imaging technology can provide coronal and sagittal tomographic images without tissue overlapping [8]. Because CT scanning equipment is expensive, it is not as popular as two-dimensional x-ray imaging equipment in poor regions of the world. In terms of imaging speed, CT scans usually takes several minutes, while x-ray scans can complete imaging scans in seconds. CT also hurt patients more than x-ray images in terms of radiation dose [9]. In some areas that lack 3D CT imaging equipment or in emergency situations, doctors lack 3D imaging equipment or lack sufficient time to obtain CT of patients as a basis for planning preoperative orthopaedic treatment programs [10]. Doctors can only get 2D x-ray spine images, such as the anteroposterior and lateral images of the spine. It is valuable

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3296838

to devise a new method capable of converting 2D spine x-ray images to 3D spine CT.

For methods of 3D reconstruction, matching surface points between multiple views is hugely challenging using dense reconstruction methods due to the transparent nature of x-ray images [11]. Based on the human spine skeleton with a stable anatomical structure, a deep learning method of converting 2D spinal x-ray images into 3D spinal CT images is designed and named 3DSRNet. Generative adversarial network (GAN) [12] is adopted as the leading architecture of 3DSRNet. Approaches employing similar architectures have been used in many medical imaging tasks [13], [14], [15], [16], [17]. In addition, x-ray images suffer from the severe overlap of internal body information. Orthogonal x-ray images are used as input to the algorithm because they can retain numerous geometric constraint information based on the dual-view imaging mode. We improve the reconstruction ability by designing some deep learning algorithms. Firstly, 3DSRNet adopts transformer [18] to enhance the learning ability of generator network. Secondly, this method makes full use of spine texture features containing local structural information and global statistical knowledge. Thirdly, some loss functions suitable for the 3D spine reconstruction application are customized for 3DSRNet. Overall, 3DSRNet can obtain rich spine feature information to achieve an excellent performance of 3D reconstruction. The contributions of this study can be listed as follows:

1. A novel deep learning-based 3DSRNet is designed to perform 3D spine reconstruction using 2D orthogonal x-ray images.

2. A CNN combined transformer learnable framework for spine reconstruction (SRCT) is proposed, which can effectively integrate the advantages of CNN focusing on the local bone surface information and transformer focusing on the global skeleton structural information.

3. 3DSRNet extracts pixel-level image texture features that can be used for spine reconstruction tasks by using a texture extraction (SRTE) method. It helps to obtain low-level feature information from spine images.

4. In order to enhance the spine 3D reconstruction capability of 3DSRNet, some customized loss functions are used that can help the model achieve sharp reconstruction results and avoid blurring.

II. Related Works

3D reconstruction is a process of reconstructing 3D information from single-view or multi-view images. This technology is fundamental to many applications, such as robot navigation, object recognition, and scene understanding, 3D modeling and animation, industrial control, and medical diagnosis. In recent years, some research results have been achieved in the fields of natural images and medical images.

A. Natural Image

For natural images, the 3D reconstruction based on images mainly includes Structure from Motion (SfM) and Multi-View System (MVS). SfM is an overall strategy for 3D reconstruction from unordered image collections [19]. MVS estimates the dense representation from overlapping images and aims to recover the dense 3D structure of a scene from a set of calibrated images [20]. SfM and MVS are the fundamental problems in computer vision and have been extensively studied for decades, because of their wide applications in 3D reconstruction, augmented reality, autonomous driving, and robotics [21].

Related methods for SfM: Cui et al. [22] proposed a hybrid SfM method to tackle the issues of efficiency, accuracy, and robustness in a unified framework. This study used an adaptive community-based rotation averaging method first to estimate camera rotations in globally. Camera centers were computed in an incrementally way based on these estimated camera rotations. Zhu et al. [23] proposed a global SfM to enhance the efficiency and robustness of large-scale motion averaging. This study divided all images into multiple partitions that preserve strong data association for well-posed and parallel local motion averaging. They proposed an internal signal averaging that determines cameras at partition boundaries, and a similarity transformation per partition to register all cameras in a single coordinate frame. Finally, local and global motion averaging were iterated until convergence. Cui et al. [24] proposed a progressive SfM method to tackle the completeness, robustness and efficiency problems in a united framework, where two loops are contained. The outer loop is a feature matching loop, where the orthogonal MSTs (maximum spanning trees) of the image similarity graph are iteratively selected to perform the image matching. The inner loop is an incremental camera calibration loop, where the initial camera poses in each iteration are inherited from those calibrated in the last one. Novotni et al. [25] used SfM to generate a supervisory signal from videos. The approach generated a partial point cloud and the relative camera parameters based on a video sequence. The different depth estimates were fused, using the estimated camera parameters, into a partial point cloud, which is further processed for completion utilizing a PointNet [26].

Related methods for MVS: Huang et al. [27] proposed a DeepMVS, a deep convolutional neural network for multi-view stereo reconstruction. Taking an arbitrary number of posed images as input, they first produced a set of plane-sweep volumes and used the proposed DeepMVS network to predict high-quality disparity maps. Yao et al. [28] suggested the MVSNet for depth map inference from multi-view images. In the network, depth visual image features were extracted to build the 3D cost volume upon the reference camera frustum via the differentiable homograph warping. 3D convolutions were applied to regularize and regress the initial depth map, which was then refined with the reference image to generate the final output. Chen et al. [29] proposed a Point-MVSNet, a novel point-based deep framework for multi-view stereo. This network leveraged 3D geometry priors and 2D texture information jointly and effectively by fusing them into a feature-augmented point cloud. It processed the point cloud to estimate the 3D flow for each point. Chen et al. [30] designed a 3D reconstruction system based on the multi-view technology. The system takes a variety of views of an object as input, and finally outputs a 3D model of the object.

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3296838

The above methods usually use depth data such as depth maps, point clouds, voxels, meshes, etc. to solve these 3D reconstruction problems. All objects appearing in the scene need to obtain their depth information and establish a global model of the scene. These 3D reconstruction methods are generally not suitable for medical reconstruction tasks. For example, obtaining depth data of a scene requires complex camera calibration, but it is not convenient to perform accurate camera calibration for medical imaging scenes.

B. Medical Image

For medical images, 3D reconstruction has been developed in recent years to aid in medical diagnosis. Tognola et al. [31] implemented image reconstruction from 2D CT scans to a 3D model for the mandibular bone for quantitative measurements on the 3D triangular mesh. Chen [32] proposed a novel FPP-based equipment to obtain dense 3D point clouds and preserve tooth details for intraoral 3D points reconstruction. Kasten et al. [33] designed an end-to-end CNN approach for the 3D reconstruction of knee bones directly from orthogonal x-ray images. In contrast to the standard statistical modeling approach, this method can learn the shape distribution of bones directly from the training images. Shen et al. [34] demonstrated the reconstruction approach of upper-abdomen, lung, and head-and-neck via deep learning. It can map projection radiographs of a patient to the corresponding 3D anatomy and subsequently generate volumetric tomographic x-ray images of the patient from a single projection view. Tognola et al. [35] obtained 3D models of the mandibular and maxilla bones and of the mandibular nerve by segmenting CT scan images of patients undergoing maxillofacial surgery. Henzlen et al. [36] devised a deep CNN method to produce whole 3D skull volumes. This study proposed firstly to learn a coarse and fixed-resolution volume which is then fused in a second step with the input x-ray into a high-resolution volume. Ying et al. [37] proposed reconstructing lung CT from two orthogonal x-ray images using a feature fusion framework based on the generative adversarial network. This novel feature fusion method was proposed to combine information from two x-ray images and increase data dimension from 2D to 3D. Aubert et al. [38] proposed an automated 3D spine reconstruction method through which a realistic statistical shape model of the spine is fitted to images using CNN. The CNNs automatically detected the anatomical landmarks controlling the spine model deformation through a hierarchical and gradual iterative process. This method is suitable for fast 3D visualization of spine structures. But a small number of marked locations are not enough to reveal the intricate details of the spinal surface. Ge et al. [39] designed the X-CTRSNet that simultaneously and accurately enables 3D cervical vertebra CT reconstruction and segmentation directly from orthogonally 2D x-ray images. This work combined the reciprocally coupled SpaDRNet for reconstruction, MulSISNet for segmentation, and an RSC Learning for task consistency. Although each method has its own characteristics, none of them can produce reconstruction results to recover the 3D geometry and structure of the spine.

III. Method

This paper aims to reconstruct 3D spine CT using 2D orthogonal projection x-ray images. We adopt a generative adversarial framework to accomplish the task of 3D spine reconstruction based on orthogonal 2D X-ray images. In this cross-modal application scenario, our method allows the generator and discriminator to compete in the training phase, which is beneficial to skillfully reconstruct rich spine 3D details through a self-learning mechanism. The architecture of the reconstruction network is shown in Fig.1. The human spine organ has texture differences in different regions. Based on this essential, SRCT and SRTE are proposed to exploit global features and low-level texture information, respectively, to enhance the reconstruction ability. In SRCT, the transformer is introduced to take full advantage of global skeleton characteristics. In SRTE, the texture extraction is used to enhance the utilization of spine texture features. These methods help the 3DSRNet achieve 3D spine reconstruction. The generator network details are depicted in Fig.2. The detailed description of these methods is as follows.

A. Generator Architecture

To enhance 3D reconstruction performance, the generator adopts an encoder-decoder network architecture. The encoder-decoder network architecture is used in many image tasks, such as denoising [40], [41], and segmentation [42], [43], [44]. The encoder-decoder architecture is designed to learn a feature mapping from 2D input to 3D target. In the encoder part, CNN's part of the generator adopts many convolutions and skip connections for feature extraction of spine images. Furthermore, as shown in Fig.2, the skip connection uses a parallel structure that includes atrous spatial pyramid pooling (ASPP) [45], transformer and texture branches simultaneously to extract features of 2D input, respectively. For global spine structure information, the transformer supplements the feature extraction results of the CNN. The texture branch is used to complete the extra enhancement work of spine low-level image feature extraction. The decoder part is composed of 3D deconvolutions to realize the generation of 3D CT. In addition, in the connection part between the encoder and the decoder, the encoder feature extraction results are fused from two inputs to perform a dimensional transformation on feature maps. The dimensional shift utilizes a one-dimensional convolution operation to obtain 3D feature maps from 2D feature maps.

Although CNN can extract some 2D features, extracted features are not enough for 3D spine reconstruction in multi-scale. The ASPP is introduced to generate multi-scale spine features. ASPP is designed to concatenate multiple atrous convolutional features into a final feature representation using different dilation rates. It can connect multi-scale convolutional layers to generate features containing the overall structure of the spine and the tiny details of the vertebrae without significantly increasing the model size.

B. SRCT

CNN and transformer are different learning paradigms. CNN relies on local convolution operations, whereas the

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3296838

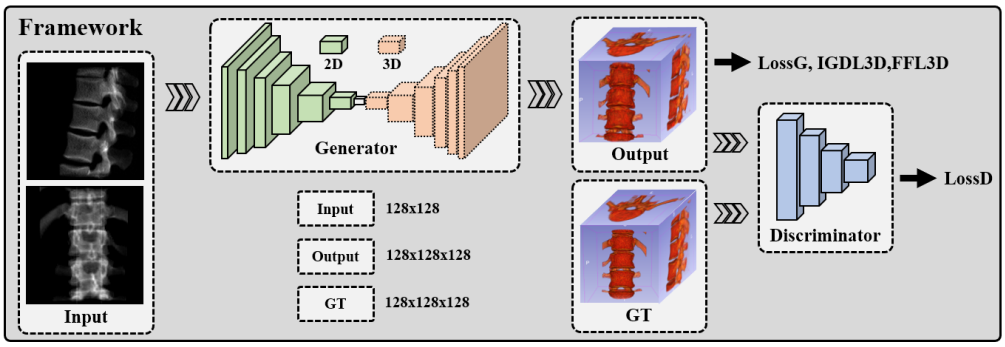


Fig. 1. The generative adversarial architecture is used as the deep learning network framework in this study. Two orthogonal projection x-ray images of 128x128 pixels as the input data. The generator consists of a 2D encoder marked with the solid line and a 3D decoder marked with the dotted line. The 3D reconstruction result of 128x128x128 pixels is output from the generator. The loss functions include LossG, IGDL3D, FFL3D, and LossD.

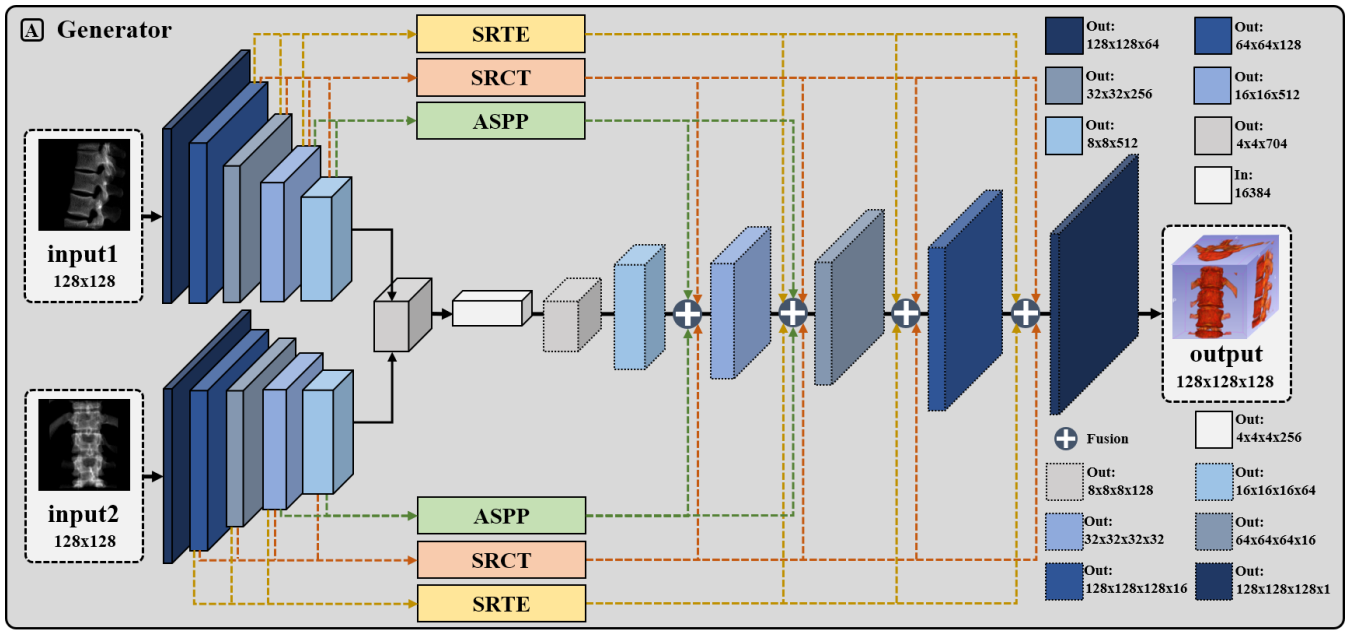


Fig. 2. The architectural details of the generator. Two parallel encoders perform feature encoding extraction on two orthogonal input images respectively. The skip connection adopts three modules (ASPP, SRCT, and SRTE) that can obtain more features of spine images. The fusion of the three modules can assist the basic generator network in achieving performance improvement. It should be noted that a linear layer marked in white is used to convert 2D feature maps to 3D feature maps.

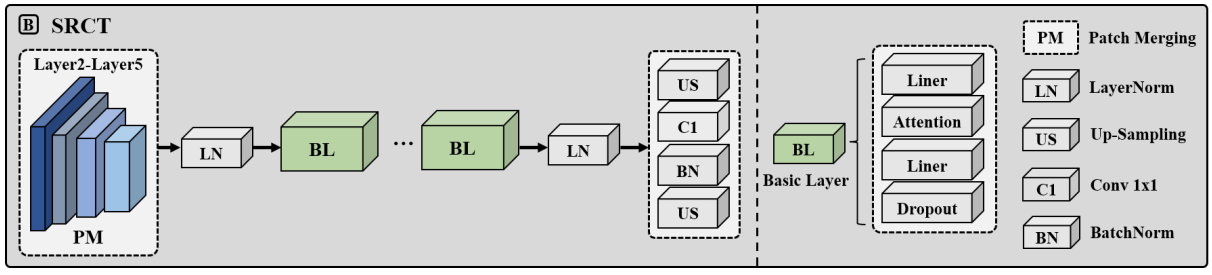


Fig. 3. The network architecture of the SRCT. The second to fifth layers of the encoder are merged as the input of SRCT. SRCT adopts multiple basic layers based on a self-attention mechanism that obtains the global representation of spine data and the up-sampling operation that obtains the same size as input.

transformer is based on long-range self-attention. For the task of spine reconstruction, the spine not only has global skeleton structure features, but also has local bone detail features. A good reconstruction method should consider both global and local features of spine data. In our framework, SRCT combines the global receptive field of Transformer with the local receptive field of CNN to obtain generalized representation of spine 3D reconstruction information.

The generator network uses a CNN encoder-decoder network, which collects different local features in a hierarchical manner similar to the pyramid structure. The feature maps of different layers are concatenated and used as the input

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3296838

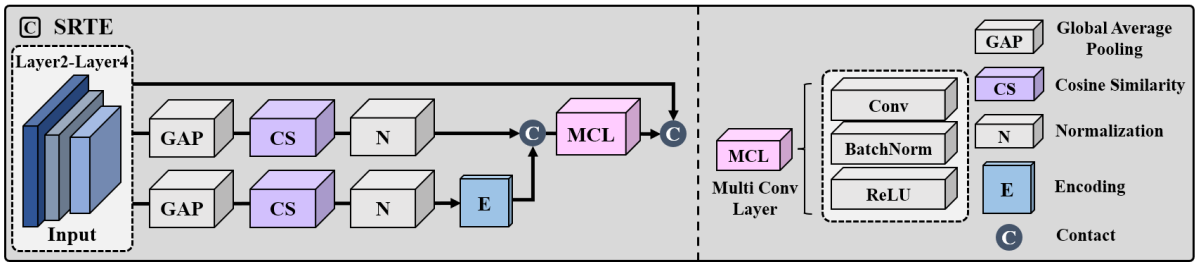


Fig. 4. The architecture details of the SRTE. The second to fourth layers of the encoder are merged into the input of SRTE. Two parallel branches complete texture feature extraction by computing the intensity information of feature maps and merging each other using multi-convolution layers. Finally, the feature map of texture information contacts with the input feature map to obtain the output.

of SRCT to reflect the spatial structure transformation and long-distance dependencies of the spine data. The feature maps from the second to fifth layers of the encoder are taken as input to the SRCT branch to obtain the global receptive field. The multiple convolution layers with small convolution kernels, normalization layers, and activation layers are used to effectively complete transformer operation as shown in Fig.3. SRCT enables the middle layer of 3DSRNet to have a larger receptive field, thereby paying more attention to the key features of global structural information.

For an input image X, the proposed approach takes the feature maps from the second to fifth layers of the encoder middle part as input to the transformer branch:

$$M_1 = f_{CNN}(X); M_2 = f_{Transformer}(M_1^{middle}) \quad (1)$$

where the input M_1^{middle} is with the feature maps of CNN middle part, M_1 and M_2 are the prediction feature map from two networks of CNN (f_{CNN}), and transformer ($f_{Transformer}$), respectively.

Based on the predictions of M_1^{middle}, the transformer is combined with the multiple convolution layers with small convolution kernels, normalization layers, and activation layers as defined below:

$$f_{Transformer}(M_1^{middle}) = SoftMax(\frac{QK^T}{\sqrt{d}})V \quad (2)$$

where Q, K and V are the query, key, and value composed from M_1^{middle}, d is the dimension of Q and K. It is worth mentioning that the transformer is also only used to supplement CNN training, not to produce final predictions alone.

C. SRTE

For spine 3D reconstruction, the high-level feature information that mainstream feature extraction deep learning methods focus on is not enough. Since using only deep convolution misses some key low-level features, 3D spine reconstruction results from deep high-level features extracted from large receptive fields may result in coarse and inaccurate. These key low-level features in multiple layers provide the image texture features necessary for 3D reconstruction. Image texture features contain local detail information and global structural information [46], [47], [48]. SRTE enables the generator network to focus on both high-level properties and low-level information of the image.

SRTE extracts the image texture information of the spine in the generator network. This is a neural network mechanism based on convolutional operations that can extract low-level image information. The feature maps of each convolutional layer contain texture information of different scales. This texture information can be extracted by encoding the feature maps. The encoding of features is performed by calculating the strength of each convolutional layer feature. To encode features to utilize image essence information, SRTE is inspired by [49] to quantitatively represent and extract different feature channels in a designed module. Matrix calculations are used in multiple convolutional layers to prevent the generation of encoding and reduce the effect of noise on the image features. The architecture is shown in Fig.4.

Specifically, the extracted feature maps I of the first three layers of the generator's encoder are concatenated and downsampled to the same size. Afterward, different scale texture details of these down sampled features I_{down} are extracted through multiple convolutional layers to get the cosine similarity matrix M_{cos} in the texture module. Finally, the original feature maps $M_{network}$ of the generator's encoder are concatenated with the extracted texture feature maps to obtain I_{cat} which is inputted into the generator network to get the final 3D reconstruction result. The operation formulas are defined below:

$$M_{cos}^{i,j} = \frac{I_{down} \cdot I_{i,j}}{\parallel I_{down} \parallel \cdot \parallel I_{i,j} \parallel} \quad (3)$$

$$M_{nor}^{n} = \frac{max(M_{cos}) - min(M_{cos})}{N} \times n + min(M_{cos}) \quad (4)$$

where the nth level M_{nor}^{n} is normalized by equally dividing N points between the minimum and maximum values of M_{cos} which can be quantized into N levels.

$$E_{i,j,n} = 1 - \parallel M_{nor}^{n} - M_{cos}^{i,j,n} \parallel \quad (5)$$

$$I_{cat} = Cat(F(Cat(M_{nor}, E)), M_{network}) \quad (6)$$

where $E_{i,j,n}$ is the encoding value of the M_{nor} and M_{cos}. M_{nor} and E are fused by function F. Cat denotes concatenation operation. F contains multiple convolutional layers and a ReLU activation layer.

D. Loss Functions

To train the 3D spine reconstruction model efficiently and stably, several loss functions are customized to constrain the

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3296838

training of 3DSRNet. These loss functions can calculate the error between actual values and predicted values from multiple perspectives in the gradient, frequency, and image domains. The detailed description is as follows.

Inspired by enhancing underwater imagery [50], a loss function is applied to penalize the 3D reconstruction model by directly computing the image gradient differences. This strategy can achieve sharpening the spine reconstruction results and avoiding the blurring problem. The 3D image gradient difference loss is chosen as the objective function and is given by Eq (7):

$$\begin{aligned} L_{IGDL3D}(X^g, X^r) = \sum_k \sum_{i,j} (&||X^g_{i,j,k} - X^g_{i-1,j,k}| \\ &- |X^r_{i,j,k} - X^r_{i-1,j,k}||^a \\ &+ ||X^g_{i,j-1,k} - X^g_{i,j,k}| \\ &- |X^r_{i,j-1,k} - X^r_{i,j,k}||^a) \end{aligned} \tag{7}$$

where X^r is the real CT volume of the spine, X^g is the generated data of the spine, a is an integer greater than or equal to 1. The loss function uses i, j and k to iterate over the entire data by length, width, and height.

The differences exist between the real and generated spine images, especially in the frequency domain. Narrowing the frequency domain difference can improve spine image reconstruction quality. A focal frequency loss function [51] is introduced, which allows the network to adaptively focus on difficult-to-synthesize frequency components. This loss function is complementary to the existing spatial loss and provides a large resistance to the loss of important frequency information due to the inherent bias of neural networks. We modify it to make it capable of 3D data calculation as follows:

$$F(u,v,w) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} \sum_{z=0}^{O-1} f(x,y,z) e^{-i2\pi(\frac{ux}{M} + \frac{vy}{N} + \frac{wz}{O})} \tag{8}$$

$$W(u,v,w) = |F_r(u,v,w) - F_f(u,v,w)|^a \tag{9}$$

$$L_{FFL3D} = \frac{1}{MNO} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} \sum_{w=0}^{O-1} W(u,v,w) |F_r(u,v,w) - F_f(u,v,w)|^2 \tag{10}$$

where $f(x,y,z)$ is the pixel value of the spine data, $F_r(u,v,w)$ is the spatial frequency value at spectrum coordinate (u,v,w) of the real data, $F_f(u,v,w)$ means the fake reconstruction result and a is the scaling factor for flexibility. FFL3D is used to calculate the weighted average of frequency distance between the real data and the fake reconstruction result.

Since the learning process of 2D x-ray images to 3D CT is a non-linear mapping, the generated 3D results should be consistent with the semantic information provided by the input 2D x-ray images. The learning process of 3DSRNet is an adversarial process, where the discriminator D and the generator G compete. To avoid vanishing gradient in training stage, the mean squared error (MSE) loss function is introduced. The loss function is defined as follows:

$$L_{GAN}(G) = \frac{1}{2n} \sum \| \, y_n - G(x_n) \, \|^2 \tag{11}$$

where x is the input of 2D x-ray images, and y is the 3D CT volume.

To optimize the final 3D reconstruction results, we effectively fuse all loss functions of the generator. The total generator loss function is formulated as follows:

$$L_{total}(G) = w_1 L_{GAN}(G) + w_2 L_{IGDL3D} + w_3 L_{FFL3D} \tag{12}$$

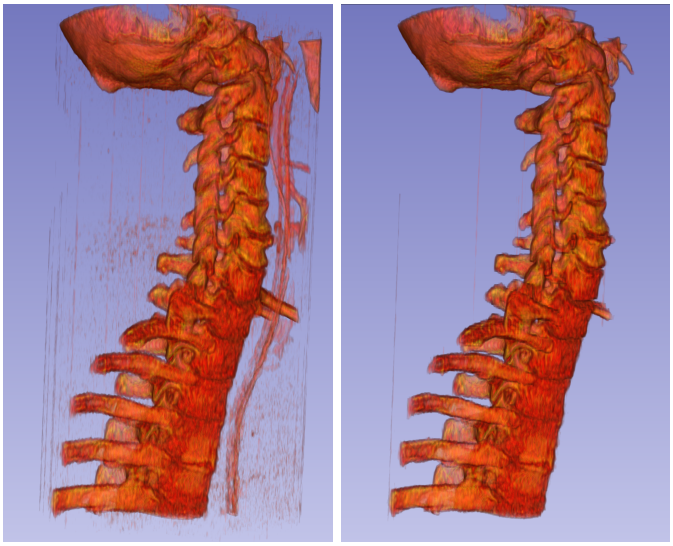
where w_1, w_2 and w_3 mean the weight factors controlling the different generator loss functions. In this study, the values of w_1, w_2 and w_3 are 0.5, 0.25 and 0.25, respectively.

The discriminator loss function is formulated as follows:

$$\begin{aligned} &L_{GAN}(D) = \\ &\frac{1}{2n} \sum \| \, D(y|x) - 1 \, \|^2 + \frac{1}{2n} \sum \| \, D(G(x)|x) \, \|^2 \end{aligned} \tag{13}$$

IV. Experimentals And Results

A. Experiments Settings



(a) (b)

Fig. 5. CT separation preprocessing. It can effectively remove the soft tissue and preserve the skeletal tissue. (a) The original spine CT without separation preprocessing. (b) The separation preprocessing effect of the spine CT.

1) Data: For the accurate CT, we introduce some public datasets including the VerSe'20 [52], VerSe'19 [52], MelihAslan [53], and BenMicrosoft [54]. A total of 729 spine CT scans were selected from these datasets. Since CT and x-ray data are difficult to obtain simultaneously in clinical, the projection processing of real CT is used to obtain simulated 2D x-ray images. Digital Reconstructed Radiograph (DRR) method [55] from the Insight Toolkit [56] is used to obtain 1458 simulated projection x-ray images. Each pair of orthogonal x-ray images and the corresponding CT serve as training data for the spine reconstruction algorithm. We collect 585 sets of data for training, 72 sets of data for validating, and 72 sets of data for testing.

2) Metrics: The peak signal-to-noise ratio (PSNR) [57], structural similarity index (SSIM) [58], Cosine Similarity (CS), mean absolute error (MAE), mean squared error (MSE) [59], and Learned Perceptual Image Patch Similarity (LPIPS)

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3296838

[60] are calculated to evaluate the reconstruction performance of the model. PSNR is used to measure the difference between two data and calculated as below:

$$MSE = \frac{1}{MNO} \sum_{m=1,n=1,o=1}^{M,N,O} [F_{real}^{m,n,o} - F_{fake}^{m,n,o}]^2 \quad (14)$$

$$PSNR(F_{real}, F_{fake}) = 10 \times \log_{10} \left(\frac{MaxValue^2}{MSE} \right) \quad (15)$$

where MSE means squared error of two data and $MaxValue$ is the maximum value of the image pixel.

SSIM is used to measure the similarity of two data in brightness, contrast, and structure. Its formula is defined below:

$$SSIM(F_{real}, F_{fake}) = \frac{(2\mu_{real}\mu_{fake} + c_1)(2\tau_{real}\tau_{fake} + c_2)}{(\mu_{real}^2 + \mu_{fake}^2 + c_1)(\sigma_{real}^2 + \sigma_{fake}^2 + c_2)} \quad (16)$$

where μ means average value, σ means standard deviation value, τ means covariance value, and c means variables to stabilize the division with a weak denominator.

CS measures the similarity between two data by calculating the cosine value of the angle between them. The formula can be defined as below:

$$\begin{aligned} CS(F_{real}, F_{fake}) &= \cos(\theta) \\ &= \frac{F_{real} \cdot F_{fake}}{\| F_{real} \| \times \| F_{fake} \|} \\ &= \frac{\sum_{i=1}^{n} F_{real}^i \times F_{fake}^i}{\sqrt{\sum_{i=1}^{n} (F_{real}^i)^2} \times \sqrt{\sum_{i=1}^{n} (F_{fake}^i)^2}} \end{aligned} \quad (17)$$

where F_{real}^i and F_{fake}^i represent the components of data F_{real} and F_{fake}, respectively.

LPIPS measures the similarity of image data by extracting deep convolution features. Deep features outperform other metrics on many datasets [60]. LPIPS obtains the distance between reference and distorted patches x, x_0 with network F. The deep features $\hat{y}^l$,$\hat{y}_0^l$ are extracted from layers L. The activations channel-wise are scaled by the vector w^l to compute the l_2 distance. Finally, LPIPS averages spatially and sums channel-wise as below:

$$LPIPS(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \| w_l \odot (\hat{y}_{hw}^l - \hat{y}_{0hw}^l) \|_2^2 \quad (18)$$

3) Implementation Details: Due to the use of different image acquisition devices in different datasets, the image pixel values at similar bone positions in different datasets will be different. For this cross-dataset situation, our research performs Gamma correction on the input image during data preprocessing to automatically avoid the above problems. The formulation of the Gamma correction is shown as follows:

$$V_{\text{out}} = AV_{\text{in}}^{\gamma} \quad (19)$$

$$\gamma = \frac{\log_{10} 0.01}{\log_{10}(m/255)} \quad (20)$$

where V_{out} is the output result, V_{in} is the input image, A is a constant, γ is the value of gamma, and m is the average value of the processed image.

A CT separation preprocessing is employed to separate skeletons from the original CT data for this study. Specifically, CT data contains many soft tissues without skeletons. Because 3DSRNet is aimed at assisting skeletal diagnosis and treatment, retaining only the skeletons in the CT data is beneficial for 3DSRNet to complete accurate reconstruction. A K-means clustering algorithm is used to complete the skeletal separation preprocessing. A comparison example of separation preprocessing results is shown in Fig.5. After the separation preprocessing, DRR technology generates the orthographic projection images from CT volumes.

All data is randomly divided into the training set, the validation set, and the test set at 8:1:1. The projection images and CT are resized to 128×128 and 128×128×128 pixels. All experiments are implemented using PyTorch on one NVIDIA 3090 GPU. We train 3DSRNet for 200 epochs and validate it every 10 epochs. In the training stage, the stochastic gradient descent (SGD) optimizer with a weight decay of 0.001 and a momentum of 0.9 is used to optimize all models. The batch-size is 2 on GPU. The learning rate starts at 0.01 and reduces by a factor of 10 after 80 and 140 epochs.

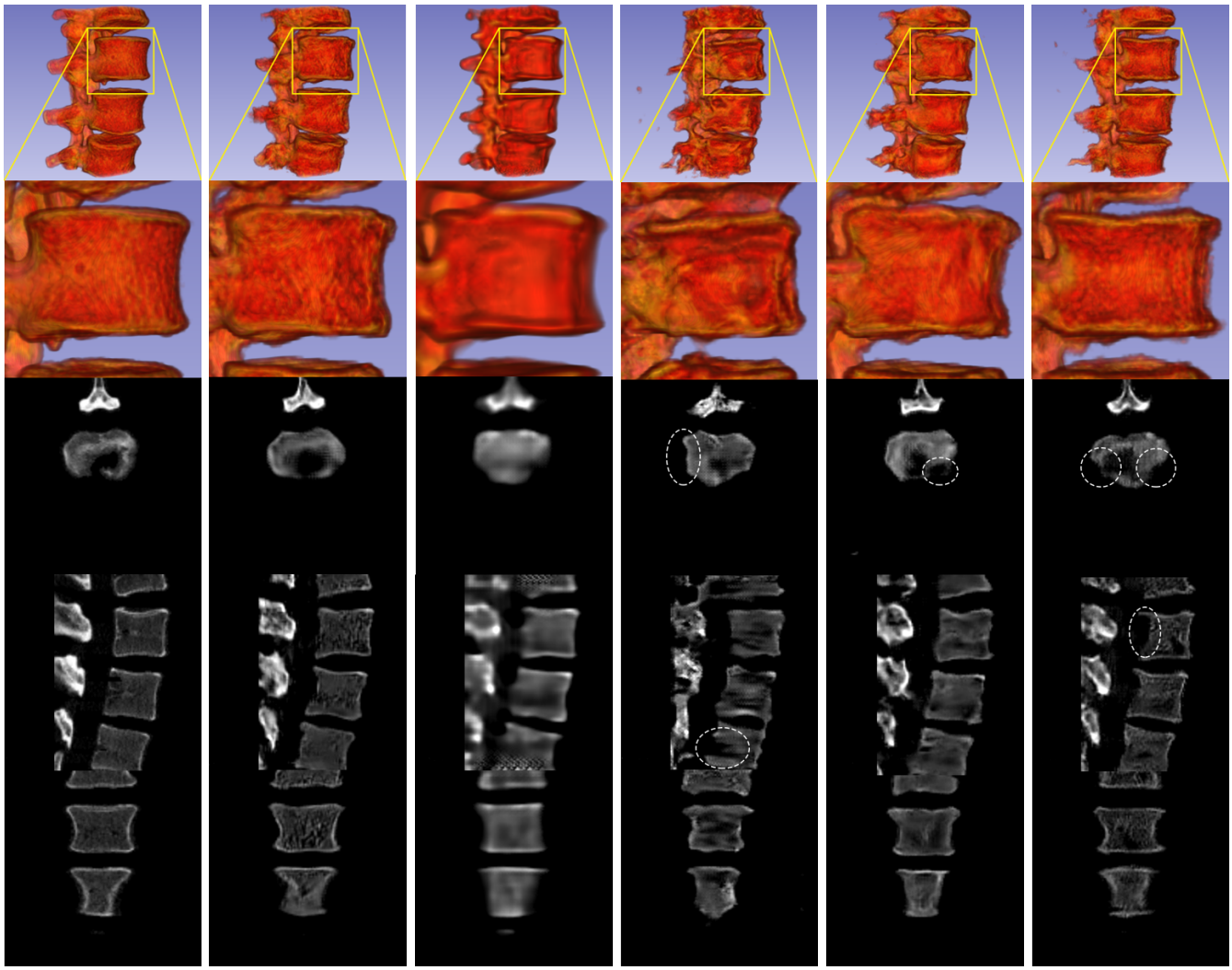
B. Evaluation on Spine Dataset

To validate the performance of 3DSRNet, four mainstream algorithms are selected to implement the comparison experiments, including PSR [34], SIT [36], X2CT [37], and ETE [33]. PSR and SIT are based on a single x-ray as the input data, while X2CT and ETE use orthogonal x-ray images as the input data in the same way as our method.

1) Quantitative Results: Experiments were performed on the spine dataset for all methods. According to the characteristics of the 3D reconstruction task, we are committed to fully concentrating the advantages of SRCT and SRTE methods in the generator of 3DSRNet. The data in Table I show that 3DSRNet achieves the best performance of 3D spine reconstruction from 2D x-ray images in all evaluation metrics. The reconstruction performance of 3DSRNet yielded PSNR of 45.4666 dB, SSIM of 0.8850, CS of 0.7662, MAE of 23.6696, MSE of 9016.1044, and LPIPS of 0.0768 in the test dataset. Experimental results show that 3DSRNet can reconstruct excellent 3D spine CT results to provide image assistance for orthopedic surgeons during diagnosis. Furthermore, since our algorithm has the most complex structure, FLOPs are the largest compared to other algorithms. The performance metrics and the runtime metric (FLOPs) of all methods are presented in Table I.

Compared with X2CT and ETE, 3DSRNet can achieve better results based on orthogonal x-ray images for all indicators. The main architecture of 3DSRNet adopts a generative adversarial network structure, which ensures the adaptability of the reconstruction network to this cross-modal task. In terms of spine reconstruction information extraction capability,

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3296838



(a) (b) (c) (d) (e) (f)

Fig. 6. 3D visualization results on spine dataset. These results demonstrate that 3DSRNet has good performance in 3D spine reconstruction. (a) represents the ground truth. (b), (c), (d), (e), and (f) represent the visualization results obtained via 3DSRNet, PSR, SIT, X2CT, and ETE, respectively. The first row to the fifth row respectively represents a 3D view, a partially enlarged 3D view, an axial view, a sagittal view, and a coronal view.

TABLE I

QUANTITATIVE RESULTS. THE QUANTITATIVE COMPARISON PERFORMANCE OF 3DSRNET AS SHOWN IN THE LAST ROW OF TABLE I. THE PERFORMANCE OF THE METHOD IS POSITIVELY CORRELATED WITH THE METRICS OF PSNR, SSIM, AND CS, AND NEGATIVELY CORRELATED WITH THE METRICS OF MAE, MSE, LPIPS AND FLOPS.

Method	PSNR	SSIM	CS	MAE	MSE	LPIPS	FLOPs
PSR	35.2928	0.759	0.6203	37.2106	15509.27	0.2169	111.631G
SIT	33.2308	0.7669	0.581	37.9719	16118.3614	0.1541	123.674G
X2CT	43.699	0.8692	0.6966	27.1507	12251.6784	0.0948	623.345G
ETE	43.958	0.8546	0.6466	29.8947	14428.335	0.1229	208.175G
3DSRNet	**45.4666**	**0.885**	**0.7662**	**23.6696**	**9016.1044**	**0.0768**	785.979G

TABLE II

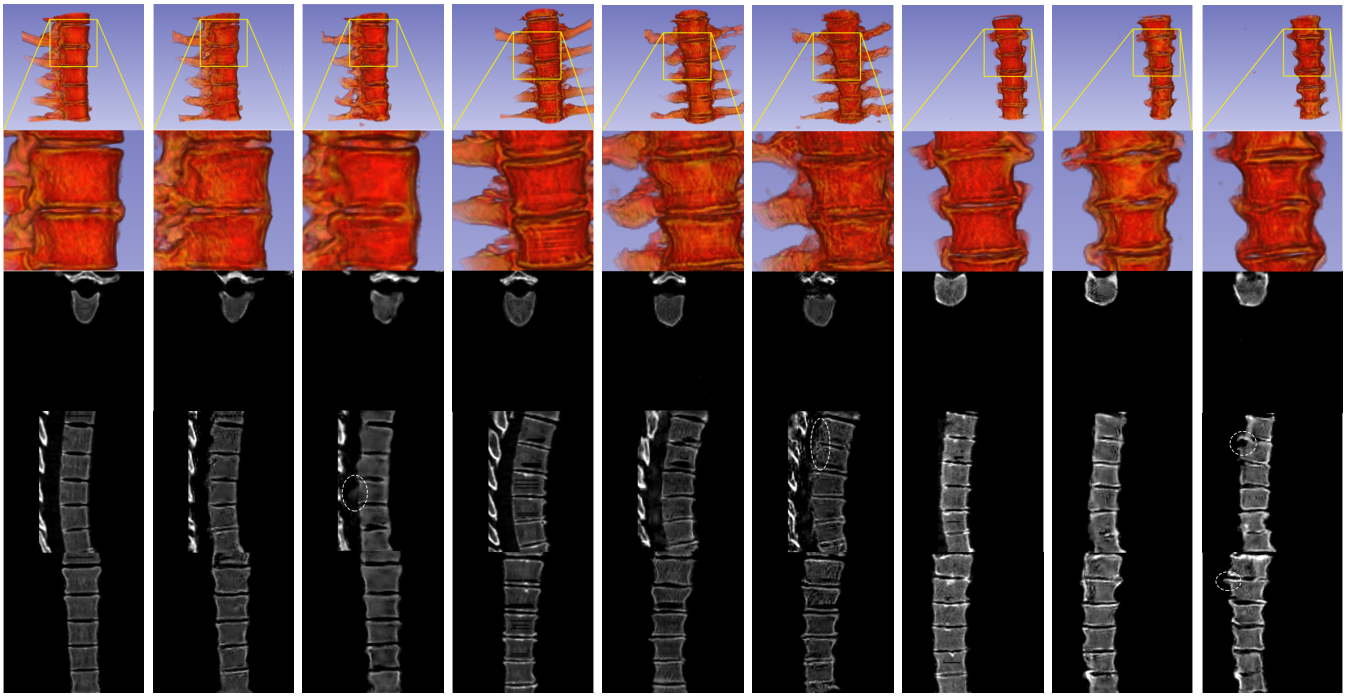
ABLATION EXPERIMENTS OF ALGORITHMS. THE ABLATION STUDY OF 3DSRNET FOR ALGORITHMIC MODULES. THE SYMBOL '+' AND '-' ARE USED TO REPRESENT ADDING AND REMOVING.

ASPP	SRCT	SRTE	PSNR	SSIM	CS	MAE	MSE	LPIPS
+	+	+	**45.4666**	**0.885**	**0.7662**	**23.6696**	**9016.1044**	**0.0768**
-	+	+	43.9712	0.87	0.7117	27.3234	12235.5622	0.1047
+	-	+	42.8017	0.8522	0.6480	30.4998	14604.7407	0.1059
+	+	-	43.3599	0.8594	0.6634	29.577	14501.7391	0.1095

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3296838

TABLE III
ABLATION EXPERIMENTS OF LOSS FUNCTIONS. THE ABLATION STUDY OF 3DSRNET FOR LOSS FUNCTION. THE SYMBOL '+' MEANS ADDING THIS LOSS FUNCTION. THE SYMBOL '-' MEANS REMOVING THIS LOSS FUNCTION.

IGDL3D	FFL3D	PSNR	SSIM	CS	MAE	MSE	LPIPS
+	+	**45.4666**	**0.885**	**0.7662**	**23.6696**	**9016.1044**	**0.0768**
-	+	44.072	0.8642	0.6811	28.5959	13466.5542	0.0993
+	-	44.2018	0.8691	0.7068	27.9719	12194.5319	0.1031



(a) (b) (c) (d) (e) (f) (g) (h) (i)

Fig. 7. 3D visualization of the evaluation of algorithms. These comparison cases show that ASPP, SRCT, and SRTE can improve the reconstruction ability, respectively. (a), (d), and (g) represent the ground truth of different original CT. (b), (e), and (h) represent 3DSRNet. (c), (f), and (i) represent the results of removing ASPP, SRCT, and SRTE, respectively. The first row to the fifth row respectively represents a 3D view, a partially enlarged 3D view, an axial view, a sagittal view, and a coronal view.

3DSRNet implements multi-scale feature extraction to improve reconstruction accuracy, benefiting from the use of an encoder-decoder generator network equipped with SRCT and SRTE. Compared with PSR and SIT which use a single x-ray image as model input, the reconstruction ability of 3DSRNet is better due to exploiting more geometric constraint information from orthogonal x-ray input images.

2) Qualitative Results: 3DSRNet provides 3D spine reconstruction results based on orthographic projection images. We qualitatively evaluate the 3D spine reconstruction results as shown in Fig.6, where visualization images are used to show the difference between the proposed method and other methods. Fig.6 (a) demonstrates the ground truth of the experimental target. Fig.6 (b) is the result of our proposed method. Fig.6 (c) to Fig.6 (f) are the results of comparison methods. There are missed skeleton image details and false skeleton structure information in the results of comparison methods. To demonstrate the effect of spine reconstruction, 3D, axial, sagittal, and coronal views of the reconstruction results are visualized. Areas with significant differences have been marked with dashed lines.

Compared with other methods, 3DSRNet has better reconstruction performance, as shown in Fig.6. For the dual-input methods including X2CT and ETE, 3DSRNet can obtain a more complete and flatter surface in terms of spine skeleton details shown in the second row of Fig.6. For single-input methods including PSR and SIT, 3DSRNet has apparent advantages in visual performance, including sharpness and detail integrity shown in the last three rows of Fig.6. Overall, the reconstruction effect of the proposed method is the best because it focuses on richer texture details and more intact spine structure information.

C. Ablation Experiments

As mentioned above, some innovative components are utilized including ASPP, SRCT, SRTE, IGDL3D, and FFL3D. To validate the effectiveness of each component, we conduct a series of ablation experiments on 3DSRNet in different settings. For better presentation, the ablation experiments are divided into two categories, including the algorithm experiments and the loss function experiments, as shown in Table II and Table III, respectively.

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3296838

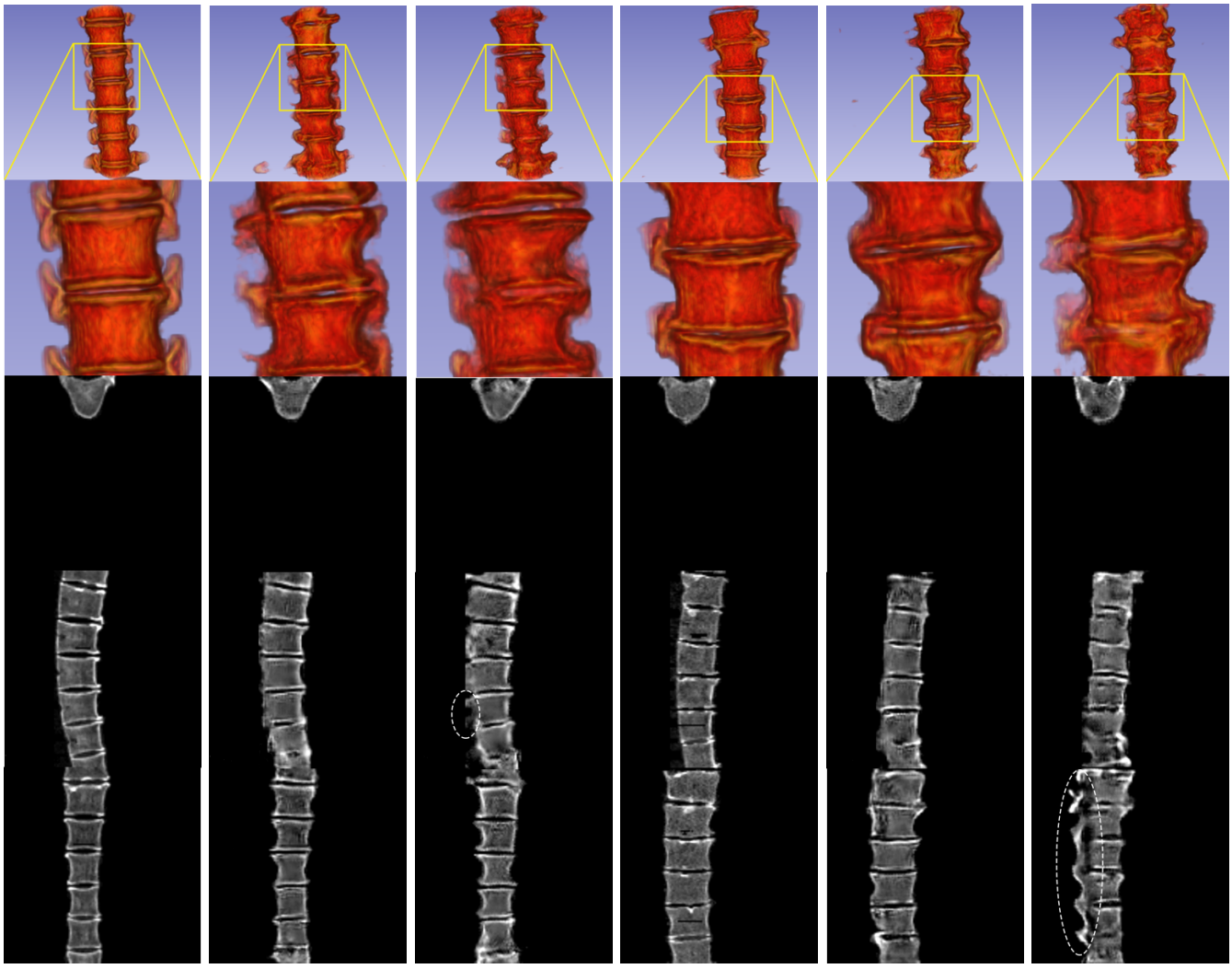
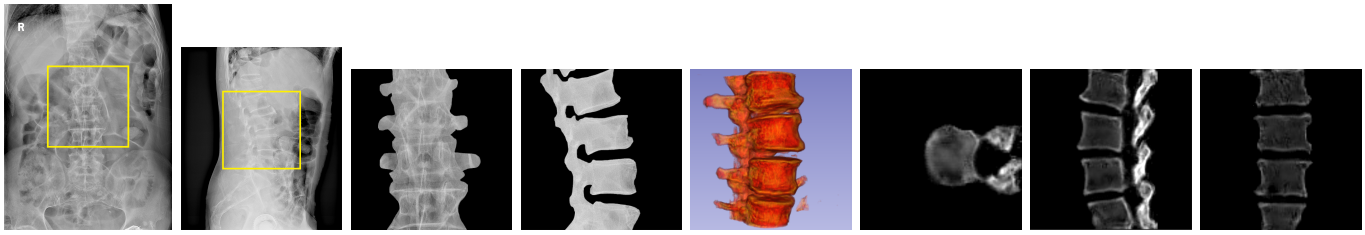


Fig. 8. 3D visualization of the evaluation of loss functions. These comparison cases show that the loss function of IGDL3D and FFL3D can improve reconstruction capabilities, respectively. (a) and (d) represent the ground truth. (b) and (e) represent 3DSRNet. (c) and (f) represent the results of removing the IGDL3D and FFL3D, respectively. The first row to the fifth row respectively represents a 3D view, a partially enlarged 3D view, an axial view, a sagittal view, and a coronal view.

1) Evaluation For Algorithmic Modules: We compare the effects of ASPP, SRCT, and SRTE as shown in Table II and Fig.7. By using ASPP, the CNN branch can use multiple parallel atrous convolutional layers with different sampling rates. Multi-scale spine information is extracted by using convolution layers with various receptive fields. Compared with 3DSRNet, if we remove ASPP, the performance is worse by 3.4%, 1.72%, 7.66%, 15.44%, 35.71%, and 36.33% on PSNR, SSIM, CS, MAE, MSE, and LPIPS, respectively. By using SRCT, it can be observed that using the transformer can reconstruct richer spine image details. It demonstrates that the joint training of CNN and transformer can promote the overall learning effect in the 3D reconstruction model. Removing SRCT decreases performance by 5.86%, 3.71%, 15.43%, 28.86%, 6.2%, and 37.89% on PSNR, SSIM, CS, MAE, MSE and LPIPS, respectively. By using SRTE, low-level texture features can be extracted to generate 3D spine results. It complements the generator's extraction of textured structures. Compared with 3DSRNet, removing SRTE can reduce the performance by 4.63%, 2.94%, 13.42%, 24.96%, 60.84%, and 42.58% on PSNR, SSIM, CS, MAE, MSE, and LPIPS, respectively.

2) Evaluation For Loss Function: We compare the loss function effect of IGDL3D and FFL3D as shown in Table III and Fig.8. The loss function of IGDL3D is used to directly penalize the difference in image gradient predictions in the generator to sharpen the generated results. Compared with 3DSRNet, if the IGDL3D is removed, the performance is worse by 3.07%, 2.35%, 11.11%, 20.81%, 49.36%, and 34.24% on PSNR, SSIM, CS, MAE, MSE, and LPIPS, respectively. The loss function of FFL3D directly optimizes the 3D spine reconstruction training stage in the frequency domain. It adaptively focuses the model on frequency components to improve the quality of the generator's reconstruction results. Compared with 3DSRNet, if the FFL3D is removed, the performance is worse by 2.78%, 1.8%, 7.75%, 18.18%, 35.25%, and 34.24% on PSNR, SSIM, CS, MAE, MSE, and LPIPS, respectively.

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3296838



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 9. 3D visualization results on the clinical data sample. The algorithm can reconstruct the rich bone structure and details of the spine. (a) and (b) represent clinical spine x-rays. (c) and (d) represent the regions marked in yellow and used as input data for 3DSRNet. (e) represents a 3D view of the result. (f-h) represent the results of an axial view, a sagittal view, and a coronal view.

D. Evaluation on Clinical Spine X-rays

In addition to experiments on the spine dataset, the experiment on clinical spine x-ray sample is also performed. The frontal and lateral images of real spine x-rays are shown in Fig. 9(a) and (b). Due to the large size of the clinical x-ray image, a part of the image is intercepted as the input data of the algorithm. Regions marked in yellow are resized to 128×128 pixels and used as input data for 3DSRNet. Clinical x-rays contain a large amount of non-skeletal image information, which is meaningless for the 3D reconstruction of spine. The skeletal separation is used to remove non-skeletal image information, as shown in Fig. 9(c) and (d). The reconstructed 3D results are shown in Fig. 9(e)-(h). Since the existing clinical x-ray image is difficult to obtain the patient's 3D CT reconstruction data at the same time, the 3D spine reconstructed by the method in this paper cannot be directly compared with the real 3D CT reconstruction data. But in terms of the visibility of the reconstruction result, the reconstruction result has rich bone shape and details. In general, the algorithm proposed in this paper can achieve 3D reconstruction of spinal bones, and have a good prospect in clinical applications related to spine orthopedics. In future work, we will focus on collecting 3D CT data corresponding to clinical x-ray data to improve our method.

V. CONCLUSIONS

This study proposed a 3DSRNet using 2D projection x-ray images to reconstruct 3D spine CT based on deep learning. 3DSRNet can use some designs to capture rich feature information based on the generative adversarial architecture. Its generator integrates the benefits of many components, including SRCT, SRTE, IGDL3D, and FFL3D. SRCT is used to obtain rich information about the spine. There are properties that CNN can effectively capture the local features of the bone images and transformer can extract the global structural relationships of the spine skeleton. These properties of CNN and transformer can complement each other learnedly during the training stage of SRCT. SRTE is devised to extract low-level features of the spine surface textures. The texture presents irregularities in the local area of the bone, but continuous regularity in the overall spine bone. SRTE can integrate these properties, which are useful for assisting 3D reconstruction into the model. The IGDL3D and FFL3D compose the loss function of 3DSRNet to obtain the spine reconstruction details more effectively. The IGDL3D sharpens 3D reconstruction predictions by directly penalizing differences in image gradient predictions in the generator. FFL3D enables 3DSRNet to optimize image 3D reconstruction directly in the frequency domain.

Experimental results show that these designs can improve the 3D spine reconstruction performance. The reconstruction performance indicators of 3DSRNet achieve PSNR of 45.4666 dB, SSIM of 0.8850, CS of 0.7662, MAE of 23.6696, MSE of 9016.1044, and LPIPS of 0.0768. Compared with many mainstream algorithms, the proposed method exhibits a better performance in the 3D reconstruction effect. Ablation experiments show that 3DSRNet suffers performance degradation in 3D reconstruction on experimental samples after removing different designs. In the future, we hope to collect more data to train the model to explore performance improvements. Furthermore, we also hope that 3DSRNet will be evaluated for application value in further clinical studies.

REFERENCES

[1] N. J. Manek and A. MacGregor, "Epidemiology of back disorders: prevalence, risk factors, and prognosis," *Current opinion in rheumatology*, vol. 17, no. 2, pp. 134–140, 2005.

[2] J. Park, D.-W. Ham, B.-T. Kwon, S.-M. Park, H.-J. Kim, and J. S. Yeom, "Minimally invasive spine surgery: techniques, technologies, and indications," *Asian spine journal*, vol. 14, no. 5, p. 694, 2020.

[3] M. D'Souza, J. Gendreau, A. Feng, L. H. Kim, A. L. Ho, and A. Veeravagu, "Robotic-assisted spine surgery: history, efficacy, cost, and future trends," *Robotic Surgery: Research and Reviews*, vol. 6, p. 9, 2019.

[4] C. D. Vo, B. Jiang, T. D. Azad, N. R. Crawford, A. Bydon, and N. Theodore, "Robotic spine surgery: current state in minimally invasive surgery," *Global Spine Journal*, vol. 10, no. 2_suppl, pp. 34S–40S, 2020.

[5] R. Müller, H. Van Campenhout, B. Van Damme, G. Van der Perre, J. Dequeker, T. Hildebrand, and P. Rüegsegger, "Morphometric analysis of human bone biopsies: a quantitative structural comparison of histological sections and micro-computed tomography," *Bone*, vol. 23, no. 1, pp. 59–66, 1998.

[6] A. U. Ibrahim, M. Ozsoz, S. Serte, F. Al-Turjman, and P. S. Yakoi, "Pneumonia classification using deep learning from chest x-ray images during covid-19," *Cognitive Computation*, pp. 1–13, 2021.

[7] M. D. Fishman and M. M. Rehani, "Monochromatic x-rays: The future of breast imaging," *European Journal of Radiology*, vol. 144, p. 109961, 2021.

[8] Y.-B. Park, H.-S. Jeon, J.-S. Shim, K.-W. Lee, and H.-S. Moon, "Analysis of the anatomy of the maxillary sinus septum using 3-dimensional computed tomography," *Journal of oral and maxillofacial surgery*, vol. 69, no. 4, pp. 1070–1078, 2011.

[9] T. R. Goodman, A. Mustafa, and E. Rowe, "Pediatric ct radiation exposure: where we were, and where we are now," *Pediatric radiology*, vol. 49, no. 4, pp. 469–478, 2019.

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3296838

[10] C. Meng, T. Wang, W. Chou, S. Luan, Y. Zhang, and Z. Tian, "Remote surgery case: robot-assisted teleneurosurgery," in *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, vol. 1. IEEE, 2004, pp. 819–823.

[11] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[13] K. Bahrami, F. Shi, I. Rekik, and D. Shen, "Convolutional neural network for reconstruction of 7t-like images from 3t mri using appearance and anatomical features," in *Deep learning and data labeling for medical applications*. Springer, 2016, pp. 39–47.

[14] N. Burgos, M. J. Cardoso, F. Guerreiro, C. Veiga, M. Modat, J. McClelland, A.-C. Knopf, S. Punwani, D. Atkinson, S. R. Arridge *et al.*, "Robust ct synthesis for radiotherapy planning: application to the head and neck region," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 476–484.

[15] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with context-aware generative adversarial networks," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 417–425.

[16] Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern Recognition*, 2018, pp. 9242–9251.

[17] Y. Zhang, D. Hu, T. Lyu, J. Zhu, G. Quan, J. Xiang, G. Coatrieux, S. Luo, and Y. Chen, "Pie-arnet: Prior image enhanced artifact removal network for limited-angle dect," *IEEE Transactions on Instrumentation and Measurement*, 2022.

[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[19] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[20] Y. Furukawa, C. Hernández *et al.*, "Multi-view stereo: A tutorial," *Foundations and Trends® in Computer Graphics and Vision*, vol. 9, no. 1-2, pp. 1–148, 2015.

[21] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-scale data for multiple-view stereopsis," *International Journal of Computer Vision*, vol. 120, no. 2, pp. 153–168, 2016.

[22] H. Cui, X. Gao, S. Shen, and Z. Hu, "Hsfm: Hybrid structure-from-motion," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1212–1221.

[23] S. Zhu, R. Zhang, L. Zhou, T. Shen, T. Fang, P. Tan, and L. Quan, "Very large-scale global sfm by distributed motion averaging," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4568–4577.

[24] H. Cui, S. Shen, W. Gao, and Z. Wang, "Progressive large-scale structure-from-motion with orthogonal msts," in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 79–88.

[25] D. Novotny, D. Larlus, and A. Vedaldi, "Capturing the geometry of object categories from video supervision," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 2, pp. 261–275, 2018.

[26] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[27] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang, "Deepmvs: Learning multi-view stereopsis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2821–2830.

[28] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 767–783.

[29] R. Chen, S. Han, J. Xu, and H. Su, "Point-based multi-view stereo network," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1538–1547.

[30] X. Chen, Q. Wu, and S. Wang, "Research on 3d reconstruction based on multiple views," in *2018 13th International Conference on Computer Science & Education (ICCSE)*. IEEE, 2018, pp. 1–5.

[31] G. Tognola, M. Parazzini, G. Pedretti, P. Ravazzani, C. Svelto, M. Norgia, and F. Grandori, "Three-dimensional reconstruction and image processing in mandibular distraction planning," *IEEE transactions on instrumentation and measurement*, vol. 55, no. 6, pp. 1959–1964, 2006.

[32] S. Chen, "Intraoral 3-d measurement by means of group coding combined with consistent enhancement for fringe projection pattern," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[33] Y. Kasten, D. Doktofsky, and I. Kovler, "End-to-end convolutional neural network for 3d reconstruction of knee bones from bi-planar x-ray images," in *International Workshop on Machine Learning for Medical Image Reconstruction*. Springer, 2020, pp. 123–133.

[34] L. Shen, W. Zhao, and L. Xing, "Patient-specific reconstruction of volumetric computed tomography images from a single projection view via deep learning," *Nature biomedical engineering*, vol. 3, no. 11, pp. 880–888, 2019.

[35] G. Tognola, M. Parazzini, G. Pedretti, P. Ravazzani, F. Grandori, A. Pesatori, M. Norgia, and C. Svelto, "Gradient-vector-flow snake method for quantitative image reconstruction applied to mandibular distraction surgery," *IEEE Transactions on instrumentation and measurement*, vol. 58, no. 7, pp. 2087–2093, 2009.

[36] P. Henzler, V. Rasche, T. Ropinski, and T. Ritschel, "Single-image tomography: 3d volumes from 2d cranial x-rays," in *Computer Graphics Forum*, vol. 37, no. 2. Wiley Online Library, 2018, pp. 377–388.

[37] X. Ying, H. Guo, K. Ma, J. Wu, Z. Weng, and Y. Zheng, "X2ct-gan: reconstructing ct from biplanar x-rays with generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 619–10 628.

[38] B. Aubert, C. Vazquez, T. Cresson, S. Parent, and J. A. de Guise, "Toward automated 3d spine reconstruction from biplanar radiographs using cnn for statistical spine model fitting," *IEEE transactions on medical imaging*, vol. 38, no. 12, pp. 2796–2806, 2019.

[39] R. Ge, Y. He, C. Xia, C. Xu, W. Sun, G. Yang, J. Li, Z. Wang, H. Yu, D. Zhang *et al.*, "X-ctrsnet: 3d cervical vertebra ct reconstruction and segmentation directly from 2d x-ray images," *Knowledge-Based Systems*, vol. 236, p. 107680, 2022.

[40] Z. Huang, J. Zhang, Y. Zhang, and H. Shan, "Du-gan: Generative adversarial networks with dual-domain u-net-based discriminators for low-dose ct denoising," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2021.

[41] M. Tajmirriahi, R. Kafieh, Z. Amini, and H. Rabbani, "A lightweight mimic convolutional auto-encoder for denoising retinal optical coherence tomography images," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–8, 2021.

[42] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "Ds-transunet: Dual swin transformer u-net for medical image segmentation," *IEEE Transactions on Instrumentation and Measurement*, 2022.

[43] L. Yang, S. Song, J. Fan, B. Huo, E. Li, and Y. Liu, "An automatic deep segmentation network for pixel-level welding defect detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–10, 2021.

[44] Z. Ling, A. Zhang, D. Ma, Y. Shi, and H. Wen, "Deep siamese semantic segmentation network for pcb welding defect detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022.

[45] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[46] K. R. Castleman, *Digital image processing*. Prentice Hall Press, 1996.

[47] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.

[48] L. G. Shapiro, G. C. Stockman *et al.*, *Computer vision*. Prentice Hall New Jersey, 2001, vol. 3.

[49] L. Zhu, D. Ji, S. Zhu, W. Gan, W. Wu, and J. Yan, "Learning statistical texture for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 537–12 546.

[50] C. Fabbri, M. J. Islam, and J. Sattar, "Enhancing underwater imagery using generative adversarial networks," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7159–7165.

[51] L. Jiang, B. Dai, W. Wu, and C. C. Loy, "Focal frequency loss for image reconstruction and synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 919–13 929.

[52] A. Sekuboyina, M. E. Husseini, A. Bayat, M. Löffler, H. Liebl, H. Li, G. Tetteh, J. Kukačka, C. Payer, D. Štern *et al.*, "Verse: A vertebrae labelling and segmentation benchmark for multi-detector ct images," *Medical image analysis*, vol. 73, p. 102166, 2021.

.

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3296838

[53] M. S. Aslan, A. Shalaby, and A. A. Farag, "Clinically desired segmentation method for vertebral bodies," in *2013 IEEE 10th International Symposium on Biomedical Imaging*. IEEE, 2013, pp. 840–843.
[54] B. Glocker, D. Zikic, E. Konukoglu, D. R. Haynor, and A. Criminisi, "Vertebrae localization in pathological spine ct via dense classification from sparse annotations," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2013, pp. 262–270.
[55] N. Milickovic, D. Baltas, S. Giannouli, M. Lahanas, and N. Zamboglou, "Ct imaging based digitally reconstructed radiographs and their application in brachytherapy," *Physics in Medicine & Biology*, vol. 45, no. 10, p. 2787, 2000.
[56] T. S. Yoo, M. J. Ackerman, W. E. Lorensen, W. Schroeder, V. Chalana, S. Aylward, D. Metaxas, and R. Whitaker, "Engineering and algorithm design for an image processing api: a technical report on itk-the insight toolkit," in *Medicine Meets Virtual Reality 02/10*. IOS press, 2002, pp. 586–592.
[57] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics letters*, vol. 44, no. 13, pp. 800–801, 2008.
[58] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.
[59] D. M. Allen, "Mean square error of prediction as a criterion for selecting variables," *Technometrics*, vol. 13, no. 3, pp. 469–475, 1971.
[60] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

Yuan Gao received the B.S. degree in computer science and technology from Shandong Jianzhu University, Shandong, China, in 2017, and the M.S. degree from the control engineering with the Taiyuan University of Technology, Taiyuan, China, in 2020. He is currently working toward the Ph.D. degree in computer science and technology with Southeast University, Nanjing, China. He is with the Lab of Image Science and Technology, Key Laboratory of Computer Network and Information Integration (Ministry of Education), Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, School of Computer Science and Engineering, Southeast University, Nanjing 210096, China. His current research interests include artificial intelligence, computer vision, and medical image processing and analysis. E-mail: yuangao@seu.edu.cn

Hui Tang received the B.S., M.S., and Ph.D. degrees in computer science and technology with Southeast University, Nanjing, China, in 2003, 2005, and 2008, respectively. She is with the Lab of Image Science and Technology, Key Laboratory of Computer Network and Information Integration (Ministry of Education), Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, School of Computer Science and Engineering, Southeast University, Nanjing 210096, China. Her research interests include medical image visualization, image processing, computer-assisted surgery planning software, and computer graphics. E-mail: corinna@seu.edu.cn

Rongjun Ge received the Ph.D. degree in computer science and technology with Southeast University, Nanjing, China, in 2020. He is with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China. His current research interests include intelligent reconstruction and analysis of medical images and medical information analysis. E-mail: rongjun.ge@nuaa.edu.cn

.

Jin Liu received the Ph.D. degrees in computer science and technology with Southeast University, Nanjing, China, in 2018. He is with the College of Computer and Information, Anhui Polytechnic University, Wuhu 241000, China. His research interests include computer vision, machine learning, medical image reconstruction and 3D visual analysis. E-mail: liujin@ahpu.edu.cn

Xin Chen received the B.S. degree in IoT Engineering from Taiyuan University of Science and Technology, Shanxi, China, in 2018, and the M.S. degree from the computer application technology with Shandong Normal University, Shandong, China, in 2021. He is currently working toward the Ph.D. degree in computer science and technology with Southeast University, Nanjing, China. He is with the Lab of Image Science and Technology, Key Laboratory of Computer Network and Information Integration (Ministry of Education), Southeast University, Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, School of Computer Science and Engineering, Nanjing 210096, China. His current research interests include biomedical signal processing and digital image analysis. E-mail: xinchen@seu.edu.cn

Yan Xi received the Ph.D. degree from the Biomedical Engineering School, Shanghai Jiao Tong University, Shanghai, China, in 2013. He is with the Jiangsu First-imaging Medical Equipment Co., Ltd., Nantong 226100, China. His current research interests include x-ray computed tomography, x-ray phase-contrast imaging, and joint tomographic reconstruction. E-mail: yanxi@first-imaging.com

Xu Ji received the Ph.D. degree from the Department of Medical Physics, University of Wisconsin-Madison, Wisconsin, U.S., in 2020. He is with the Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, Southeast University, Nanjing, China, and also with the Laboratory of Image Science and Technology, the School of Computer Science and Engineering, Southeast University, Nanjing 210096, China. His current research interests lies in digital image reconstruction and processing, fousing on imaging algorithms related to x-ray biomedical imaging systems. E-mail: xuji@seu.edu.cn

Huazhong Shu received the Ph.D. degree from the Department of Mathematics, University of Rennes 1, Rennes, France, in 1992. He is with the Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, Southeast University, Nanjing, China, and also with the Laboratory of Image Science and Technology, the School of Computer Science and Engineering, Southeast University, Nanjing 210096, China. His current research interests include signal and image processing, pattern recognition. E-mail: shu.list@seu.edu.cn

Jian Zhu received the Ph.D.degree in signal processing and communication from the University of Rennes 1, Rennes, France, and computer science and technology from Southeast University, Nanjing, China, in 2013. He is currently the deputy director of the Department of Radiation Physics Technology at the Cancer Hospital Affiliated to Shandong First Medical University (Shandong Cancer Institute, Shandong Cancer Hospital), Jinan 250117, China. His current research interests include oncology, radiation physics, and medical physics. E-mail: zhu-

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3296838

Gouenou Coatrieux received the Ph.D. degree in signal processing and telecommunication from the University of Rennes I, Rennes, France, in collaboration with the Institut Mines Telecom, Telecom Paris-Tech, Paris, France, in 2002. His Ph.D. focused on watermarking in medical imaging. He is currently an Associate Professor in the Information and Image Processing Department, Institut Mines-Telecom, Telecom Bretagne, Brest, France, and his research is conducted in the Laboratoire de Traitement de l'Information Médicale, Unité INSERM U1101, Brest. His primary research interests include medical information system security, watermarking, electronic patient records, and healthcare knowledge management. E-mail: gouenou.coatrieux@telecom-bretagne.eu

Jean-Louis Coatrieux received the Ph.D. and State Doctorate degrees in sciences from the University of Rennes 1, Rennes, France, in 1973 and 1983, respectively. He is the chief researcher of the French National Institute of Health and Medical Research (INSERM), a professor in Rennes, France, and an adjunct professor at the New Jersey Institute of Technology in the United States. He is with the Centre de Recherche en Information Biomédicale Sino-Francais, Inserm, University of Rennes 1, Rennes 35042, France. His current research interests include 3-D images, signal processing, computational modeling, and complex systems with applications in integrative biomedicine. E-mail: jean-louis.coatrieux@univ-rennes1.fr

Yang Chen received the B.S. and Ph.D. degrees in biomedical engineering with First Military Medical University, Guangzhou, China, in 2001 and 2007. He is with the Lab of Image Science and Technology, Key Laboratory of Computer Network and Information Integration (Ministry of Education), Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, School of Computer Science and Engineering, Southeast University, Nanjing 210096, China. His current research interests include artificial intelligence, computer vision, and medical signal image processing and analysis. E-mail: chenyang.list@seu.edu.cn

.