DeepOrganNet: On-the-Fly Reconstruction and Visualization of 3D / 4D Lung Models from Single-View Projections by Deep Deformation Network

Yifan Wang, Zichun Zhong, and Jing Hua

Abstract—This paper introduces a deep neural network based method, i.e., DeepOrganNet, to generate and visualize fully high-fidelity 3D / 4D organ geometric models from single-view medical images with complicated background in real time. Traditional 3D / 4D medical image reconstruction requires near hundreds of projections, which cost insufferable computational time and deliver undesirable high imaging / radiation dose to human subjects. Moreover, it always needs further notorious processes to segment or extract the accurate 3D organ models subsequently. The computational time and imaging dose can be reduced by decreasing the number of projections, but the reconstructed image quality is degraded accordingly. To our knowledge, there is no method directly and explicitly reconstructing multiple 3D organ meshes from a single 2D medical grayscale image on the fly. Given single-view 2D medical images, e.g., 3D / 4D-CT projections or X-ray images, our end-to-end DeepOrganNet framework can efficiently and effectively reconstruct 3D / 4D lung models with a variety of geometric shapes by learning the smooth deformation fields from multiple templates based on a trivariate tensor-product deformation technique, leveraging an informative latent descriptor extracted from input 2D images. The proposed method can guarantee to generate high-quality and high-fidelity manifold meshes for 3D / 4D lung models; while, all current deep learning based approaches on the shape reconstruction from a single image cannot. The major contributions of this work are to accurately reconstruct the 3D organ shapes from 2D single-view projection, significantly improve the procedure time to allow on-the-fly visualization, and dramatically reduce the imaging dose for human subjects. Experimental results are evaluated and compared with the traditional reconstruction method and the state-of-the-art in deep learning, by using extensive 3D and 4D examples, including both synthetic phantom and real patient datasets. The efficiency of the proposed method shows that it only needs several milliseconds to generate organ meshes with 10K vertices, which has great potential to be used in real-time image guided radiation therapy (IGRT).

Index Terms—Deep deformation network, organ meshes, 3D / 4D shapes, 2D projections, single-view

1 Introduction

Cone beam computed tomography (CBCT) has become increasingly important in cancer radiotherapy for understanding the anatomical structure of organs and pinpointing tumors during the treatments. Integrated CBCT is an important and convenient tool for patient positioning, verification, and visualization in image guided radiation therapy (IGRT). Traditional high-quality CBCT image reconstruction requires near hundreds of projections, which consequently deliver undesired high imaging / radiation dose to patients as well. The high imaging dose to healthy organs in CBCT scans [23, 25, 46] is a crucial clinical concern. Practically, the imaging dose in CBCT can be reduced by reducing the number of X-ray projections and lowering tube voltage setting. On the other hand, due to the limited number of projections, the image quality is highly degraded in 3D-CBCT reconstructed by conventional methods, such as Feldkamp-Davis-Kress (FDK) [16] (plenty of artifacts and noises with lower accuracy). Several strategies have been proposed to enhance the image quality of reconstructed CBCT. One major kind of approaches is to use iterative image reconstruction algorithms, such as simultaneous algebraic reconstruction technique (SART) [1], total variation (TV) minimization [45], and prior image constraint techniques [4, 8]. Thus, the accuracy of subsequent 3D organ modeling does highly depend on the quality of the reconstructed images. Currently, the doctors and clinicians need to use some post-processing methods / tools to segment, reconstruct, and visualize the 3D organ models, which are quite time-consuming and cumbersome.

Recently, there is an emerging trend to generate 3D models by deep neural network in computer vision, computer graphics, and visualization communities, in which the 3D shapes can be captured and represented from the input raw data in different formats such as 3D meshes, 3D point clouds, 3D volumes, multi-view 2D images, etc. Among them, deriving the 3D shape from a single view is fundamental and very challenging. Recently, deep learning techniques have been developed to generate / reconstruct 3D shapes from a single RGB natural image (e.g., photograph) [9, 14, 28, 50]. Their 3D shape outputs from the neural network can be represented in different formats, such as a volume [9], point loud [14, 28], or surface mesh [50]. However, these methods either require complicated post-processing to generate the surface models [9, 14], or have non-manifold and invalid surface elements [50].

In order to build the bridge to directly generate the 3D shape meshes from a single 2D medical grayscale image on the fly, in this work, we present a deep neural network based method, i.e., *DeepOrganNet*, to generate and visualize high-fidelity fully 3D / 4D organ geometric models from single-view medical images, e.g., 3D / 4D-CBCT projections, by learning the smooth deformation fields based on a trivariate tensor-product deformation technique. Experimental results are evaluated and compared with the traditional reconstruction method and the state-of-the-art in deep learning, by using extensive 3D and 4D examples, including both synthetic phantom and real patient datasets. The key *contributions* of our work are as follows:

- It proposes an end-to-end deep learning method with a lightweight but effective neural network to reconstruct multiple high-fidelity 3D organ meshes with a variety of geometric shapes from a single-view medical image with complicated background and noises.
- The proposed organ reconstruction network simultaneously learns the optimal selection and the best smooth deformation from multiple templates via a trivariate tensor-product deformation technique, i.e., free-form deformation (FFD), to match the query 2D image.
- To our knowledge, it is the first time using deep learning framework to generate multiple 3D organ meshes (such as left and right lungs in our application) from a single-view medical image.

- *Yifan Wang, Zichun Zhong, and Jing Hua are with the Department of Computer Science, Wayne State University, Detroit, MI 48202. E-mail: {yifan.wang2,zichunzhong,jinghua}@wayne.edu.*
- *Corresponding author is Zichun Zhong.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

- The application and user study on IGRT demonstrate that the accurate on-the-fly tracking and reconstruction of 3D / 4D organ shapes facilitated by our method have the potential in improving the current IGRT procedure and practice.

2 Related Work

In this section, we only review some most related work on 3D shape reconstruction from single images in computer vision / graphics, visualization, and medical imaging domains.

2.1 3D Shape from Single-View Image in Computer Vision

In computer vision, graphics, and visualization, 3D reconstruction is the process of capturing the shape and appearance of real objects.

2.1.1 Traditional Learning Based Methods

Hoiem et al. [20] and Saxena et al. [39] started to use statistic and learning based approaches for 3D shape reconstruction from a single image several decades ago. Recently, Kar et al. [26] proposed to learn category-specific 3D shape models from object silhouettes and then capture intra-class shape variation from a single image. Carreira et al. [5] proposed a method to estimate the camera viewpoint using rigid structure-from-motion and then reconstruct object shapes by optimizing over visual hull proposals guided by loose within-class shape similarity assumptions. Fouhey et al. [18] demonstrated to learn their proposed primitives to infer 3D surface normals given a single image. Eigen et al. [13] presented a method to estimate and find 3D depth relations from a single stereo image by using a multi-scale deep network.

With the help of ShapeNet [7], a richly-annotated and large-scale repository of 3D CAD models, there are several 3D reconstruction approaches presented in the recent few years. For instances, Huang et al. [22] proposed a joint analysis method for shape reconstruction by estimating the camera pose, computing dense pixel-level correspondences between image patches, and finally creating a 3D model for each image by an optimization.

2.1.2 Deep Learning Based Methods

Most recently, using deep learning methods to analyze and represent 3D objects is becoming a popular trend, inspired by the successes of these techniques in 2D images and 1D texts. Choy et al. [9] proposed a 3D recurrent reconstruction neural network (3D-R2N2) to output a reconstruction of the object with a 3D occupancy grid format, which cannot well preserve the surface geometry of a 3D shape. In order to predict a nicer surface space, Fan et al. [14] explored the generative networks for 3D geometry based on a point cloud representation. Kuryenkov et al. [28] proposed a DeformNet to achieve smooth geometric deformations on point clouds for 3D shape reconstruction. However, it is well-known that a 3D point cloud may not be as efficient and effective in representing the underlying continuous 3D geometry as a 3D surface mesh. It needs some non-trivial post-processings to generate the valid surface meshes (to guarantee the manifold property).

Wang et al. [50] adopted a graph-based convolutional neural network to produce the 3D geometry by progressively deforming an ellipsoid with leveraging perceptual features extracted from an input image. However, this method can only reconstruct a single genus-0 topology shape, since the initial shapes are all deformed from an ellipsoid. Another limitation is that their deformation is defined on the surface space with a linear transformation model, which is difficult for the network to compute high-fidelity large deformation to accurately capture the shape geometry and they need several regularization terms to control the shape smoothness and local consistency. Smith et al. [44] extended Wang et al.'s work [50] by using an adaptive face splitting strategy in order to better capture the local surface geometry, but it still has the problems of having non-manifold elements and topological constraint (by using a sphere as the initial shape). The above methods are based on *surface deformation*. However, one of the major limitations of surface deformation, whose deformation field is directly defined on the shape surface, is that its computational effort and numerical robustness are highly related to the complexity and quality of the surface tessellation [3]. In the presence of the degenerate or poor quality triangles, the local transformations on these triangles are not well defined and thus lead to topological or non-manifold errors [50], as shown in Sec. 5. Even with quite some efforts for adding regularization terms, such as Laplacian regularization, edge length regularization, etc. [44, 50], it is still difficult to fully guarantee the deformation consistency in the local vertex neighborhood.

From the mathematical aspect, this problem can be avoided by *space deformation*. The key idea is to deform the ambient space (i.e., 3D volume space) enclosing the shapes, and thus implicitly deform the embedded surface shape (i.e., 2-manifold) [3]. Compared with the surface-based deformation methods, space deformation approaches apply a trivariate deformation function to transform all the points of the original surface. One major advantage of the space deformation is that it does not depend on any particular surface representation, so that it can be used to deform all kinds of explicit surface representations, such as vertices of meshes or samples of point clouds [3]. Classical free-form deformation (FFD) [40] represents the space deformation by a trivariate tensor-product spline function. Pontes et al. [35] proposed a learning framework (i.e., Image2Mesh) to reconstruct a single 3D object mesh from a 2D natural image by first deforming a selected template using the symmetric FFDs and then linearly combining a few more strongly related templates. However, their method predominantly relies on a complicated and pre-computed graph embedding of templates and their framework is not end-to-end trainable. Jack et al. [24] proposed a method to learn FFDs for multiple templates to infer a 3D shape reconstruction from a single natural image with a plain background, but their framework is also limited to generate a single object, without considering multi-object scenario.

Besides that, all the aforementioned deep learning based methods for 3D shape reconstruction from a single image are not designed and applied to medical image reconstruction and visualization.

2.2 3D Volume (Shape) from Single-View Image in Medicine

In medical image, 3D reconstruction is the process of computing the structure and tissue of real objects (not only the shape).

2.2.1 Volumetric Image Reconstruction Methods

Li et al. [31, 32] utilized a deformable image registration method to compute deformation vector fields (DVFs) for the reference of a lung motion model. Then, a principal component analysis (PCA) based lung motion model has been applied to generate a motion vector field so as to reconstruct a volumetric image and locate 3D tumor from a single CBCT / X-ray projection. The algorithm was implemented on graphics processing unit (GPU) to achieve real-time efficiency. However, there are limitations of their method, such as a linear relationship between the image intensity of the computed and measured projection images may not be accurate. Some pre-processings for DVF computation are needed. The framework settings are not fully automatical and practical for clinical use. The single-view reconstruction suffers from an ill-posed problem because only one angle data is used in the reconstruction. To alleviate this issue, Liu et al. [33] tried a wavelet-based reconstruction approach to the acquired singe-view measurements, but the reconstruction quality is still not satisfactory for clinical applications. Recently, Henzler et al. [19] proposed a convolutional encoder-decoder network to reconstruct a 3D volume from a 2D single-view cranial X-ray image. The direct coarse output is then improved to the higher resolution by a post fusion. The resulting 3D shape structure is still embedded in a 3D volume and the 3D shape can only be shown by the volume rendering with the manual-setting isosurface threshold.

2.2.2 Shape Reconstruction Methods

There are few works on directly reconstructing the 3D shapes (meshes) from medical images (grayscale pixels), since it is a cross-modality problem, which is relatively challenging. The traditional solution is to reconstruct the 3D volumetric images from multiple 2D view images at first [1,4,8,16,45], and then use image segmentation methods to extract the region of interest (ROI), such as organs or tumors; and finally generate the 3D shape meshes (i.e., isosurface) by using Marching

Cubes algorithm [34]. For instance, iso2mesh [15] is an open-source toolbox for generating 3D surficial and volumetric meshes from binary and grayscale images, but it needs to undergo the tedious procedure due to the complicated substeps.

Some researchers investigated to fill the gap to directly build the 3D shape from a limited / sparse number of 2D medical images. Fleute et al. [17] proposed to use a few X-ray images generated from a C-Arm and to build the 3D shape of the patient bones or organs by deforming a statistical 3D model to the contours segmented on the X-ray views. Tang et al. [48] used a hybrid 3D atlas shape model to reconstruct or deformably register the surface of an object from two to four 2D X-ray projections of the object. Lamecker et al. [30] presented a method to reconstruct 3D shapes from few digital X-ray images on the basis of 3D-statistical shape models; however, there are some empirical pre-processings needed, such as thickness of the shape model, silhouette extraction, etc. Sadowsky et al. [38] presented a method for volume rendering of unstructured grids, which was applied in visualizing "2D-3D" deformable registration of anatomical models. Ehlke et al. [12] proposed a novel GPU-based approach to render virtual X-ray projections of deformable tetrahedral meshes, and applied the method to improve the geometric reconstruction of 3D anatomy (e.g., pelvic bone) from few 2D X-ray images.

To our knowledge, there is no existing method to reconstruct the 3D mesh models from a single-view 2D medical image, which is a very challenging problem by using the traditional model-driven or statistical techniques. In this paper, we propose a deep learning based data-driven approach to solve this difficult but inspiring problem.

3 DEEPORGANNET

In this work, we propose an end-to-end deep neural network, named as DeepOrganNet, to generate 3D / 4D surface meshes of multiple organs from single-view medical image projections by learning the optimal deformations upon the best-selected mesh templates. This strategy not only prevents the poor quality of the reconstructed result with coarse voxelization or non-manifold surface mesh (widely existing in previous methods as discussed in Sec. 2.1.2), but also preserves fine and smooth surface details for 3D shape generation and visualization. In this section, we introduce main technique components of the DeepOrganNet model, including dataset generation, free-form deformation (FFD) on mesh, and organ reconstruction network with the loss functions.

3.1 Dataset Generation

3D object reconstruction is one of the most complicated tasks in computer vision / graphics and visualization fields, compared with object classification, segmentation, retrieval, etc., not to mention 3D object reconstruction from a single-view image. Such work usually needs a large-size dataset in support of the deep learning networks to learn the correct mapping from the 2D projection image to the corresponding 3D object. The current 3D shape reconstruction tasks mainly rely on some public large-size and well-established synthetic 3D shape dataset, such as ShapeNets [7], ModelNet10, and ModelNet40 [51]. Additionally, the input image in most tasks is restricted to the 2D projection of a certain object under an identical lighting condition with a uniform background in order to filter out as much unrelated information as possible.

Challenges: Unlike such natural-image-to-3D-object tasks, the proposed organ reconstruction from a single X-ray image (e.g., 3D / 4D-CBCT projection) is more challenging due to following aspects. First, the dataset of 3D organ shapes, such as human lungs (in this paper, we use lung organs as illustrative examples), is very limited, and there exists neither established synthetic dataset nor available clinical dataset with a reasonable scale, which can be adopted in our task. Second, an X-ray image is essentially different from a 2D natural image, which contains obvious object profile and appearance (with the clear background in most previous works); instead the X-ray image contains the structures and details inside the object or occluded from the viewpoint (with the complicated background and noises). Third, the proposed framework does reconstruct multiple objects simultaneously, such as left and right lungs as an example. However, the previous existing approaches could work on reconstructing only one object.

Generated Organ Meshes: Firstly, we address the limited dataset challenge on both 3D lung models and the corresponding 2D medical image projections. We propose a feasible strategy to generate a large number of synthetic 3D-CBCT projection images along with their corresponding 3D two-lung (left and right lungs) surface meshes with various geometric shapes by using a small amount of 3D / 4D phantom data. Given a 3D digital phantom (i.e., a volumetric image) I, we first apply the Snakes segmentation method [27] to segment the 3D lung mask images and then extract the isosurface mesh S from the segmented lung mask image by Marching Cubes algorithm [34]. After that, we employ a variety of shape deformations and spatial translations on S to get a new mesh S'. For the organ shape deformation, we first manipulate the coordinates by multiplying a scaling ratio to globally stretch or compress the lung shape in S. The scaling ratios are either constant or gradually change. We then semi-automatically apply local distortions (e.g., dents / concaves, convexes, abnormal parts) on each lung shape using Blender software tool. Both the above global and local deformations are manipulated under the guidance of our collaborative doctors to resemble and cover the real lung shape variations. All the procedures are performed based on 3D / 4D phantoms within different respiration phases to capture the real lung breathing motions. For the organ spatial arrangement, we randomly disturb the distance between each left / right lung's bounding box center and origin within a reasonable range based on the original lung positions in the phantom to resemble various real lung shape cases, and the final new mesh is S'.

Generated Volumetric Images: Once we have the deformed surface mesh S' along with the original 3D digital phantom image I, the deformed 3D digital phantom image I' can be computed. Then, we can generate the corresponding single-view (e.g., front-view) 3D-CBCT projections. It is noted that we can obtain the deformation vector field (DVF) $\Delta\mathbf{D}_V$ of the mesh vertices from surface S to S' as follows:

$$\Delta\mathbf{D}_V = \mathbf{V}_{S'} - \mathbf{V}_S, \tag{1}$$

where $\mathbf{V}_{S'}, \mathbf{V}_S \in \mathbb{R}^{N\times 3}$ are the positions of mesh vertices in S' and S and N is the number of mesh vertices. As a result, for each voxel $\boldsymbol{\alpha}_i$ in I, we can estimate its deformation vector $\Delta\mathbf{d}_{\boldsymbol{\alpha}_i}$ by incorporating the DVF of its k-nearest neighboring vertices on mesh S. Such process can be written as follows:

$$\Delta\mathbf{d}_{\boldsymbol{\alpha}_i} = \mathbf{H}_{\boldsymbol{\alpha}_i}\mathbf{J}_{\boldsymbol{\alpha}_i}\Delta\mathbf{D}_V, \tag{2}$$

where $\mathbf{J}_{\boldsymbol{\alpha}_i} \in \mathbb{R}^{K\times N}$ is an one-hot encoding matrix for the indices of the K neighbors (in this work, K is 4). $\mathbf{H}_{\boldsymbol{\alpha}_i} \in \mathbb{R}^K$ is a weight vector in which:

$$\begin{cases} h_{\boldsymbol{\alpha}_i(\kappa)} = \frac{1}{K} & \text{if } ||\mathbf{v}_\kappa - \boldsymbol{\alpha}_i|| \le \psi, \\ h_{\boldsymbol{\alpha}_i(\kappa)} = \frac{1}{K||\mathbf{v}_\kappa - \boldsymbol{\alpha}_i||} & \text{otherwise}, \end{cases} \tag{3}$$

where $\kappa \in 1, ..., K$ and ψ is the norm of the maximum DVF of mesh vertex in $\Delta\mathbf{D}_V$. We then can calculate the resulting I' with all voxels' DVF using the reconstruction method with the deformation field map [36, 52].

Generated 2D Projection Images: Finally, we apply the Siddon ray tracing algorithm [43] to generate the desired 2D front-view 3D / 4D-CBCT images of S' by tracing the path of light through voxels in the 3D volumetric image I'. To better simulate the realistic raw target CBCT projections from the digital phantom data and test the sensitivity of our method to the realistic complications, after the noise-free ray line integrals are computed according to the above ray tracing method, the noisy signal at each pixel on the CBCT projections is generated based on the noise model with Poisson and normal distributions [29, 49, 52].

In this way, we can generate as many 3D lung meshes and their corresponding 3D / 4D-CBCT medical projections as possible, which is quite crucial for our proposed data-driven deep learning framework. Accordingly, we can cover various types of abnormalities caused by lesions, injuries, or singularities through applying different kinds of global and local deformations in the dataset generation and intentionally increase the ratio of such abnormalities in the dataset to provide our network with adequate prior knowledge to deal with potential unusual cases. Fig. 1 demonstrates the flowchart of the dataset generation. More detailed configuration is in the implementation section (Sec. 4).

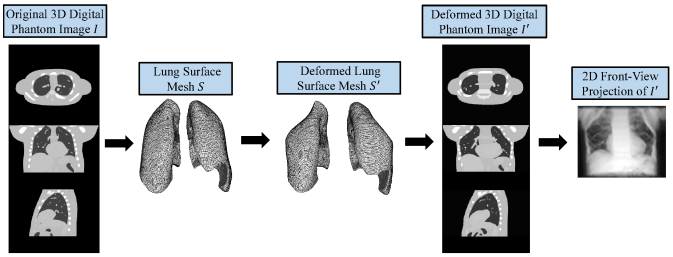

Fig. 1: The flowchart of dataset generation.

3.2 Free-Form Deformation (FFD) on Mesh

A 3D template mesh $\Omega = (\mathbf{V}, \mathbf{F})$ consists of a set of N vertices $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_N\}$ and a set of M faces $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_M\}$. A high-quality 3D mesh object usually requires dense vertices to represent fine details and thus it is computationally unfriendly, if one intends to deform it pointwisely. Instead, FFD [40] deforms the 3D mesh object through a small amount of control points. FFD introduces a 3D control point grid of size $(l+1)\times(m+1)\times(n+1)$, which encloses the target 3D mesh and performs the deformation in a trivariate tensor-product spline function, where the position of each vertex on the target mesh can be calculated:

$$\mathbf{v}(s,t,u) = \sum_{i=0}^{l}\sum_{j=0}^{m}\sum_{k=0}^{n} B_{i,l}(s)B_{j,m}(t)B_{k,n}(u)\mathbf{p}_{i,j,k}, \quad (4)$$

where $\mathbf{v}(s,t,u)$ is an arbitrary mesh vertex coordinate in the coordinate system defined by three orthogonal axes s, t, and u. $B_{p,q}(x) = \binom{p}{q}(1-x)^{p-q}x^q$ is a binomial function called Bernstein polynomial of degree q, and $\mathbf{p}_{i,j,k}$ is the control point at the node (i,j,k) on the grid. From Eq. (4), we notice that the vertex placement of the target mesh is essentially a weighted sum of the control points. Denote $\mathbf{V} \in \mathbb{R}^{N\times 3}$ as the matrix form of vertices on mesh Ω, then the mesh vertex representation can be converted:

$$\mathbf{V} = \mathbf{B}\mathbf{P}, \quad (5)$$

where $\mathbf{B} \in \mathbb{R}^{N\times\Psi}$ is the matrix form of the trivariate Bernstein tensor for all N vertices, $\mathbf{P} \in \mathbb{R}^{\Psi\times 3}$ is the matrix form of control point coordinates, and Ψ is the number of control points, i.e., $(l+1)\times(m+1)\times(n+1)$. Suppose given the displacement of these control points $\triangle\mathbf{P}$, the corresponding deformed mesh $\Omega' = (\mathbf{V}', \mathbf{F})$ in which $\mathbf{V}'$ is:

$$\mathbf{V}' = \mathbf{B}(\mathbf{P} + \triangle\mathbf{P}). \quad (6)$$

As shown in Fig. 2, in this way, the objective of the proposed deep learning network (DeepOrganNet) is to infer a $\triangle\mathbf{P}$ such that the resulting mesh S' best matches the shape of 3D lung organ surface according to the input 2D X-ray image.

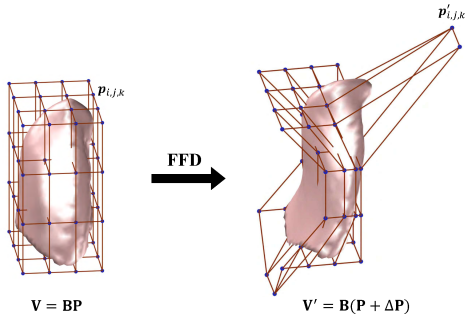

Fig. 2: FFD process on a 3D lung shape: it is deformed according to the displacement of the control points on a $4\times4\times4$ grid.

3.3 Organ Reconstruction Network

3.3.1 Network Architecture Design

The pipeline of DeepOrganNet framework consists of three functional components: feature encoder block, deformation block, and spatial arrangement block. The overall architecture is shown in Fig. 3.

Given a single-view 3D / 4D-CBCT projection image, our network first encodes it into a latent descriptor, which contains effective information for different purposes in the reconstruction stage. Due to the dataset availability and the specific training objective, the image encoder should be lightweight to alleviate the overfitting risk but still quite efficient for the reconstruction task. Jack et al. [24] has justified the adequate ability for MobileNets in the intra-class deformation, so we apply and fine-tune the pre-trained MobileNets [21] to encode the input medical image. MobileNets are compact and Inception style [47] network, which factorizes a standard convolution into a depthwise convolution and a 1×1 pointwise convolution. In addition, MobileNets introduce a width multiplier, which can reduce the width for each layer by a constant ratio and thus give us more freedom to adjust our network to best fit a relatively small amount of data in medical image scenarios. In our work, we only adopt the convolution layers in MobileNets and add a 1×1 convolutional layer after that to generate the image descriptor with a reasonable dimension. The detailed network configuration is shown in Fig. 3. Through our extensive experiments in Sec. 5, we find that the lightweight MobileNets is sufficient and robust to extract the informative features from a single-view medical image with complicated background and noises. Our input image is quite different from the one in most of the current natural-image-to-3D-object tasks, since their inputs are 2D images with clear object profile and boundary, which are generated based on the light illumination and reflection; however, our X-ray images are generated based on ray tracing technique to compute the attenuation of the energy absorption. One of the advantages in the X-ray-image-to-3D-object task is that the input X-ray images can include some shape information, which is occluded by the natural 2D images. It can make the viewer to see through the front shape surface and thus alleviate the occlusions. We will show examples in the experiment section (Sec. 5).

Furthermore, another advantage of our DeepOrganNet, compared with current natural-image-to-3D-object tasks in which the prediction is only designed for a single object, is that our task is essentially designed for reconstruction of multiple separate objects, along with a spatial arrangement between each other. As a result, we split different branches from the whole image descriptor to reconstruct different organs (i.e., different disconnected components), such as the left and right lungs independently, instead of learning all the objects together. Based on our observations and explorations, this scheme is more effective for the network to learn discriminative features from different objects than the one to learn everything together (such as using one branch scheme with a set of different two-lung templates). Each branch is responsible for the corresponding single lung generation by deforming differently-geometric templates. In this work, we use left and right lung organs as the testbed for our study, but the proposed framework can be easily extended to multi-organ (more than two organs) scenarios. Suppose we select n_l left lung templates and n_r right lung templates for corresponding branches, left lung branch learns a set of the deformation parameters for all of the n_l templates $\{\triangle\mathbf{P}_{L_i}\}_{i=1}^{n_l}$ according to the left lung shapes from the input 2D image simultaneously, where $\triangle\mathbf{P}_{L_i} \in \mathbb{R}^{\Psi\times3}$ is the deformation parameter (i.e., the control points' displacements) for a single template $L_i = (\mathbf{V}_{L_i}, \mathbf{F}_{L_i})$. The deformed template L'_i can be achieved by:

$$\mathbf{V}'_{L'_i} = \mathbf{B}_{L_i}\left(\mathbf{P}_{L_i} + \triangle\mathbf{P}_{L_i}\right), \quad (7)$$

$$\mathbf{F}'_{L'_i} = \mathbf{F}_{L_i}, \quad (8)$$

where $\mathbf{B}_{L_i}$ and $\mathbf{P}_{L_i}$ are the pre-computed transformation matrix and control point position matrix for the template L_i. The mesh connectivity does not change during the deformation as shown in Eq. (8). In addition, the left lung branch also learns a set of selection weights $\{w_{L_i}\}_{i=1}^{n_l}$ for each left lung template, which are determined along with the template deformations as shown in Eq. (13) and Eq. (15). The final left lung prediction is then selected by:

$$L_{pred} = L'_{i_{max}}, \quad (9)$$

where $i_{max} = \arg\max\{w_{L_i}\}_{i=1}^{n_l}$. Similarly, right lung branch applies the same procedure as the left lung branch to obtain the final right

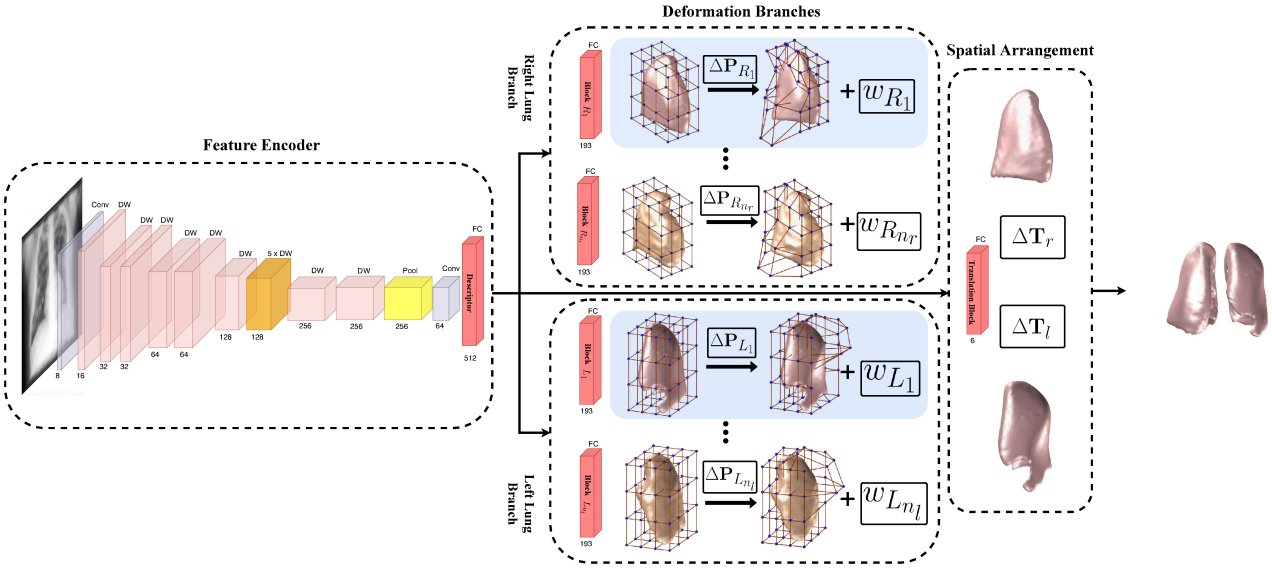

Fig. 3: The architecture of our DeepOrganNet. The DeepOrganNet first encodes the input image into a descriptor using MobileNets (without fully-connected layers) followed by a 1×1 convolution layer (dimension reduction). DW refers to the depthwise separable convolution block (two separable convolutional layers, functionally equivalent to a standard convolutional layer) and the numbers are output channel sizes (i.e., widths) of each layer / block. Every template in either left or right lung branch learns its own selection weight w and deformation parameters $\triangle\mathbf{P}$ through an independent fully-connected layer with dimension 193, including 192 for $\triangle\mathbf{P}$ and 1 for w. The deformed templates with the highest selection weights (e.g., templates L_1 and R_1 are selected in this example) in both branches are arranged according to the translation vector $\triangle\mathbf{T}_l$ and $\triangle\mathbf{T}_r$ learned from another fully-connected layer to generate the final combined multi-organ meshes.

lung prediction R_{pred}. By splitting two branches to extract the corresponding effective information from the image descriptor, the learning objectives become more specific and clearer. At this stage, the network only focuses on how to deform the templates with respect to the lung geometry from the input image. The spatial information, such as the gap / distance and the relative positions between left and right lungs, are reserved for the next stage.

As long as we have L_{pred} and R_{pred} ready, the next step is to combine them together so as to generate final left and right lung meshes with the correct relative spatial arrangement according to the input image. In order to achieve this, we learn two translation vectors $\triangle\mathbf{T}_l$, $\triangle\mathbf{T}_r$ from the image descriptor. Then the entire prediction of the new organ meshes $\Omega' = (\mathbf{V}'_{\Omega'}, \mathbf{F}'_{\Omega'})$ of both lungs is:

$$\mathbf{V}'_{\Omega'} = \left\{ \mathbf{V}'_{L_{pred}} + \triangle\mathbf{T}_l, \mathbf{V}'_{R_{pred}} + \triangle\mathbf{T}_r \right\}, \tag{10}$$

$$\mathbf{F}'_{\Omega'} = \left\{ \mathbf{F}'_{L_{pred}}, \mathbf{F}'_{R_{pred}} \right\}. \tag{11}$$

3.3.2 Loss Functions

In this subsection, we define three kinds of losses in our network not only to constrain the output shape results but also to optimize the training process.

Deformation Loss: To ensure the deformation accuracy, we choose Chamfer loss [14] to regulate the accuracy of the vertex locations on a single lung prediction. The Chamfer loss is defined as:

$$C(\mathbf{P}, \mathbf{Q}) = \sum_{\mathbf{p}\in\mathbf{P}} \min_{\mathbf{q}\in\mathbf{Q}} \|\mathbf{p}-\mathbf{q}\|_2^2 + \sum_{\mathbf{q}\in\mathbf{Q}} \min_{\mathbf{p}\in\mathbf{P}} \|\mathbf{p}-\mathbf{q}\|_2^2, \tag{12}$$

where $\mathbf{p}$ and $\mathbf{q}$ are points from two mesh vertex sets $\mathbf{P}$ and $\mathbf{Q}$. Essentially, for each point in $\mathbf{P}$ or $\mathbf{Q}$, the Chamfer loss finds the nearest vertex in the other point set and sums up all pair-wise distances. In our framework, we apply weighted Chamfer loss for both lung branches as:

$$\begin{aligned} \mathfrak{L}_{deform} = &\sum_{i=1}^{n_l} w_{L_i} C(\mathbf{V}'_{L_{pred}}, \mathbf{V}_{L_{gt}}) + \\ &\sum_{i=1}^{n_r} w_{R_i} C(\mathbf{V}'_{R_{pred}}, \mathbf{V}_{R_{gt}}), \end{aligned} \tag{13}$$

where $\mathbf{V}_{L_{gt}}$ and $\mathbf{V}_{R_{gt}}$ are the ground truth for left and right lung meshes (aligned at the origin), respectively. In this way, the proposed network is enforced to give the highest weight to the template, which can be deformed best to match the ground truth. Now, we can select the best template among all potential candidates in the datasets for predicting each organ individually and automatically.

Translation Loss: The second loss term $\mathfrak{L}_{trans}$ is intended to learn the translation vectors $\triangle\mathbf{T}_l$ and $\triangle\mathbf{T}_r$. It is defined as:

$$\mathfrak{T}_{trans} = \|\triangle\mathbf{T}_l - \mathbf{ctr}_l\|_2^2 + \|\triangle\mathbf{T}_r - \mathbf{ctr}_r\|_2^2, \tag{14}$$

where $\mathbf{ctr}_l$ and $\mathbf{ctr}_r$ are the ground truth translation vectors (i.e., two global translation vectors between the origin and the bounding box centers of left and right lungs in all ground truth meshes).

Regularization Loss: Our network deforms all templates according to input 2D images. Sometimes, the reconstruction results are achieved by tremendous deformations from a template that is not the closest one in the template pool. We introduce a weight regularization term similar to the one in [24] to encourage the network to give higher weight to the template closer to the ground truth. In this way, the overall performance of the network becomes more rational and intuitive:

$$\mathfrak{L}_w = \sum_{i=1}^{n_l} w_{L_i} \|\triangle\mathbf{P}_{L_i}\|_2^2 + \sum_{i=1}^{n_r} w_{R_i} \|\triangle\mathbf{P}_{R_i}\|_2^2, \tag{15}$$

where this loss is defined on the deformations of the control points.

The **total loss** is a weighted sum of all the above three kinds of losses as follows:

$$\mathfrak{L}_{total} = \mathfrak{L}_{deform} + \lambda_1 \mathfrak{L}_{trans} + \lambda_2 \mathfrak{L}_w, \tag{16}$$

where $\lambda_1 = 50$ and $\lambda_2 = 1$ in experimental settings, which are determined based on the corresponding order of the magnitude and balanced by the optimal network performance via our extensive experiments.

It is worth mentioning that through the strategy of integrating the deformation weights in the loss function, the proposed DeepOrganNet can automatically select the proper templates so that the network has a good prior information to start with for each organ. The risk of non-manifold issue in the reconstructed shape meshes, such as [14, 44, 50], is dramatically alleviated. In addition, FFD deforms the templates with a small amount of control points compared with vertex-wise deformation, e.g., 64 vs 10K deformation parameters, which is quite efficient.

Furthermore, FFD can realize the high-order interpolation for the deformation computations, so that the mesh surface smoothness is well maintained and no additional loss term as in [44,50] is required beyond the Chamfer loss (fidelity term) to yield a good inference.

4 Implementation Details

In this section, we introduce our dataset preparation and network training details followed by evaluation metrics which we use to measure the experimental results.

Dataset Preparation: In order to evaluate the proposed DeepOrganNet, we use following phantoms, patient studies in lung imaging and motion datasets. There are two 3D / 4D digital phantoms, i.e., a dynamic NURBS-based cardiac-torso (4D NCAT) phantom (4D images and motions are provided) and 4D extended cardiac-torso (XCAT) [41], being used as basis models to generate a reasonable number of 3D lung surface meshes and corresponding 3D / 4D-CBCT projections. They both have 10 breathing phases in 3D volumetric images (e.g., $256 \times 256 \times 150$ with voxel size of $1mm \times 1mm \times 1mm$). We generate 542 pairs of (left and right) lungs with various shapes and different spatial arrangements together with their corresponding 2D single front-view CBCT projections (see Sec. 3.1). We use the first five phases of NCAT and XCAT of 4D-CBCTs to build our training dataset and leave the rest for testing evaluation purpose. All two-lung models are normalized along the sagittal axis and translated to the origin. The bounding box size of all these models is within $1.35 \times 1.25 \times 1$ along the transverse, coronal, and sagittal axes. We then compute the bounding box centers of left and right lungs in the two-lung meshes and translate them to the origin to form the ground truth for the two deformation branches. The input 2D front-view CBCT projections are grayscale images of size 192×256 with pixel size of $1mm \times 1mm$. In the experiments, we randomly split the dataset by 446 pairs for training and 96 pairs for testing, respectively. We also test our model performance on deformable image registration (DIR)-Lab (ten lung cancer patient 4D-CT datasets with ten respiration phases each) [6], Japanese Society of Radiological Technology (JSRT) database (247 chest X-ray images) [42] for lung shape reconstruction.

Training Details: Our task is to generate left and right lung shapes from an image with noisy background and limited dataset, in order to reach a good balance between the prediction accuracy and the network overfitting risk. We set the MobileNets [21] (pre-trained on ImageNet dataset [11]) width multiplier to be 0.25 and the width (i.e., channel number) for each layer is shown in Fig. 3. For each lung branch in the network, we have two single lung templates from XCAT and NCAT, respectively. The 3D control point grids for every template from two branches are set to be $4 \times 4 \times 4$, which yields to 64 control points per template. We train the network for 65K steps using Adam optimizer with learning rate as 1×10^{-3}. The batch size is 32. The total training time is 4 hours on a single Nvidia GTX 1080 GPU with 8 GB GDDR5X.

Evaluation Metrics: Following the standard 3D shape reconstruction evaluation method, we use five different kinds of numeric metrics to evaluate the performance of our model and compare it with the existing state-of-the-art techniques.

Chamfer distance (CD) is applied in both training and testing processes. The formal expression is shown in Eq. (12). It measures bidirectional overall vertex-wise distance between two meshes.

Earth mover's distance (EMD) [37] is designed to compute the minimal sum of distances over all possible one-to-one mappings between points in $\mathbf{P}$ and points in $\mathbf{Q}$, where $\mathbf{P}$ and $\mathbf{Q}$ are two point sets of the same size. The EMD can be written as:

$$EMD(\mathbf{P}, \mathbf{Q}) = \min_{\phi:\mathbf{P}\to\mathbf{Q}} \sum_{\mathbf{p}\in\mathbf{P}} \|\mathbf{p} - \phi(\mathbf{p})\|_2 , \tag{17}$$

where ϕ is a bijection from $\mathbf{P}$ to $\mathbf{Q}$.

Hausdorff distance (HD) is adopted to measure the largest inconsistency between the reconstruction result and the ground truth. A lot of previous 3D reconstruction works did not list it as their evaluation metric because they mainly focused on point cloud reconstruction, which is insensitive to small amount of outliers. In geometric modeling and computer graphics, Hausdorff distance is a widely-used indicator to check the reconstructed mesh quality since even small amount of outliers may undermine the mesh surface consistency and quality, especially for visualization and rendering. In our experiments, we measure the Hausdorff distance between prediction and ground truth with respect to both point clouds and surface meshes [10]. Suppose $\mathbf{p}$ and $\mathbf{q}$ are the points sampled from point clouds (or surface meshes) of $\mathbf{P}$ and $\mathbf{Q}$ accordingly, the HD in terms of point clouds (or surface meshes) can be written as:

$$HD(\mathbf{P}, \mathbf{Q}) = \max\left[\max_{\mathbf{p}\in\mathbf{P}}\left(\min_{\mathbf{q}\in\mathbf{Q}}\|\mathbf{p}-\mathbf{q}\|_2\right), \max_{\mathbf{q}\in\mathbf{Q}}\left(\min_{\mathbf{p}\in\mathbf{P}}\|\mathbf{q}-\mathbf{p}\|_2\right)\right]. \tag{18}$$

F-score [50] is used as the harmonic mean of precision and recall regarding how many points in prediction or ground truth can find the nearest neighbor from the other within a threshold (ϵ). We set $\epsilon = 0.001$ in our experiments.

Intersection over union (IoU) is used to examine the volumetric similarity between the voxelized prediction and ground truth. The IoU is defined as:

$$IoU(\mathbf{P}, \mathbf{Q}) = \frac{|\mathbf{P} \cap \mathbf{Q}|}{|\mathbf{P} \cup \mathbf{Q}|}, \tag{19}$$

where $\mathbf{P}$ and $\mathbf{Q}$ are the voxelized 3D models.

Among the above five metrics, for CD, EMD and HD, the smaller the better; while for F-score and IoU, the larger the better.

5 Experiments

In this section, we conduct extensive experiments of our model on the 3D and 4D synthetic data as well as real patient data. The results are qualitatively and quantitatively compared with several state-of-the-art in deep learning based 3D shape generation (from a single-view image) methods and the traditional reconstruction method. It is noted that for comparison experiments, best results in tables are shown in bold font.

5.1 3D Lung Shape Reconstruction from Synthetic Images

Fig. 4 shows the reconstruction results of 3D lungs with different shapes based on the synthetic data. Our network is capable of dealing with drastic variations (even though the real-world medical scenarios are far less challenging). For each input image, our network is able to pick the template which most resembles the ground truth model within the corresponding branch, and predict the accurate spatial arrangement between left and right lungs to generate the final high-fidelity 3D lung shape pair. The reconstruction error (HD on mesh) is mapped into a unified colormap range and it shows that the reconstruction results are pretty good qualitatively and quantitatively.

Since our network learns the deformation parameters which are essentially applied on the control points instead of directly on the template model surface, one can generate the final 3D mesh models with arbitrary resolutions in real-time (e.g., 1K vertices: 20 ms, 2.5K vertices: 21 ms, 5K vertices: 22 ms, 10K vertices: 25 ms) according to the users' needs without re-training the network.

5.2 Comparison with Deep Learning Based Methods

Comparison with Pixel2Mesh [50]: Intuitively, according to the Pixel2Mesh (P2M) experiment setting, we first use their network to predict a single lung from a single-view input image. We train P2M on our synthetic dataset following their training details. The quantitative evaluation result is shown in Tab. 1. We fairly set our output mesh vertex number to be the same as the output of P2M network (i.e., 2466). The CD and EMD are both computed between the uniformly sampled 1024 points from the prediction and ground truth such that the comparison can be made not only between the single lung reconstruction from P2M and our network, but also between the lung pair reconstruction (see Tab. 2) and single lung reconstruction of our network. From Tab. 1, our network outperforms P2M on all metrics.

We also present the qualitative comparison results in Fig. 5 and Fig. 6. Both (P2M and our) networks yield predictions of smooth surface but our model performs better in well preserving the mesh surface geometry without non-manifold issue. The reason may be that the input single-view 3D-CBCT projection image is essentially different from a 2D natural input image. P2M has no mechanism to

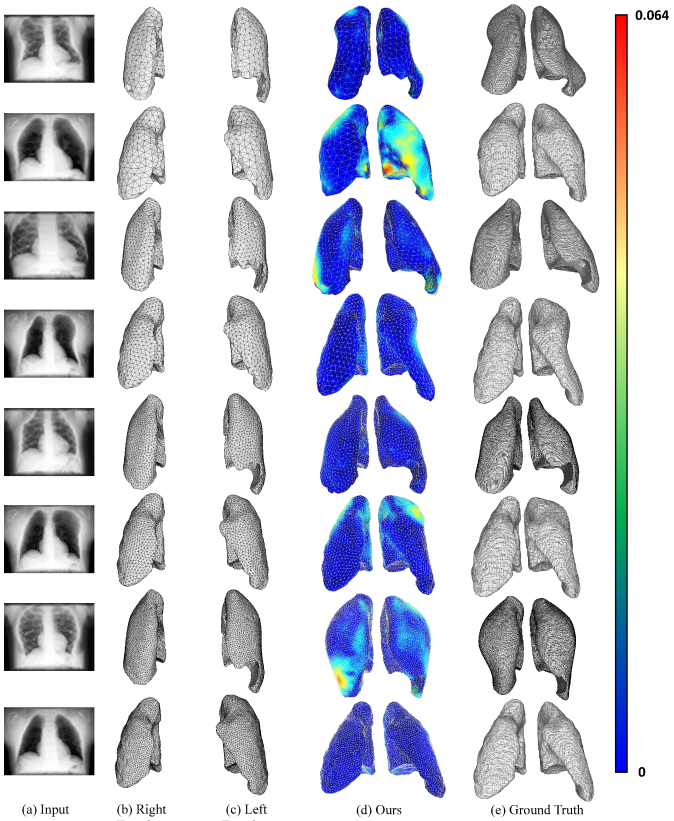

Fig. 4: Qualitative reconstruction and visualization results of some lung shapes with drastic variations. The reconstruction error (HD on mesh) is mapped into a unified colormap range (hotter colors indicate larger errors and colder colors indicate smaller errors) and the mesh resolution increases from top to bottom (e.g., 1K, 2.5K, 5K, 10K vertices).

Table 1: Quantitative comparison between P2M and our method on our synthetic dataset.

Method	CD	EMD	F-score (ϵ / 1.5ϵ)	IoU	HD (Mesh)
P2M (Left Lung)	2.4609	76.2620	0.5983 / 0.7799	0.7190	0.1300
Ours (Left Lung)	**1.7018**	**57.0856**	**0.7293 / 0.8910**	**0.8352**	**0.0672**
P2M (Right Lung)	2.3399	69.7205	0.6111 / 0.8014	0.7661	0.1022
Ours (Right Lung)	**1.7300**	**59.9497**	**0.7293 / 0.8892**	**0.8423**	**0.0786**

deal with the ambiguity caused by such ill-posed problem like PSGN. However, our network have more specific templates to start with so as to rule out some uncertainties or local minima, while the initial ellipsoid template in P2M network is too general for this task; and how to modify their network to fit for an initial lung shape is beyond the scope of this work.

We also attempt to infer two lungs together with P2M network by replacing the single ellipsoid with a pair of two ellipsoids. However, the network tends to fuse two separate lungs as a single object. The original weights for each regularization terms need to be further determined to reach a good performance. It seems to be non-trivial to extend P2M framework to multi-object reconstruction scenario.

Comparison with Point Set Generation Network [14]: We use our synthetic dataset (i.e., 542 pairs of left and right lungs) with the same training and testing splits to train the Point Set Generation Network (PSGN) [14] as our model (discussed in Sec. 4) and generate meshes from corresponding prediction point clouds using Ball Pivoting Algorithm [2]. Since our model can generate meshes with arbitrary densities, we set the output mesh vertex number as 1024 to fairly compare HD with the mesh generated from PSGN predictions. The CD and EMD are both computed between the prediction and uniformly sampled 1024 points from the ground truth (denser isosurface meshes).

Tab. 2 shows the quantitative evaluation of six different metrics and

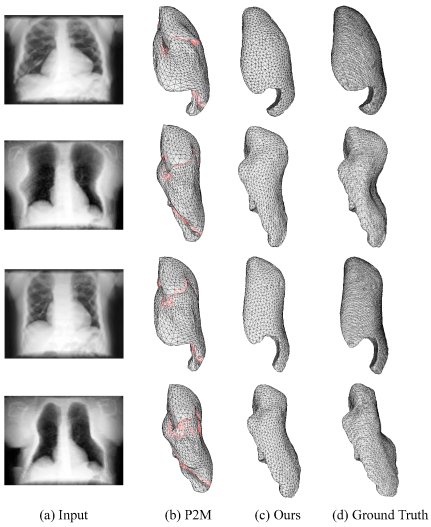

Fig. 5: Qualitative comparison between P2M and our method on left lung model. Our results generate meshes with no non-manifold issue, while the results from P2M have self-intersections (highlighted in red).

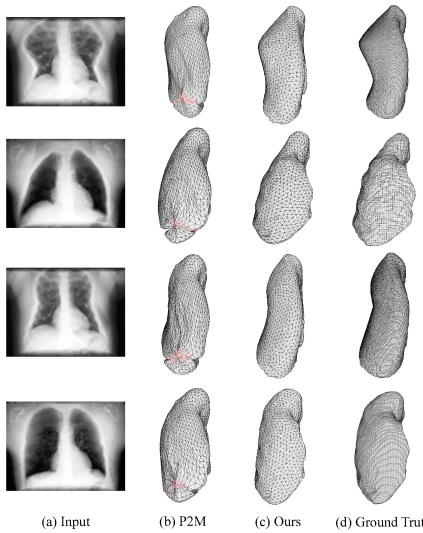

Fig. 6: Qualitative comparison between P2M and our method on right lung model. Our results generate meshes with no non-manifold issue, while the results from P2M have self-intersections (highlighted in red).

Fig. 7 provides the qualitative comparison. Our network outperforms PSGN in most metrics. In terms of point-wise HD and F-score(ϵ) evaluation, PSGN tends to get slightly better numeric results since the PSGN generates points independently, thus it has more degrees of freedom. However, the EMD of PSGN is much larger since the point cloud inference from PSGN is irregularly distributed, and sometimes the points from one lung are much denser than those from the other. Although PSGN has comparable performance in most of the point-wise evaluation metrics, it does not guarantee a high-quality 3D surface mesh. The mesh-based HD is nearly 50% higher than ours since there are a lot of meshing failures (e.g., self-intersecting triangles, holes, non-manifold triangles, etc.) and bumpy details in the generated surface meshes. In addition, PSGN learns a lung pair as a single object, when the gap between two lungs is small, the (post-processing) meshing algorithm is difficult to separate them.

Table 2: Quantitative comparison between PSGN and our method on our synthetic dataset.

Method	CD	EMD	F-score (ϵ / 1.5ϵ)	HD (Point)	IoU	HD (Mesh)
PSGN	3.0122	186.8821	**0.4384** / 0.6377	**0.0960**	0.8002	0.1491
Ours	**2.8955**	**70.7083**	0.4375 / **0.6650**	0.0980	**0.8148**	**0.1000**

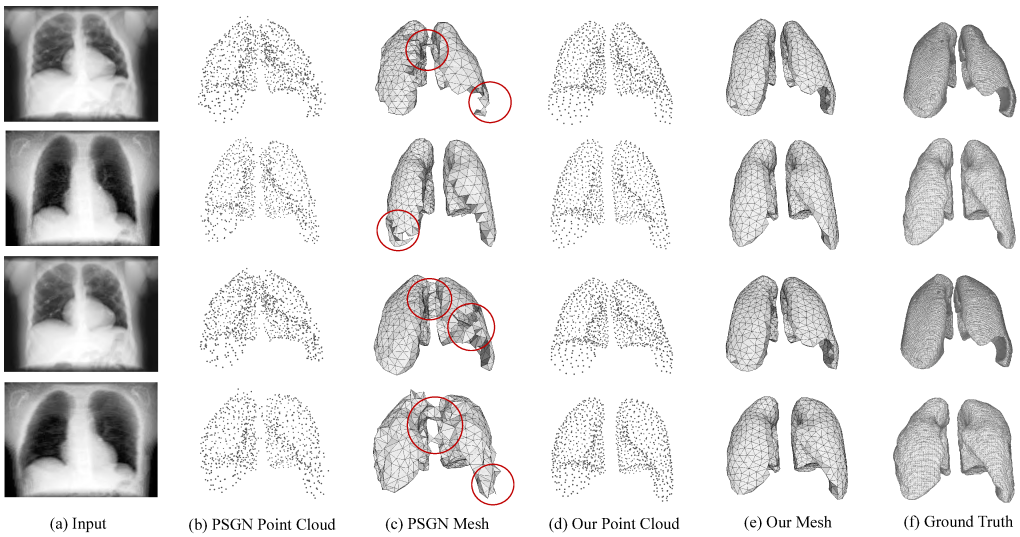

Fig. 7: Qualitative comparison between PSGN and ours. Both point clouds and solid surface meshes are given. The failure parts (e.g., self-intersecting triangles, holes, non-manifold triangles, etc.) of PSGN meshes are red-cycled.

5.3 Comparison with Traditional Reconstruction Method

Before deep learning methods are applied to 3D reconstruction area, the most common way to acquire 3D organ models from a patient is to first reconstruct 3D-CBCT volumetric image from multiple 2D projections from different views and then segment the organ models from the reconstructed volumetric image. The segmentation quality heavily depends on the number of projections. Very few views severely undermines the reconstructed 3D-CBCT image accuracy, while increasing the views impairs patient health due to a higher imaging dose as well as consumes a longer computational time. Our network offers satisfiable 3D lung shape models with only a single-view 3D-CBCT projection. Tab. 3 shows that the traditional Simultaneous Algebraic Reconstruction Techniques (SART) [1] requires at least 10-view projections to reconstruct the 3D lung mesh model and up to 50-view projections to reconstruct the 3D-CBCT volumetric image so as to segment the clean, smooth, and complete 3D lung models to reach the comparable result as ours. Fig. 8 shows the qualitative comparison between two methods. It is worth mentioning that in clinical studies (or during the therapy), it is common to use hundreds of projections to reconstruct a high-quality 3D volumetric image and then process a good-quality 3D organ model.

Table 3: Quantitative comparison between SART (with different numbers of views) and our method.

Method	CD	EMD	F-score (ϵ / 1.5ϵ)	IoU	HD (Mesh)
1-view	N/A	N/A	N/A	N/A	N/A
5-view	N/A	N/A	N/A	N/A	N/A
10-view	3.3143	92.3019	0.3999 / 0.6381	0.7277	0.1888
20-view	2.3458	66.1249	0.5107 / 0.7470	0.9099	0.1332
50-view	1.6694	43.6351	0.6753 / 0.8423	0.9433	0.0593
Ours	2.2458	48.5677	0.4931 / 0.7567	0.8880	0.0662

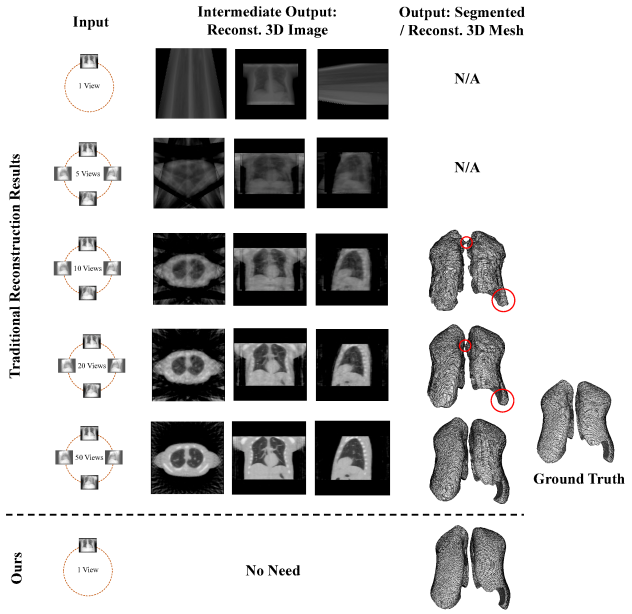

Fig. 8: Qualitative comparison between the traditional SART and our method. The reconstructed 3D-CBCT images by SART from one and five views are unable to be used to segment lungs. SART requires CBCT projections from at least 50 views to reconstruct a good-quality 3D volumetric image such that the corresponding segmented lung model is comparable to our result. Failure parts (e.g., wrong connectivities and not-good shape preserving parts) of SART-based meshes are red-cycled.

5.4 Applications and User Study on Patient Datasets

To further evaluate the accuracy and usability of the proposed method, our DeepOrganNet has been evaluated in the following studies by some domain experts, including our collaborative radiation oncologists and physicians. The efficiency and accuracy of our method demonstrate its capabilities to explicitly track, reconstruct, and visualize 3D / 4D organ shapes on the fly during the dynamic procedure and therefore it can be employed in the real-time image guided radiation therapy (IGRT).

5.4.1 3D Lung Shape Reconstruction from Patient Images

We first use ten cases of 4D-CTs from DIR-LAB datasets to evaluate the robustness of our method in real applications. For each case, we select the phase-0 of 4D-CTs to compute the front-view CBCT projection using the method in Sec. 3.1. All the generated front-view projections are histogram-equalized. We test our network directly on images of all the cases without any fine-tuning. We also further test our model on some single front-view X-ray images from JSRT database [42] and some ill-positioned single-view CBCT projections from real lung cancer patient datasets. We can see that our network is capable of describing the shape geometric property and providing a reasonable spatial arrangement in real case even though the images appear to be different from synthetic inputs. Fig. 9 shows qualitative visualization results of the above datasets, which are examined by domain experts.

5.4.2 4D Lung Shape Reconstruction

Instead of inferring the different 3D lung shapes, our network shows potential capability to track and visualize lung shapes along with the

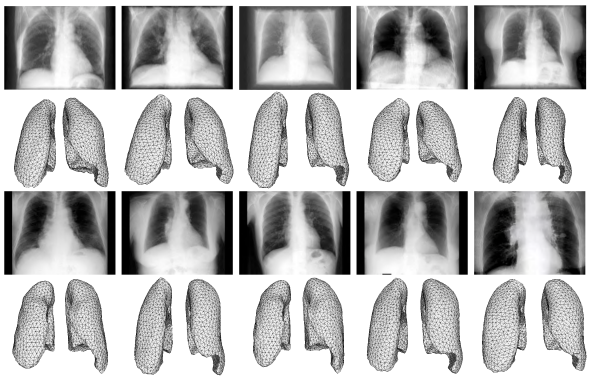

Fig. 9: Top: qualitative visualization results of 3D lung shape reconstruction from single-view phase-0 projections of five cases in DIR-LAB dataset. Bottom: qualitative visualization results of 3D lung shape reconstruction from single-view real X-ray images in JSRT database and real 3D-CBCT patient datasets. These sample results are picked from the challenging cases with large variations of the lung shapes.

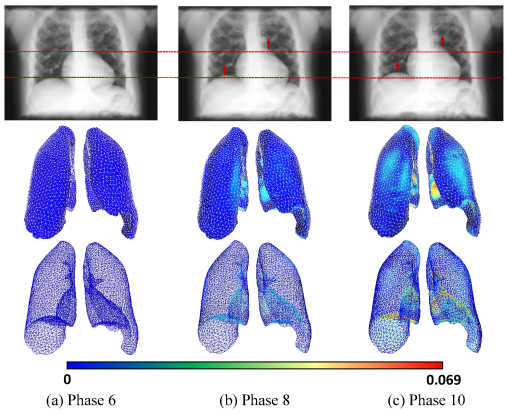

Fig. 10: Three expiration phases of 4D NCAT phantom model. Maximal deformation can be traced according to the red dashed lines across the input 2D images, and the corresponding deformations on the reconstructed 3D mesh models are mapped into a unified colormap range (hotter colors indicate larger deformations). The solid surface and wireframe meshes show the front-view and occluded (diaphragm) deformations, respectively. The deformations of Phases 8 and 10 are computed based on Phase 6 as the reference.

dynamic process of breathing. It is extremely important for IGRT procedure to understand the anatomical changes and pinpoint the location of the diseased regions on the fly. By sending a series of 2D front-view 4D-CBCT projections with different phases, our network is capable of capturing the minor changes between phases to describe the breathing tendency and maintaining the shape consistency simultaneously. It is interesting to discover that even some occluded deformations (in the natural images) in the diaphragm areas (bottom part of the lungs) can be extracted and reconstructed from the input single-view X-ray or 4D-CBCT projections. To our knowledge, this is the first time that a single-view reconstruction method can capture that. Fig. 10 and Fig. 11 show three expiration phases of a phantom case and a real case in DIR-LAB dataset. The colormaps represent the deformation magnitudes during the breathing. The solid surface meshes and wireframe meshes are used to visualize the front-view and occluded (diaphragm) deformations, correspondingly. Furthermore, the proposed method only takes about 22 milliseconds to generate 4D lung meshes with 5K vertices at each phase, which has great potential to be used in an on-the-fly targeting system on dynamic scenes in IGRT; however, there is no current method, which can make it on-the-fly.

Since our proposed method outperforms the current methods, an official clinical trial is under arrangement with our collaborative hospital.

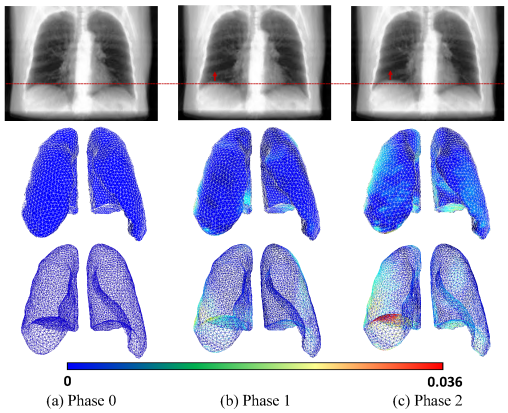

Fig. 11: Three expiration phases of Case 8 in 4D-CT DIR-LAB dataset. Maximal deformation can be traced according to the red dashed line across the input 2D images, and the corresponding deformations on the reconstructed 3D mesh models are mapped into a unified colormap range (hotter colors indicate larger deformations). The solid surface and wireframe meshes show the front-view and occluded (diaphragm) deformations, respectively. The deformations of Phases 1 and 2 are computed based on Phase 0 as the reference.

6 Conclusion

In this work, we have proposed the DeepOrganNet, a deep neural network, to generate and visualize high-fidelity 3D / 4D organ shape geometry from single-view medical images in real time. DeepOrganNet has three major components, i.e., feature encoder block, independent deformation block, and spatial arrangement (translation) block. By using the multi-organ template selection and the smooth FFD strategies in the proposed framework, our method can generate high-quality manifold meshing models, which outperforms the previous deep learning methods as well as the traditional method from the single-view image reconstruction. In medical practice, this work can be used as the key functions for real-time IGRT in order to accurately visualize the patients' organ shapes on the fly, significantly improve the procedure time for patients and doctors, and dramatically reduce the imaging dose during the treatment. Some further interactive techniques based on DeepOrganNet will be developed in collaboration with domain experts.

Discussion and Future Work: In the current framework, the lightweight MobileNets are computational efficient but limit the power of feature extraction in the encoder block. In the future, we will explore some more powerful deep neural networks for the encoder part and collect more 4D lung cancer patient datasets to improve the diversity and scalability of the training and testing for our DeepOrganNet. Although the proposed DeepOrganNet aims to reconstruct multiple 3D organs simultaneously, our current work does only implement for left and right lung organs as an example for justifying the feasibility and extendability of the proposed method. We will extend the framework into more organ reconstructions, such as heart, liver, pancreas, etc., in order to build a real fully DeepOrganNet system at a complicated 3D / 4D scene-level reconstruction. As for 4D scenarios, we have reconstructed each phase independently in the current system, and we will consider to use recurrent neural network and attention-based models to construct a 4D dynamic organ shape reconstruction deep neural network. It is worth mentioning that the quality of the reconstructed shapes can be further improved by including 2D-view projections from more viewpoints as the input to alleviate shape over- / under-estimation; we will accordingly explore how to balance the computational time (imaging dose) and reconstructed accuracy in the clinical study.

Acknowledgments

We would like to thank the reviewers for their valuable comments. This work was partially supported by the National Science Foundation under Grant Numbers IIS-1816511, CNS-1647200, OAC-1657364, OAC-1845962, OAC-1910469, the Wayne State University Subaward 4207299A of CNS-1821962, NIH 1R56AG060822-01A1 and ZJNSF LZ16F020002.

REFERENCES

[1] A. Andersen and A. Kak. Simultaneous algebraic reconstruction technique (SART): a superior implementation of the ART algorithm. *Ultrasonic Imaging*, 6(1):81–94, 1984.

[2] F. Bernardini, J. Mittleman, H. Rushmeier, C. Silva, and G. Taubin. The ball-pivoting algorithm for surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 5(4):349–359, 1999.

[3] M. Botsch, L. Kobbelt, M. Pauly, P. Alliez, and B. Lévy. *Polygon mesh processing*. AK Peters/CRC Press, 2010.

[4] R. Brock, A. Docef, and M. Murphy. Reconstruction of a cone-beam CT image via forward iterative projection matching. *Medical Physics*, 37(12):6212–6220, 2010.

[5] J. Carreira, S. Vicente, L. Agapito, and J. Batista. Lifting object detection datasets into 3D. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(7):1342–1355, 2016.

[6] R. Castillo, E. Castillo, R. Guerra, V. Johnson, T. McPhail, A. Garg, and T. Guerrero. A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. *Physics in Medicine & Biology*, 54:1849–1870, 2009.

[7] A. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. ShapeNet: an information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[8] G.-H. Chen, J. Tang, and S. Leng. Prior image constrained compressed sensing (PICCS): a method to accurately reconstruct dynamic CT images from highly undersampled projection data sets. *Medical Physics*, 35(2):660–663, 2008.

[9] C. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In *Proceedings of the European Conference on Computer Vision*, pp. 628–644, 2016.

[10] P. Cignoni, C. Rocchini, and R. Scopigno. Metro: measuring error on simplified surfaces. In *Computer Graphics Forum*, vol. 17, pp. 167–174, 1998.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[12] M. Ehlke, H. Ramm, H. Lamecker, H.-C. Hege, and S. Zachow. Fast generation of virtual X-ray images for reconstruction of 3D anatomy. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2673–2682, 2013.

[13] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, pp. 2366–2374, 2014.

[14] H. Fan, H. Su, and L. Guibas. A point set generation network for 3D object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 605–613, 2017.

[15] Q. Fang and D. Boas. Tetrahedral mesh generation from volumetric binary and grayscale images. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1142–1145, 2009.

[16] L. Feldkamp, L. Davis, and J. Kress. Practical cone-beam algorithm. *Journal of the Optical Society of America A-Optics Image Science and Vision*, 1(6):612–619, 1984.

[17] M. Fleute and S. Lavallée. Nonrigid 3-D / 2-D registration of images using statistical models. In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 138–147, 1999.

[18] D. Fouhey, A. Gupta, and M. Hebert. Data-driven 3D primitives for single image understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3392–3399, 2013.

[19] 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, pp. 377–388, 2018.

[20] D. Hoiem, A. Efros, and M. Hebert. Automatic photo pop-up. *ACM Transactions on Graphics*, 24(3):577–584, 2005.

[21] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[22] Q. Huang, H. Wang, and V. Koltun. Single-view reconstruction via joint analysis of image and shape collections. *ACM Transactions on Graphics*, 34(4):87, 2015.

[23] M. Islam, T. Purdie, B. Norrlinger, H. Alasti, D. Moseley, M. Sharpe, J. Siewerdsen, and D. Jaffray. Patient dose from kilovoltage cone beam computed tomography imaging in radiation therapy. *Medical Physics*, 33(6 Part 1):1573–1582, 2006.

[24] D. Jack, J. K. Pontes, S. Sridharan, C. Fookes, S. Shirazi, F. Maire, and A. Eriksson. Learning free-form deformations for 3D object reconstruction. In *Proceedings of the Asian Conference on Computer Vision*, 2018.

[25] M. Kan, L. Leung, W. Wong, and N. Lam. Radiation dose from cone beam computed tomography for image-guided radiation therapy. *International Journal of Radiation Oncology* Biology* Physics*, 70(1):272–279, 2008.

[26] A. Kar, S. Tulsiani, J. Carreira, and J. Malik. Category-specific object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1966–1974, 2015.

[27] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[28] A. Kurenkov, J. Ji, A. Garg, V. Mehta, J. Gwak, C. Choy, and S. Savarese. DeformNet: free-form deformation network for 3D shape reconstruction from a single image. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 858–866, 2018.

[29] P. La Riviere and D. Billmire. Reduction of noise-induced streak artifacts in X-ray computed tomography through spline-based penalized-likelihood sinogram smoothing. *IEEE Transactions on Medical Imaging*, 24(1):105–111, 2005.

[30] H. Lamecker, T. Wenckebach, and H.-C. Hege. Atlas-based 3D-shape reconstruction from X-ray images. In *Proceedings of IEEE International Conference on Pattern Recognition*, vol. 1, pp. 371–374, 2006.

[31] R. Li, X. Jia, J. Lewis, X. Gu, M. Folkerts, C. Men, and S. Jiang. Real-time volumetric image reconstruction and 3D tumor localization based on a single x-ray projection image for lung cancer radiotherapy. *Medical Physics*, 37(6 Part 1):2822–2826, 2010.

[32] R. Li, X. Jia, J. Lewis, X. Gu, M. Folkerts, C. Men, and S. Jiang. Single-projection based volumetric image reconstruction and 3D tumor localization in real time for lung cancer radiotherapy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 449–456, 2010.

[33] X. Liu, H. Wang, M. Xu, S. Nie, and H. Lu. A wavelet-based single-view reconstruction approach for cone beam x-ray luminescence tomography imaging. *Biomedical Optics Express*, 5(11):3848–3858, 2014.

[34] W. Lorensen and H. Cline. Marching cubes: A high resolution 3D surface construction algorithm. In *ACM SIGGRAPH Computer Graphics*, vol. 21, pp. 163–169, 1987.

[35] J. Pontes, C. Kong, S. Sridharan, S. Lucey, A. Eriksson, and C. Fookes. Image2Mesh: A learning framework for single image 3D reconstruction. In *Proceedings of the Asian Conference on Computer Vision*, 2018.

[36] L. Ren, J. Zhang, D. Thongphiew, D. Godfrey, Q. Wu, S.-M. Zhou, and F.-F. Yin. A novel digital tomosynthesis (DTS) reconstruction method using a deformation field map. *Medical Physics*, 35(7Part1):3110–3115, 2008.

[37] Y. Rubner, C. Tomasi, and L. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[38] O. Sadowsky, J. Cohen, and R. Taylor. Projected tetrahedra revisited: A barycentric formulation applied to digital radiograph reconstruction using higher-order attenuation functions. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):461–473, 2006.

[39] A. Saxena, M. Sun, and A. Ng. Make3D: Learning 3D scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):824–840, 2009.

[40] T. Sederberg and S. Parry. Free-form deformation of solid geometric models. *ACM SIGGRAPH Computer Graphics*, 20(4):151–160, 1986.

[41] W. Segars. *Development and application of the new dynamic NURBS-based Cardiac-Torso (NCAT) phantom*. Ph.D. dissertation, University of North Carolina, 2001.

[42] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, 174(1):71–74, 2000.

[43] R. Siddon. Fast calculation of the exact radiological path for a three-dimensional CT array. *Medical Physics*, 12(2):252–255, 1985.

[44] E. Smith, S. Fujimoto, A. Romero, and D. Meger. GEOMetrics: Ex-

ploiting geometric structure for graph-encoded objects. *arXiv preprint arXiv:1901.11461*, 2019.

[45] J. Song, Q. Liu, G. Johnson, and C. Badea. Sparseness prior based iterative image reconstruction for retrospectively gated cardiac micro-CT. *Medical Physics*, 34(11):4476–4483, 2007.

[46] W. Song, S. Kamath, S. Ozawa, S. Alani, A. Chvetsov, N. Bhandare, J. Palta, C. Liu, and J. Li. A dose comparison study between XVI and OBI CBCT systems. *Medical Physics*, 35(2):480–486, 2008.

[47] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

[48] T. Tang and R. Ellis. 2D/3D deformable registration using a hybrid atlas. In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 223–230, 2005.

[49] J. Wang, T. Li, H. Lu, and Z. Liang. Penalized weighted least-squares approach to sinogram noise reduction and image reconstruction for low-dose X-ray computed tomography. *IEEE Transactions on Medical Imaging*, 25(10):1272–1283, 2006.

[50] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang. Pixel2Mesh: Generating 3D mesh models from single RGB images. In *Proceedings of the European Conference on Computer Vision*, pp. 52–67, 2018.

[51] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1912–1920, 2015.

[52] Z. Zhong, X. Guo, Y. Cai, Y. Yang, J. Wang, X. Jia, and W. Mao. 3D-2D deformable image registration using feature-based nonuniform meshes. *BioMed Research International*, 2016.