IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 25 February 2024, accepted 24 March 2024, date of publication 5 April 2024, date of current version 15 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3385668

RESEARCH ARTICLE

X2V: 3D Organ Volume Reconstruction From a Planar X-Ray Image With Neural Implicit Methods

GOKCE GUVEN[1], HASAN F. ATES[1], (Senior Member, IEEE), AND H. FATIH UGURDAG[2], (Senior Member, IEEE)
[1]Department of Computer Science, Özyeğin Üniversitesi, 34794 Istanbul, Turkey
[2]Department of Electrical and Electronics Engineering, Özyeğin Üniversitesi, 34794 Istanbul, Turkey

Corresponding author: Hasan F. Ates (hasan.ates@ozyegin.edu.tr)

ABSTRACT In this work, an innovative approach is proposed for three-dimensional (3D) organ volume reconstruction from a single planar X-ray, namely X2V network. Such capability holds pivotal clinical potential, especially in real-time image-guided radiotherapy, computer-aided surgery, and patient follow-up sessions. Traditional methods for 3D volume reconstruction from X-rays often require the utilization of statistical 3D organ templates, which are employed in 2D/3D registration. However, these methods may not accurately account for the variation in organ shapes across different subjects. Our X2V model overcomes this problem by leveraging neural implicit representation. A vision transformer model is integrated as an encoder network, specifically designed to direct and enhance attention to particular regions within the X-ray image. The reconstructed meshes exhibit a similar topology to the ground truth organ volume, demonstrating the ability of X2V in accurately capturing the 3D structure from a 2D image. The effectiveness of X2V is evaluated on lung X-rays using several metrics, including volumetric Intersection over Union (IoU). X2V outperforms the state-of-the-art method in the literature for lungs (DeepOrganNet) by about 7-9% achieving IoU's between 0.892-0.942 versus DeepOrganNet's IoU of 0.815-0.888.

INDEX TERMS 3D reconstruction, X-ray, 3D organ topology, neural implicit methods, vision transformers.

I. INTRODUCTION

Medical imaging, particularly computed tomography (CT), plays a vital role in providing three-dimensional (3D) views of internal organ positions and shapes. This patient-specific topological information is crucial for accurate diagnosis, pre-treatment planning and interventional surgical operations. The higher radiation dose to healthy organs in CT scans [1], [2] is a clinical concern, and the cost associated with CT scans presents challenges in resource allocation. The imaging dose in CT can be reduced by reducing the number of X-ray projections. Due to the limited number of projections, this approach significantly compromises the image quality in 3D-CT reconstructions obtained through conventional methods.

The associate editor coordinating the review of this manuscript and approving it for publication was Yi Zhang.

Image-based navigation, poised as the next evolution in minimally invasive surgery, aims to enhance access to safe, precise, and reproducible surgery by integrating with modern workflows and enabling mixed reality and autonomous, robot-assisted operations. Scientific efforts have been directed towards refining solutions to derive real-time 3D organ shapes from X-rays during Image-Guided Radiation Therapy (IGRT). The primary focus has been on the 2D/3D registration of organ templates and learning deformations from average organ shapes [3].

2D/3D image registration serves as a crucial role in various image-guided interventions, including radiation therapy, radiosurgery, and minimally invasive therapy to estimate the spatial relationships between different dimensionalities of the information present in the body [4]. This process involves aligning 3D images as pre-interventional data with 2D images as intra-interventional data, creating a

fused representation that enhances intervention precision while minimizing invasiveness. Soft tissue deformations and respiratory motions during thoracic and abdominal surgeries are common problems that occur during IGRT procedures. Different positioning of the patient during the preoperative CT scan and the continuous movement of the organs due to respiration can lead to inaccuracies in the 2D/3D registration process. Another challenge is the computational complexity of the 2D/3D image registration algorithms [3]. Depending on the registration method used, the analytical solutions can be computationally intensive and time-consuming, especially for real-time applications.

FIGURE 1. a) CT; b) DRR; c) triangular mesh representations of a patient's thorax region.

The emergence of machine learning-oriented strategies for imaging issues, which utilize highly adaptive parametric models to approximate the desired functional mapping, presents a promising solution to some of the well-known difficulties in 2D/3D registration [3]. Wu et al. [5] and Wang et al. [6] introduced methods for 3D reconstruction of lung models using convolutional neural networks (CNNs) and data augmentation, with the latter focusing on learning space deformation fields from 2D images. Tong et al. [7] employed real CT-DRR (Digital Reconstructed Radiograph) pairs and 2D CNNs to reconstruct liver shapes. Nakao et al. later enhanced this approach with 2D deformation mapping and introduced the Image-to-Graph Convolutional Network (IGCN) [8], which was further developed into the IGCN+ network [9], adeptly predicting 3D organ mesh deformations and demonstrating clinical accuracy in anticipating organ motion and deformation in radiation therapy.

In medical imaging, deep learning models that rely on deformation-based learning are constrained when it comes to predicting the morphology of organs with disrupted topology. This limitation stems from their dependence on the mean shape derived from the dataset, which does not account for significant topological variations often found in pathological cases. This paper introduces the X2V network, which does not learn from the initial template shape of the organ. X2V network utilizes the occupancy probability to learn whether a point lies inside the surface of the organ or not. A total of 6,392 CT cases were curated from the National Lung Screening Trial (NLST) dataset. For each case, DRRs were generated using a ray-tracing algorithm tailored for this purpose. Corresponding lung meshes were extracted via a U-Net-based architecture (see Figure 1). The resulting pairs of DRRs and lung meshes were then utilized to train the neural network. The X2V network can predict any topological variations at infinite resolution.

This research leverages the occupancy network framework [10], enhancing the image encoder's ability to efficiently extract both global and local shape deviation for accurate computation of the actual volume of the examined 3D organ shape. Our approach can generate multiple organs from a planar x-ray by encoding topological features and implicitly representing the 3D organ volume surface as a continuous decision boundary through a deep neural network classifier with an en-to-end framework. This allows for the encoding of 3D structures at an infinite resolution without a significant memory impact. The aim of our study is to reconstruct a 3D organ shape from a patient's 2D planar X-ray image, eliminating the need for the relevant organ's 3D mean template shapes. Our method employs neural implicit representations, learning the organ's 3D geometry and computing the continuous 3D mapping.

In summary, the contributions of this study are as follows:

- To our knowledge, this is the first study to reconstruct 3D organ volume from a planar X-ray by utilizing neural implicit representations, using a lightweight and effective architecture and without a 3D template organ volume reconstruction.
- For this study, we created the largest available real patient dataset of 3D mesh organ volumes with their corresponding DRRs.
- Since it is not a template based model, the proposed network is capable of extracting the targeted volume when it is adapted to other organs, even when the default body has lost its topological properties such as bone fraction and tumor volume extraction.
- This paper is the first to use occupancy functions in deep learning-based 2D-3D medical reconstruction.
- The proposed model achieves state-of-the-art results in 3D reconstruction accuracy, using metrics such as Chamfer-L1 distance, IoU, and F-score, when compared to recent existing approaches.

II. RELATED WORK

In this study, the literature pertaining to three pivotal domains within the realm of image analysis will be explored: single image reconstruction networks, 2D-3D registration methods, and deformation learning. Each of these thematic areas presents unique methodologies and challenges in the context of 3D shape reconstruction from X-ray images, thereby providing a comprehensive framework to understand and innovate upon existing paradigms in the field.

Addressing the notable challenges of substantial radiation exposure and computational complexity in extracting 3D topology and volumetric data of human anatomy from Computed Tomography (CT), a wealth of research has been conducted in recent years. Investigations have spanned various anatomical structures, including skeletal elements (such as the knee, femur, spine, tibia, fibula, pelvis, lower extremity, and knee joint) [11], [12] and soft organs (like

the lung, liver, stomach, duodenum, kidneys, and pancreas) [5], [6], [7], [8], [9], each with distinct X-ray transmittance or Hounsfield Unit (HU) values and subject to unique topological deformations in various diseases and temporal phases, thereby necessitating the exploration of diverse solution architectures to accommodate phenomena like soft organ deformations and bone fractures. The single image reconstruction methods are more favored to study since it is clinically almost impossible to take 2 or more correspondent x-ray imagery to reconstruct the 3D shape.

A. SINGLE IMAGE 2D-3D RECONSTRUCTION METHODS

In recent times, the field of 3D reconstruction has witnessed a substantial influx of research efforts, culminating in the emergence of single-image reconstruction as a pivotal and popular domain within image processing.

Various methods exist for 3D shape representation in a scene, including depth and normal maps for partial geometry and orientation, and point clouds, meshes, and voxels for more comprehensive 3D depictions, each with their own advantages and limitations in portraying surface and solid models. While depth and normal maps offer a ''2.5D'' perspective due to their single viewpoint, point clouds, meshes, and voxels provide a fuller 3D representation, albeit with challenges related to computational memory and data downsampling [13].

Bednarik et al. [14] and Patch-Net [15] employ similar architectures to reconstruct 2.5D shapes from textureless surfaces using normal and depth maps, with the latter implementing a patch-based strategy to notably enhance depth and angular accuracy. Shimada et al. introduced HDM-Net [16] and IsMo-GAN [17], which utilize various network improvements and adversarial training, respectively, to enhance 3D mesh reconstruction, with IsMo-GAN also demonstrating proficiency in 3D point cloud reconstruction through additional network modifications. Pixel2Mesh (P2M) [18] and VANet [19] utilize dual-lane architectures for 3D mesh reconstruction, with VANet achieving a substantial reduction in Chamfer distance error by enhancing its feature extraction capabilities. Lastly, Salvi et al. [20] introduced an attentioned occupancy network that employs self-attention blocks within its encoder to reconstruct 3D shapes as continuous functions, achieving significantly improved Chamfer-L1 distances compared to P2M.

B. 2D/3D REGISTRATION-BASED APPROACHES

Initial studies in the domain of 2D-3D registration predominantly focused on deriving 3D volumes of skeletal system elements, potentially due to bones having higher HU values compared to other tissues and their boundaries not being subject to immediate, continuous deformations. The pivotal role of image registration in numerous medical image analysis applications has been notably enhanced by the advent of deep learning [21], leading to substantial improvements in algorithmic performance across various computer vision tasks. 2D/3D registration in deformable contexts grapples with the complexity and ambiguity arising from high-dimensional parameter spaces and challenging parameter modifications due to scarce information [21].

Humbert et al. [22] and Chaibi et al. [23] presented parametric models to represent spine and femur bones from biplanar X-rays. Cresson et al. [24] devised an algorithm to infer obscured spinal regions using a two-step process involving a 2D/3D registration procedure and refining reconstructions with a Statistical Shape Model (SSM) with the 2D bone contours extracted from the patients planar radiograph. Zhang et al. [25] developed a technique that automatically identifies vertebral orientations and positions by aligning vertebral contours, employing an extended generalized Hough transform method and a strategy that tolerates deformation. This method seeks optimal alignment of 3D primitive projections with biplanar radiographs.

2D/3D Registration methods align 2D images with 3D models, which might not always result in high-resolution models, especially when the available 2D images are of low quality or when they do not cover all the aspects of the 3D shape.

C. DEFORMATION LEARNING BASED APPROACHES

As foundational studies in 3D reconstruction, P2M and PointNet have paved the way for generating 3D structures from 2D images in computer vision, each introducing unique methodologies and subsequent challenges in the field, particularly in the context of deformation learning and medical applications.

Recent advancements in deep learning have enhanced 2D/3D deformable registration, with Pointnet [26] creating 3D point clouds from single-viewpoint images, albeit losing some topological information. Conversely, Wang et al.'s P2M [18] generates a 3D mesh from a 2D image by deforming an ellipsoid template using latent image features.

Wu et al. [5] introduce a new 3D shape reconstruction method tailored for lung surgeries, addressing the challenge of data scarcity in the medical field by utilizing a deformable statistical model of the 3D lungs and employing data augmentation techniques alongside CNNs. Despite its innovations, the reliance on point clouds leads to significant frailties, notably the neglect of essential surface details and the loss of topological data concerning vertex relationships. Wang et al. [6] utilized the P2M network and developed a CNN method to reconstruct 3D/4D lung models from single 2D images, such as 3D/4D-CT projections or X-ray images, by learning space deformation fields from synthetically crafted multiple initial templates.

Tong et al. [7] have utilized the real CT-DRR pairs for the first time to reconstruct the liver shape of the actual patients using 2D CNN and a graph convolutional surface deformation network. Since low contrast or invisible contours can generate problems for 2D/3D registration, in their next paper [8], Nakao et al. have added 2D deformation mapping

and stated a mean distance error of 3.6mm for reconstruction of liver shapes from a planar DRR and called the network image-to-graph convolutional network (IGCN). The authors have also introduced the IGCN+ network [9] that adeptly predicts multi-3D organ mesh deformations from 2D images of abdominal organs, demonstrating clinically acceptable accuracy in anticipating respiratory motion and deformation across multiple abdominal organs and pancreatic cancer in radiation therapy contexts.

Template deformation methods deform a template mesh to fit the target shape, which might introduce artifacts or lose details if the template is not sufficiently detailed or if it does not closely resemble the target shape. Depending on the complexity of the target shape and the template, template deformation might result in self-intersections, especially in regions with high curvature or complex geometry.

In contrast to the 2D/3D registration and template deformation learning methods, this paper shows that utilizing the occupancy function leads to high resolution closed surfaces without self intersections and achieves highly accurate reconstructions without requiring any template meshes from the same object class.

III. PROPOSED APPROACH

This section introduces the X2V deep network model architecture for X-ray to 3D lung mesh reconstruction and explains the details of network training and inference.

A. X2V MODEL ARCHITECTURE

Neural implicit methods, leveraging neural networks, represent 3D shapes indirectly through a continuous function, typically providing a smooth and differentiable shape representation beneficial for gradient-based optimization. On the other hand, The ViT (Vision Transformer) model has revolutionized the field of computer vision by applying transformers, traditionally used in natural language processing, to image classification tasks [27]. This paper applies a novel ViT based neural implicit model, conditioned on the input X-ray image, for 3D lung mesh reconstruction.

The X2V model architecture is shown in Figure 2a. The network accepts as input a DRR image of size 224×224 and a set of 3D-coordinates of T random points in space. The network predicts the occupancy probability (i.e. the probability that a point in space is occupied) of each point at its output. To reconstruct a 3D surface from a DRR image, the occupancy function is conditioned on the antero-posterior DRR image. The occupancy function [10], central to our methodology, implicitly represents the lung surface as a decision boundary. This approach generates high-resolution, non-self-intersecting closed surfaces without relying on input template meshes.

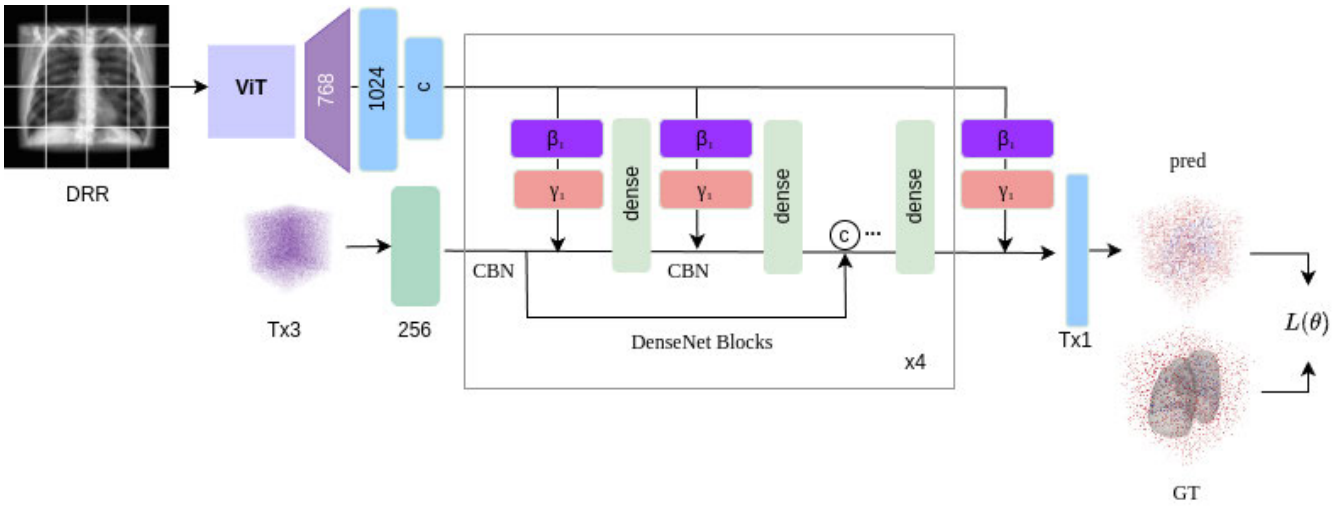
In the X2V network, a low complexity Vision Transformer (ViT) model is employed as a feature extractor to condition the occupancy probability of 3D points in space on the input DRR image. The transformer divides the image into 16×16 pixel patches; these patches are linearly embedded and augmented with absolute positional encodings to retain their spatial information. The core of the model consists of a transformer encoder that employs multi-head self-attention and feed-forward layers, leveraging the transformer's capacity to capture long-range dependencies within the image. A classification head, typically connected to the output associated with a special "class token," generates the final prediction. This model strikes a balance between computational efficiency and performance, making it suitable for a broad range of image-related tasks.

In our approach, the ViT model extracts rich, high-dimensional features from the input DRR image, hence transforms the input image into a latent feature representation, which effectively summarizes the visual information necessary for 3D reconstruction. This integration of ViT network with neural implicit models represents a novel approach to harnessing the power of transformers in image reconstruction tasks, aiming to improve the quality and fidelity of the generated 3D models.

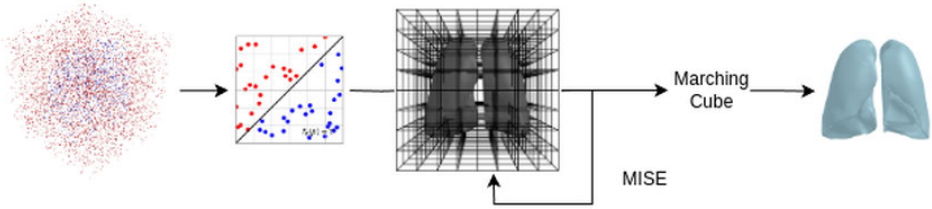
The Vision Transformer (specifically, ViT-B/16) model [28], [29] for feature extraction is altered by removing its original classification head. This modification shifts the model's focus from classifying images to providing a dense feature representation. Following this, a projection layer linearly transforms the 768-dimensional feature vector into a higher-dimensional space of 1024 features. The purpose of this transformation is to expand the feature space, potentially capturing more complex relationships within the data and making the representation more suitable for downstream tasks that require a richer feature set.

The occupancy network model takes the output of ViT-B/16 (represented as codevector c) and a batch of T 3D points (formatted as a $T \times 3$ dimensional matrix) as input and produces the occupancy probability of each 3D point at the output. The input points are passed through a fully connected layer to produce a 256 dimensional feature vector for each point. The feature vectors are passed through 4 DenseNet [30] blocks that first apply conditional batch normalization enabling dynamic adjustment of normalization parameters based on the conditional codevector c (See Figure 2a). Conditional Batch Normalization (CBN) [31] is an advanced normalization technique that modifies the traditional batch normalization process to be dependent on additional external information, often in the form of a conditional vector c. Linear layers ($\gamma(c)$ and $\beta(c)$) generate scale and shift factors from c, modulating the normalized output to adapt to different conditions dynamically. This approach allows the normalization process to be finely tuned based on contextual information, making it highly versatile for applications requiring responsive adjustments to input variations, thereby enhancing the model's ability to process conditionally varied data efficiently.

Initially, the conditioning vector c that represents external information such as a 2D image, derived from the encoder network, passes through two separate fully-connected layers,



(a) Our model architecture employs a neural network to approximate the occupancy function for 3D objects, interpreting each point in space as an occupancy probability. The X2V framework enhances this process with ViT encoder, densely connected layers and conditional batch normalization, improving the network's ability to reconstruct complex 3D structures. Training involves processing DRR-CT pairs and evaluating loss at sampled points within the object's bounding volume, using a structure that integrates feature vectors through multiple processing stages for deep feature representation. This approach ensures efficient 3D reconstruction from diverse inputs.



(b) For 3D lung mesh inference with *X2V* model, a modified Multiresolution IsoSurface Extraction (MISE) algorithm for high-resolution mesh extraction is integrated, starting with a base resolution and evaluating against the occupancy network. The occupancy threshold is set at $\tau = 0.2$ for balance in accuracy and completeness. The process involves subdividing voxels until the desired resolution is reached, using Marching Cubes for mesh generation, and refining the mesh with Fast - Quadric - Mesh - Simplification and gradient optimization. Our method achieves efficient and accurate mesh inference, optimized for an initial resolution of 32^3, and is capable of extracting mesh normals effectively.

FIGURE 2. X2V network architecture training and inference phases.

resulting in 256-dimensional vectors $\beta(c)$ and $\gamma(c)$. Subsequently, the 256-dimensional input feature vector f_{in} of each CBN layer is normalized using its mean and variance, followed by a scaling operation with $\gamma(c)$ and an addition of the bias term $\beta(c)$. CBN layer applies the following normalization to the input feature vectors f_{in}:

$$\text{CBN}_i(c, f_{\text{in}}) = \gamma_i(c)\left(\frac{f_{\text{in}} - \mu}{\sqrt{\sigma^2 + \epsilon}}\right) + \beta_i(c) \qquad (1)$$

The densenet blocks integrate conditional batch normalization and convolutional layers, including a ReLU activation function, to process and transform the input tensor. The blocks facilitate a two-step transformation—initial normalization and transformation of input, followed by concatenation with the original input, another transformation and projection to the desired output dimension.

As seen in Figure 2a, each DenseNet block employs skip connections in order to transfer shallow layer features to deeper layers and to facilitate better gradient flow during back-propagation. The dense connectivity in the architecture ensures comprehensive feature integration, as each layer is fed inputs from all previous layers, thus enhancing the depth and subtlety of feature fusion. The inclusion of CBN allows for dynamic adjustment of normalization parameters, tailored to the specific attributes of the input DRR image. Overall, the joint use of ViT-based encoding, CBN layers and DenseNet architecture make substantial contributions to the 2D-to-3D reconstruction performance of the proposed X2V model.

B. TRAINING

Ideally, our aim is to discern the occupancy not merely at fixed, discrete 3D locations but at every conceivable 3D point $p \in \mathbb{R}^3$. The function

$$o : \mathbb{R}^3 \rightarrow \{0, 1\}$$

is termed the occupancy function of the 3D object. A key insight is that this 3D function can be approximated by a neural network, assigning each location p an occupancy probability between 0 and 1, effectively utilizing the decision boundary of a binary classification network to implicitly represent the object's surface. The network, represented as $f : \mathbb{R}^3 \times X \rightarrow \mathbb{R}$, conditions the task of 3D reconstruction on an observation $x \in X$ by mapping the pair (p, x) to a real number that indicates the probability of occupancy.

In the context of a binary-classification network, training involves evaluating the mini-batch loss $L_B(\theta)$ at randomly sampled points within the 3D bounding volume of the object. This is defined as:

$$L_B(\theta) = \frac{1}{|B|} \sum_{i=1}^{|B|} L(f_\theta(p_i, x_i), o_i) \tag{2}$$

where x_i is the i^{th} observation of batch B, $o_i \equiv o(p_i)$ denotes true occupancy at point p_i, and $L(\cdot, \cdot)$ is binary cross-entropy classification loss. In this case, the method's effectiveness is influenced by the scheme used for sampling the locations p_i employed during training. Optimal results are obtained with uniform sampling inside the object's bounding box [10].

In all conducted experiments, the Adam optimizer [32] was employed for training, configured with a learning rate (η) of 10^{-4}. Regarding the Adam optimizer's additional hyperparameters, the default settings provided by PyTorch were adopted: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$.

C. INFERENCE

Figure 2b depicts the inference of 3D lung meshes from occupancy predictions at the X2V network output. In the *X2V* implementation, a version of the Multiresolution IsoSurface Extraction (MISE) algorithm [10], originally detailed in the occupancy network inference approach, is integrated. This integration facilitates efficient high-resolution mesh extraction from our network. Our process begins with the initial discretization of the volumetric space, followed by an evaluation against the occupancy values at this base resolution. The voxelization involves setting a threshold to determine whether each voxel, or 3D pixel, is inside or outside the object. Essentially, the threshold helps in classifying voxels based on their likelihood of being part of the object, as predicted by the deep learning model. This step is crucial for transforming the model's probabilistic outputs into a clear, discrete 3D representation, enabling the accurate reconstruction of objects from single image viewpoint. A critical aspect of our approach is the identification and marking of grid points exceeding a certain occupancy threshold, which is the sole parameter determining the thickness of the extracted 3D surface. To choose the threshold parameter τ for our method, we opt for the optimum value $\tau = 0.2$, which yields a good trade-off between accuracy and completeness, as indicated in the ONet article [10].

Subsequently, active voxels are detected and subdivided, which are characterized by adjacent grid points with differing occupancy predictions. This iterative subdivision continues until the desired resolution is reached. At this final resolution, the Marching Cubes algorithm is employed to generate the initial mesh.

To refine the initial mesh, Fast-Quadric-Mesh-Simplification algorithm [33], which is based on iterative edge contraction and quadric error metrics, is employed, followed by a gradient-based optimization process. This ensures both the simplification and enhancement of the mesh quality. Our method ensures convergence to an accurate mesh, contingent on the initial resolution encapsulating all necessary components of the mesh's interior and exterior. In our application, an initial resolution of 32^3 voxels has proven to be effective.

IV. DATASET

In this paper, X2V network, strategically designed to isolate various organs from X-ray images using a singular network, is trained for lung organ extraction due to the high availability of lung data and its higher HU contrast compared to other tissues, since it is filled with air. Early experimentation with various organ segmentation models revealed the superior accuracy of U-net(R231) [34], guiding our focus towards lung segmentation. The selection of the organ and architectural approach was thus influenced by data accessibility and the accuracy of the mesh segmented from its CT. The dataset generation steps are summarized as below:

- Data collection through The Cancer Imaging Archive (TCIA) [35].
- Lung segmentation from CT images as watertight manifold meshes via U-net (R231) network [34].
- Occupancy value calculation for the points in the lung meshes.
- Collecting corresponding DRR images of the CT scans from an antero-posterior (AP) view.
- Applying CLAHE contrast enhancement to obtain enhanced DRR images.

A part of the NLST dataset [36], [37] is downloaded from TCIA [35] using NBIA Data Retriever Tool. A volumetric analysis and visualization in 3D Slicer (https://www.slicer.org/) is performed [38], [39], [40], [41] via the Lung CT Analyzer project with U-net(R231) network architecture [34], [42] on NLST dataset leveraging SimpleITK library. For the purpose of training and validating our proposed CT reconstruction approach, an ideal dataset would consist of a large collection of paired X-rays and corresponding CT reconstructions. However, such datasets are not readily available and the process of collecting them can be expensive. To overcome this challenge, the digitally reconstructed radiographs (DRR) technology [43] is utilized to synthesize the corresponding X-rays using a real CT volume, as depicted in Fig. 1. In our study, approximately 2560 subjects from the publicly available National Lung Screening Trial (NLST) dataset are selected [37] that comprises the low-dose CT scans from 26,254 of the subjects in TCIA [35].

After eliminating CT scans that do not encompass the entire lungs and dicom cases with flawed CT images, suitable rotation matrices are employed to align CT scans that are not similarly oriented, ensuring uniform direction and alignment with respect to the anterior, posterior, left, and right anatomical planes of the patients.The second elimination step involves removing low-quality segmentations following the 3D organ mesh creation phase. In the segment of the NLST dataset [37] that is compiled, for each patient, one random CT scan from various years for their examination are selected, in cases where multiple CT scans per year are available.

A. 3D ORGAN MESHES

6392 3D-CT scans were collected from 2560 selected subjects, and 3D organ meshes were segmented via the Lung CT Analyzer which has a Dice Similarity Coefficient (DSC) of 0.98 $\pm$ 0.03 on the test dataset [35]. A Python script was crafted to streamline the segmentation process, employing the Python interactor of the 3D Slicer software. This script automatically loads the DICOM file into 3D Slicer, triggering the Lung CT Analyzer module—a 3D Slicer extension—that segments the left and right lungs independently and amalgamates them according to their CT orientation, ensuring that the segmented lungs' position and volume align with the CT without necessitating further object orientation [38], [39], [40], [41]. Essentially, the code resamples the DICOM to a 1.5mm $\times$ 1.5mm $\times$ 1.5mm spacing using B-spline interpolation, segments the lung regions, exports the closed surface representation of multiple segments to files, and merges all segments into a singular mesh.

B. PLANAR DRR IMAGES

6,392 3D-CT scans are collected from 2,560 subjects and resampled to a resolution of 512 $\times$ 512 $\times$ 512 pixels with a voxel size of 1.0mm $\times$ 1.0mm $\times$ 1.0mm. 512 $\times$ 512 DRR images with 1.0 $\times$ 1.0 spacing are generated using a Siddon Ray Tracing algorithm on a GPU [43], [44]. In this study, Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied to DRR images to enhance the contrast and resolution of local details.

C. DATA PREPROCESSING

The trimesh Python library is utilized to confirm the watertightness of the acquired meshes, a necessary condition for determining whether a point resides within a mesh's interior (e.g., for calculating IoU) [10]. All meshes are centered and rescaled, ensuring the 3D bounding box of each mesh is centered at (0,0,0) and the longest edge across the dataset measures 1 so that the meshes are aligned with the voxelization forms. The maximum value among the x, y, and z dimensions of the bounding box measurements of the acquired 3D lung volumes is used to uniformly normalize all meshes. As in the context of Occupancy Networks (ONet) [10], voxel representations facilitate efficient querying of occupancy probabilities in the 3D space, enabling the generation of detailed and accurate 3D reconstructions [10].

Employing the voxelization method from Choy et al. [45], the voxelizations of 32^3 are generated. The dataset is partitioned into training, validation, and test subsets using a 7:2:1 ratio, with the training subset further divided into additional training and validation sets. All voxels intersecting the mesh surface are identified and labeled as occupied. Subsequently, an evaluation is performed by selecting an arbitrary point within each voxel to determine its position relative to the mesh, either inside or outside. If the point is situated inside, the corresponding voxel is labeled as occupied. 100k points are generated within a padded unit cube and their positions are determined relative to the watertight mesh by counting intersections with a z-axis parallel ray; an even count signifies an external point, odd indicates internal (see Figure 4). Both the 100k point coordinates and their occupancies are saved, and a 2048-point subset is randomly selected during the training phase.

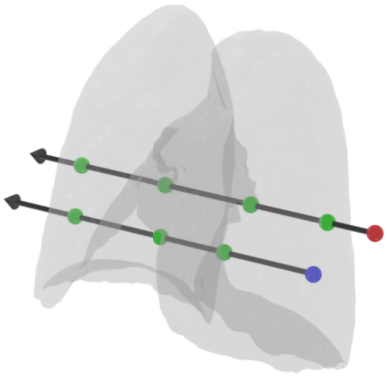


FIGURE 3. Illustration of the method for calculating occupancy values. A ray is drawn from a 3D point along the z-axis, which intersects the mesh surface. Points of intersection are marked in green. If the ray intersects the mesh an even number of times, the point is considered as external, indicated by a red dot. Conversely, an odd number of intersections make the point labelled as internal, shown by a blue dot.

The ground truth and predicted meshes and their corresponding distance heatmaps are shown in Figure 3. X2V model outputs are displayed sequentially from left to right, starting with the Digital Reconstructed Radiographs (DRRs). These are followed by the anterior views of the Ground Truth (GT) and the predicted 3D structures. Adjacent to these are the Euclidean distance maps, which quantify the spatial discrepancy between the GT and predicted models. Completing the sequence, the posterior views of both the GT and predicted structures are presented for comprehensive visual comparison.

V. EXPERIMENTS

This section covers the simulation results and comparative analysis of X2V network. Evaluation metrics are explained and the results are discussed in detail. The 3D reconstruction performance of X2V is compared against two recent 2D-3D

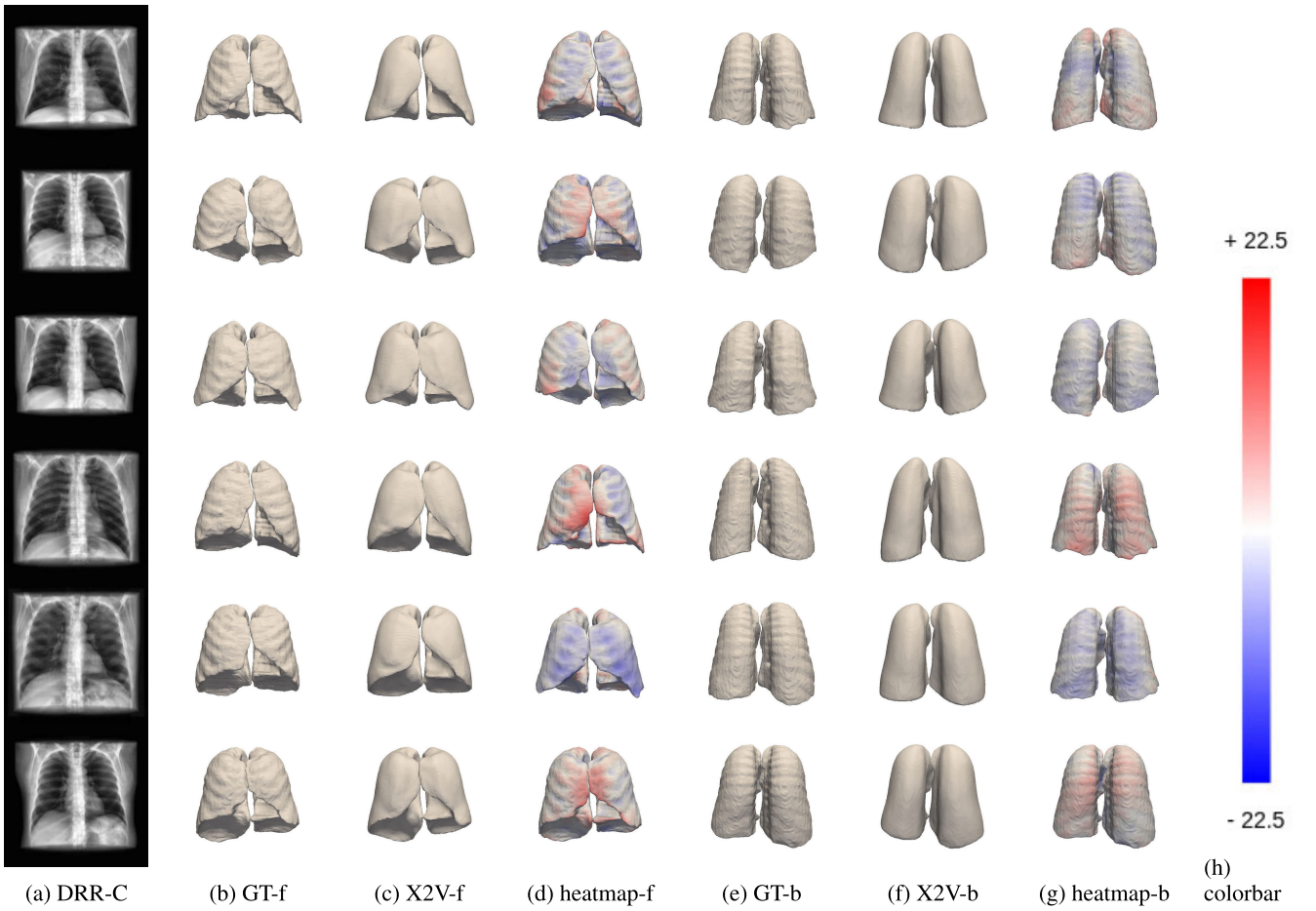


FIGURE 4. Real-size lung volume analysis of GT and X2V network output: (a) Single anteroposterior DRR with CLAHE of lungs as input to networks; (b) Ground truth lung mesh front-view(-f); (c) Front side of the lung high resolution mesh generated by the X2V network; (d) Distance heatmap between GT and predicted mesh front-view; (e) Ground truth lung mesh back-view(-b); (f) Back side of the lung mesh generated by the X2V network; (g) Distance heatmap between GT and predicted mesh from back-view; (h) color scalar bar (blue min −22.5 mm, red +22.5 mm with mean −0.328mm and std 2.825mm).

reconstruction models from the literature, namely, Point Set Generation Network (PSGN) [46] and P2M Network [18].

A. COMPARATIVE STUDIES

This section explains the details of PSGN and P2M models, which are used for performance comparison.

1) POINT SET GENERATION NETWORK

The Point Set Generation Network (PSGN) [46], a 3D reconstruction model, generates 2048 3D point coordinates from a single image input. It utilizes a ResNet-50 encoder and a decoder with four fully connected layers. Models are trained for 10 hours on an Nvidia GTX 4090 GPU, with a batch size of 64. In the PSGN [46], normalization of the 3D ground truth point clouds is achieved by moving their centroid to the origin and scaling them to fit within a standard volume, such as a unit cube, ensuring consistent scale. Uniformity in point density across these point clouds is attained by sampling a fixed number of points using the farthest point sampling method.

2) PIXEL2MESH

P2M [18] applies a mesh-focused technique to reconstruct 3D structures from individual images, iteratively adjusting an initial ellipsoid mesh to approximate the object by harnessing perceptual features from the input. The network operates with a sequence of three consecutive mesh deformation stages. Each stage enhances the mesh's resolution and refines the positions of the vertices. These updated vertex coordinates serve to gather perceptual features from the 2D image via a CNN, informing the subsequent deformation stage with Graph based ResNet (G-ResNet). In our experiments, P2M is trained using our DRR-CT lung dataset, while conforming to the slightly altered implementation from the P2M paper [18] via changing the perceptual feature pooling layer from VGG-16 to ResNet-50 network architecture. Additionally, it is important to mention that our method employs uniformly sampled points from CAD models for ground truth, in contrast to the P2M approach, which uses Poisson-disk sampling. The comparative analysis in the subsequent section demonstrates that the proposed X2V

model significantly outperforms the P2M method in terms of IoU, Chamfer-L1 distance, and F-score metrics.

A starting template mesh, typically an ellipsoid or sphere, is selected for the deformation process in P2M [18]. This mesh is then scaled to the target mesh's dimensions for consistency. To enhance the starting configuration, features from the input image are superimposed onto the mesh. Connectivity between vertices is depicted through an adjacency matrix, essential for the graph convolutional network (GCN) layers, while the graph Laplacian is calculated to embody the mesh structure, aiding deformation learning. The vertices' coordinates are incorporated into the feature vector for network input. Additionally, camera parameters are employed to transform the initial mesh to align with the input image's perspective.

B. EVALUATION METRICS

In this research, the proposed method and established baselines are quantitatively evaluated using the IoU, Chamfer-L_1 (C-L_1) distance, F-score and normal consistency score metrics. The Chamfer L_1 distance [10], denoted as C-L_1, serves as a pivotal metric for assessing the similarity between two 3D shapes. It quantifies the discrepancy by computing the average nearest-point Euclidean distance between corresponding points on the surfaces of both a predicted 3D model and a ground truth model. This metric is inherently symmetric and normalized, facilitating a comprehensive and robust comparison of 3D reconstruction accuracy. To calculate the Chamfer L_1 distance, for each vertex p in the predicted mesh (PR), $p \in \partial M_{\text{PR}}$, the nearest vertex $q \in \partial M_{\text{GT}}$ within the ground truth (GT) mesh is identified, and conversely, for each vertex in the GT mesh, the nearest vertex in the predicted mesh is determined. The sum of these minimum distances is then averaged, providing a measure that encapsulates the overall spatial deviation between the two meshes.

$$\text{C-L}_1(M_{\text{PR}}, M_{\text{GT}}) \equiv \frac{1}{2|\partial M_{\text{PR}}|}\int_{\partial M_{\text{PR}}} \min_{q\in\partial M_{\text{GT}}} \|p-q\|dp + \frac{1}{2|\partial M_{\text{GT}}|}\int_{\partial M_{\text{GT}}} \min_{p\in\partial M_{\text{PR}}} \|p-q\|dq. \tag{3}$$

IoU for 3D meshes is the ratio of their intersecting volume to the volume of their union. In evaluating 3D point cloud similarity, the voxelized IoU is calculated by first converting each point cloud into a voxel grid, marking voxels as occupied or not. The IoU is then determined by dividing the count of voxels occupied in both grids (intersection) by the count of voxels occupied in at least one grid (union), where P_{PR} and P_{GT} are the sets of occupied voxels for predicted and ground truth meshes, respectively. This metric ranges from 0 (no overlap) to 1 (perfect overlap), providing a standardized measure of spatial congruence between two 3D models.

$$\text{IoU}(P_{\text{PR}}, P_{\text{GT}}) \equiv \frac{|P_{\text{PR}} \cap P_{\text{GT}}|}{|P_{\text{PR}} \cup P_{\text{GT}}|}. \tag{4}$$

A higher IoU indicates greater overlap and similarity between the meshes, useful for evaluating accuracy in 3D reconstructions. The meshes and point clouds are voxelized with voxel size of 0.5 to calculate IoU values. Voxelized approach is adopted due to the inherent limitations of the P2M and PSGN methods used for comparison, which do not generate watertight meshes. Normally, the PSGN method produces point clouds; however, for comparison with our method, meshes are generated from point clouds using the ball pivoting algorithm [47].

Normal consistency (NC) between two meshes is assessed by calculating the absolute dot product of surface normals from one mesh with those of the nearest neighbors in the other mesh. This measure reflects the alignment accuracy of corresponding surface elements, focusing solely on geometric correspondence without being affected by the positive or negative directionality of the normals.

The F-score is a widely recognized metric for assessing the accuracy of 3D mesh reconstructions, encapsulating both precision and recall into a singular measure. The F-score between two meshes is a balanced measure combining precision and recall, calculated as the harmonic mean of the proportion of correctly identified points in the predicted mesh (i.e. precision) and the proportion of actual mesh points that are correctly identified (i.e. recall). This score effectively evaluates the accuracy of 3D mesh reconstructions, with a higher F-score indicating better accuracy and similarity. The F-score is computed as follows:

$$F\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall} + \epsilon} \tag{5}$$

Precision and recall are defined based on the Euclidean distance between corresponding points in the ground truth and predicted meshes, with respect to a predefined threshold t. Specifically,

$$\text{Precision} = \frac{\sum_{i=1}^{N_{\text{PR}}} \mathbb{1}(d_{\text{PR}}^i < t)}{N_{\text{PR}}} \tag{6}$$

$$\text{Recall} = \frac{\sum_{i=1}^{N_{\text{GT}}} \mathbb{1}(d_{\text{GT}}^i < t)}{N_{\text{GT}}} \tag{7}$$

In these equations, N_{PR} and N_{GT} denote the number of points in the predicted and ground truth meshes, respectively. The function $\mathbb{1}(\cdot)$ is an indicator function that returns 1 if the condition within is true (i.e., the Euclidean distance d between a point in one mesh and its nearest neighbor in the other mesh is less than the threshold t), and 0 otherwise. This threshold t is crucial as it defines the sensitivity of the metric to the distance between corresponding points, thereby influencing the assessment of reconstruction accuracy. In the experiments, t is set as 0.02. A small constant $\epsilon = 10^{-8}$ is added in F-score calculation to avoid division by zero.

C. SIMULATION RESULTS AND DISCUSSIONS

This section compares the performance three tested methods, X2V, PSGN and P2M, first visually and then using the objective evaluation metrics.

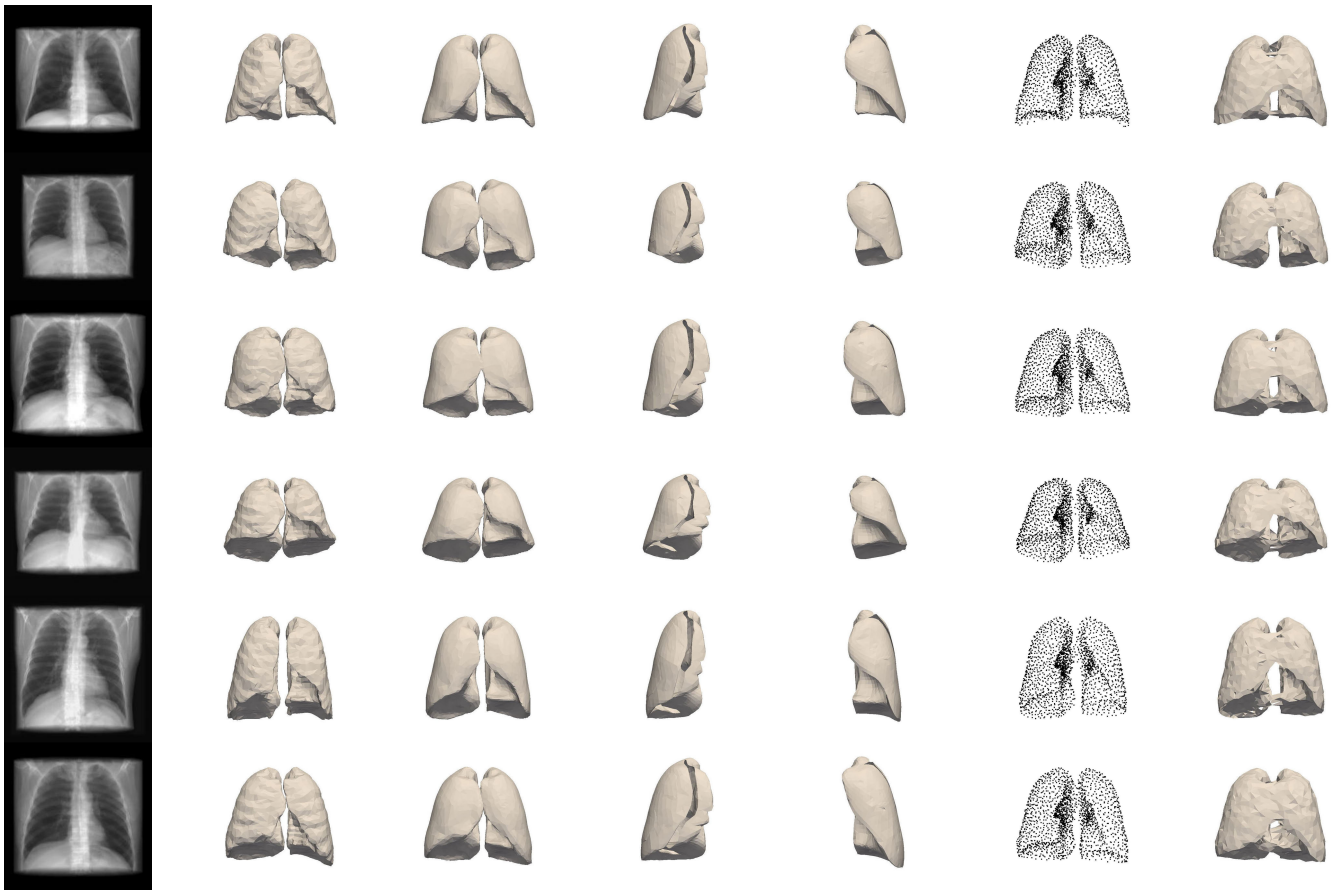


FIGURE 5. Comparison for lung volume analysis. (a) Single antero-posterior DRR of lungs as input to networks. (b) Ground truth lung mesh. (c) high-accuracy, real-size, watertight lung mesh generated by the X2V network. (d) and (e) separately processed left (e) and right (d) lung meshes, modified from initial ellipsoid shapes due to deformations, using the P2M algorithm. (f) PSGN derived point cloud lacking connectivity, subsequently converted into a mesh (g) using the Ball Pivoting Algorithm, exhibiting non-manifold characteristics.

Figure 5 shows that each tested method effectively replicates the 3D geometry from an antero-posterior DRR. The PSGN method demonstrates the ability to produce relatively high-quality outputs; however, it exhibits limitations in maintaining structural connectivity. This shortcoming necessitates the implementation of lossy post-processing techniques, such as the ball pivoting algorithm, for mesh reconstruction. Conversely, P2M shows proficiency in generating visually appealing meshes but encounters difficulties in accurately rendering openings in the presence of complex topologies. This challenge is particularly evident around the adjacent corresponding surface areas in the predicted P2M-R and P2M-L lung meshes, as illustrated in Figure 5, which displays non-manifold surface properties and holes throughout the surface of the meshes. In P2M, the inherent complexity of lung shapes poses challenges when attempting to model them with a basic deformable mesh originating from a standard lung. Especially, capturing precise details of complex or slender structures is difficult, which leads to inaccuracies such as holes or meshes that are not watertight.

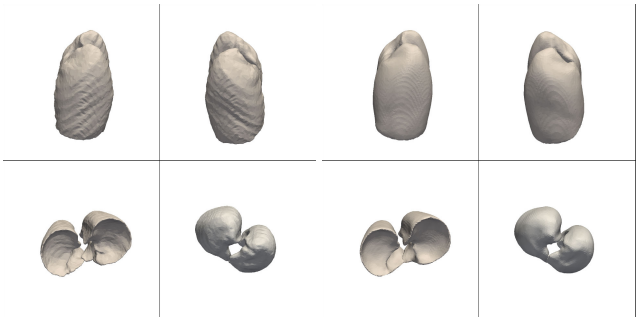
The X2V model demonstrates the capability to accurately predict the left and right lungs within their respective 3D spaces, achieving a real-size mesh prediction accuracy of 0.5 mm, as illustrated in Figure 3. Note that, the results depicted in Figure 5 are normalized within a unit cube. Close examination of X2V-derived meshes reveals that corners of the lung meshes and detailed topological features align closely with the ground truth. Furthermore, the meshes produced by the X2V model are watertight, indicating no gaps or holes, which is crucial for maintaining the integrity of the 3D representations.

In Figure 6, the ground truth mesh and the X2V predicted mesh share nearly identical global topological characteristics, although some disparities in local properties are evident when viewed from the side of the meshes. Specifically, the rib imprints on the lung meshes are not as discernible on the predicted meshes.

Tables 1 and 2 present quantitative comparison of our X2V approach against standard benchmarks in single image 3D reconstruction, utilizing our test set. The values in Tables 1 and 2 are calculated by averaging metrics over 632 test data.

IoU, Chamfer-L1 distance, F-Score and normal consistency (NC) metrics are calculated for X2V, P2M and PSGN network model outputs, with respect to the ground truth meshes.

In the P2M experimental framework, the P2M model is deployed for the prediction of a singular (left or right) lung mesh from a single-view input image. Utilizing the initial template provided by P2M, the process generates a single mesh in alignment with the methodologies established in the foundational work.



(a) GT (b) X2V

FIGURE 6. Different views for a sample lung mesh. (a) ground truth lung mesh. (b) lung mesh generated by X2V network.

The resolution of the output mesh for PSGN predictions stands at 2048 points. In the case of P2M-L and P2M-R, the resolution is 2466 for each, attributable to the default initial mesh in P2M featuring 2466 vertices. Meanwhile, the X2V model employs 2048 and 4932 vertices for comparison in Tables 2 and 1 respectively, matching the resolutions of PSGN and P2M predictions.

TABLE 1. Comparison of X2V and P2M in terms of IoU, Chamfer-L1, F-score and NC metrics for mesh reconstruction accuracy.

Method	*IoU*	*Chamfer-L1*	*F-score(t=0.02)*	*NC*
P2M-L	0.857 ± 0.042	0.013 ± 0.001	0.902 ± 0.053	0.708 ± 0.092
P2M-R	0.849 ± 0.045	0.011 ± 0.001	0.891 ± 0.053	0.685 ± 0.053
X2V	0.922 ± 0.028	0.010 ± 0.001	0.964 ± 0.030	0.930 ± 0.013

TABLE 2. Comparison of X2V and PSGN in terms of IoU, Chamfer-L1, F-score and NC metrics for mesh reconstruction accuracy.

Method	*IoU*	*Chamfer-L1*	*F-score(t=0.02)*	*NC*
PSGN	0.699 ± 0.059	0.016 ± 0.002	0.802 ± 0.071	0.860 ± 0.012
X2V	0.892 ± 0.035	0.013 ± 0.001	0.853 ± 0.055	0.920 ± 0.012

In the assessment, lower values of Chamfer-L1 are indicative of improved performance, whereas for the IoU, F-score and NC, higher values are favorable. Our approach outperforms the two state-of-the-art methods across each of these metrics. In evaluating mesh quality using IoU, Chamfer-L1, and F-score metrics, the mean and standard deviation offer key insights. The mean calculation involves summing each metric's values across the test dataset and dividing by the number of observations, indicating average performance. The standard deviation, calculated as the square root of the averaged squared deviations from the mean, gauges result variability. A lower standard deviation suggests consistent mesh quality, while a higher one indicates variability. These statistics collectively provide a succinct overview of mesh evaluation, highlighting both average performance and consistency in terms of IoU, Chamfer-L1, F-score and NC metrics.

In Table 3, X2V assessment of lung meshes at higher resolution (10,000 points) reveals significant enhancement in IoU metrics, surpassing previous research benchmarks. Specifically, the IoU attains an impressive score of 0.94, exceeding the prior standard of 0.88 in [6]. This improvement is indicative of our model's superior accuracy in capturing the overlapping regions of lung meshes, demonstrating a considerable advancement over existing methodologies.

TABLE 3. X2V performance evaluation at different mesh resolutions.

Metric	*X2V* (10,000 pts)	*X2V* (4,932 pts)	*X2V* (2,048 pts)
IoU	0.942 ± 0.025	0.922 ± 0.028	0.892 ± 0.035
Chamfer-L1	0.008 ± 0.001	0.010 ± 0.001	0.013 ± 0.001
F-score (t=0.02)	0.976 ± 0.024	0.964 ± 0.030	0.853 ± 0.055
NC	0.938 ± 0.012	0.930 ± 0.013	0.920 ± 0.012

The F-score, which measures model accuracy, displays remarkable stability especially at medium-to-high resolutions. The consistency of an F-score of about 0.97, whether at a resolution of 4,932 points or the higher resolution of 10,000 points, emphasizes the reliability of our model. Such uniformity in performance, irrespective of mesh detail, underscores the model's robustness, a critical attribute for precision in complex simulations. Additionally, our model achieves an exceptional NC score of 0.938. As a pivotal metric for evaluating the alignment precision of mesh normals, this high score reflects a significant level of congruence in our model's mesh normals. It corroborates the model's effectiveness in accurately replicating the intricate geometric structures of lung tissues, thereby reinforcing its applicability in detailed anatomical simulations.

As for the time complexity, X2V model, utilizing marching cubes, exhibits an average inference time of 0.21 s, which is slightly higher in comparison to the baseline algorithms, where PSGN operates at 0.08 s and Pixel2Mesh at 0.17 s.

VI. CONCLUSION

Our X2V model has distinctly established itself as a superior tool in the domain of single-image 3D reconstruction, particularly in generating lung meshes, outperforming existing methods like PSGN and P2M in crucial metrics like IoU, Chamfer-L1 distance, F-Score and normal consistency. Its exceptional performance is most notable in high-resolution mesh evaluations, where it achieves an IoU of 0.94 and maintains consistent F-scores across different resolutions. This success underscores X2V's adeptness in accurately capturing complex geometries, a key aspect for realistic anatomical

simulations. The model's high normal consistency score further reinforces its ability to precisely replicate the intricate structures of lung tissues.

Future enhancements of the X2V approach will primarily focus on expanding its applications beyond lung tissue reconstruction to a broader range of complex anatomical structures. This extension aims to evaluate and boost the model's versatility and broaden its application in medical simulations. Efforts will also be directed towards fostering closer collaborations with medical professionals to incorporate clinical data, enhancing the model's real-world accuracy and relevance. Another significant area of development will involve integrating X2V with various imaging modalities, such as MRI, CT scans, and ultrasound. This integration seeks to offer a more comprehensive view of internal anatomy, thereby improving the efficacy of diagnosis and treatment planning. Furthermore, the integration of automated features for detecting pathological conditions in segmented organs will be a key goal. This feature aims to facilitate early diagnosis and timely medical intervention, potentially improving patient outcomes. These strategic enhancements are poised to further establish the X2V model as an invaluable asset in the field of medical imaging and analysis.

ACKNOWLEDGMENT

During the preparation of this work the authors used ChatGPT language model in order to rephrase author-written text into more concise and grammarly correct content. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

[1] M. K. Islam, T. G. Purdie, B. D. Norrlinger, H. Alasti, D. J. Moseley, M. B. Sharpe, J. H. Siewerdsen, and D. A. Jaffray, "Patient dose from kilovoltage cone beam computed tomography imaging in radiation therapy," *Med. Phys.*, vol. 33, no. 6, pp. 1573–1582, Jun. 2006, doi: 10.1118/1.2198169.

[2] W. Y. Song, S. Kamath, S. Ozawa, S. Al Ani, A. Chvetsov, N. Bhandare, J. R. Palta, C. Liu, and J. G. Li, "A dose comparison study between XVI and OBI CBCT systems," *Med. Phys.*, vol. 35, no. 2, pp. 480–486, Feb. 2008, doi: 10.1118/1.2825619.

[3] M. Unberath, C. Gao, Y. Hu, M. Judish, R. H. Taylor, M. Armand, and R. Grupp, "The impact of machine learning on 2D/3D registration for image-guided interventions: A systematic review and perspective," *Frontiers Robot. AI*, vol. 8, Aug. 2021, doi: 10.3389/frobt.2021.716007.

[4] P. Markelj, D. Tomaževič, B. Likar, and F. Pernuš, "A review of 3D/2D registration methods for image-guided interventions," *Med. Image Anal.*, vol. 16, no. 3, pp. 642–661, Apr. 2012, doi: 10.1016/j.media.2010.03.005.

[5] S. Wu, M. Nakao, J. Tokuno, T. Chen-Yoshikawa, and T. Matsuda, "Reconstructing 3D lung shape from a single 2D image during the deaeration deformation process using model-based data augmentation," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Inform. (BHI)*, May 2019, pp. 1–4, doi: 10.1109/BHI.2019.8834454.

[6] Y. Wang, Z. Zhong, and J. Hua, "DeepOrganNet: On-the-fly reconstruction and visualization of 3D/4D lung models from single-view projections by deep deformation network," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 1, pp. 960–970, Jan. 2020, doi: 10.1109/TVCG.2019.2934369.

[7] F. Tong, M. Nakao, S. Wu, M. Nakamura, and T. Matsuda, "X-ray2Shape: Reconstruction of 3D liver shape from a single 2D projection image," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 1608–1611, doi: 10.1109/EMBC44109.2020.9176655.

[8] M. Nakao, F. Tong, M. Nakamura, and T. Matsuda, "Image-to-graph convolutional network for deformable shape reconstruction from a single projection image," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, vol. 12904, Sep. 2021, pp. 259–268, doi: 10.1007/978-3-030-87202-1_25.

[9] M. Nakao, M. Nakamura, and T. Matsuda, "Image-to-graph convolutional network for 2D/3D deformable model registration of low-contrast organs," *IEEE Trans. Med. Imag.*, vol. 41, no. 12, pp. 3747–3761, Dec. 2022, doi: 10.1109/TMI.2022.3194517.

[10] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3D reconstruction in function space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4455–4465, doi: 10.1109/CVPR.2019.00459.

[11] P. Maken and A. Gupta, "2D-to-3D: A review for computational 3D image reconstruction from X-ray images," *Arch. Comput. Methods Eng.*, vol. 30, no. 1, pp. 85–114, Jan. 2023, doi: 10.1007/s11831-022-09790-z.

[12] B. Goswami and S. Kr., "3D modeling of X-ray images: A review," *Int. J. Comput. Appl.*, vol. 132, no. 7, pp. 40–46, Dec. 2015.

[13] M. S. U. Khan, A. Pagani, M. Liwicki, D. Stricker, and M. Z. Afzal, "Three-dimensional reconstruction from a single RGB image using deep learning: A review," *J. Imag.*, vol. 8, no. 9, p. 225, Aug. 2022, doi: 10.3390/jimaging8090225.

[14] J. Bednarik, P. Fua, and M. Salzmann, "Learning to reconstruct texture-less deformable surfaces from a single view," in *Proc. Int. Conf. 3D Vis.*, Sep. 2018, pp. 606–615, doi: 10.1109/3DV.2018.00075.

[15] A. Tsoli and Antonis. A. Argyros, "Patch-based reconstruction of a textureless deformable 3D surface from a single RGB image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 4034–4043, doi: 10.1109/ICCVW.2019.00498.

[16] V. Golyanik, S. Shimada, K. Varanasi, and D. Stricker, "HDM-Net: Monocular non-rigid 3D reconstruction with learned deformation model," in *Proc. 15th EuroVR Int. Conf., Virtual Reality Augmented Reality*, Oct. 2018, pp. 51–72.

[17] S. Shimada, V. Golyanik, C. Theobalt, and D. Stricker, "IsMo-GAN: Adversarial learning for monocular non-rigid 3D reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 2876–2885.

[18] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y. G. Jiang, "Pixel2Mesh: Generating 3D mesh models from single RGB images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 52–71, doi: 10.1007/978-3-030-01252-6_4.

[19] Y. Yuan, J. Tang, and Z. Zou, "VANet: A view attention guided network for 3D reconstruction from single and multi-view images," in *Proc. IEEE Int. Conf. Multimedia and Expo. (ICME)*, Jul. 2021, pp. 1–6, doi: 10.1109/ICME51207.2021.9428171.

[20] A. Salvi, N. Gavenski, E. Pooch, F. Tasoniero, and R. Barros, "Attention-based 3D object reconstruction from a single image," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8, doi: 10.1109/IJCNN48605.2020.9206776.

[21] X. Chen, A. Diaz-Pinto, N. Ravikumar, and A. Frangi, "Deep learning in medical image registration," *Prog. Biomed. Eng.*, vol. 3, no. 1, Feb. 2021, Art. no. 012003, doi: 10.1088/2516-1091/abd37c.

[22] L. Humbert, J. A. De Guise, B. Aubert, B. Godbout, and W. Skalli, "3D reconstruction of the spine from biplanar X-rays using parametric models based on transversal and longitudinal inferences," *Med. Eng. Phys.*, vol. 31, no. 6, pp. 681–687, Jul. 2009, doi: 10.1016/j.medengphy.2009.01.003.

[23] Y. Chaibi, T. Cresson, B. Aubert, J. Hausselle, P. Neyret, O. Hauger, J. A. de Guise, and W. Skalli, "Fast 3D reconstruction of the lower limb using a parametric model and statistical inferences and clinical measurements calculation from biplanar X-rays," *Comput. Methods Biomech. Biomed. Eng.*, vol. 15, no. 5, pp. 457–466, May 2012, doi: 10.1080/10255842.2010.540758.

[24] T. Cresson, D. Branchaud, R. Chav, B. Godbout, and J. A. de Guise, "3D shape reconstruction of bone from two X-ray images using 2D/3D non-rigid registration based on moving least-squares deformation," *Proc. SPIE*, vol. 7623, Mar. 2010, Art. no. 76230F, doi: 10.1117/12.844098.

[25] J. Zhang, L. Lv, X. Shi, Y. Wang, F. Guo, Y. Zhang, and H. Li, "3-D reconstruction of the spine from biplanar radiographs based on contour matching using the Hough transform," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 7, pp. 1954–1964, Jul. 2013.

[26] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 77–85, doi: 10.1109/CVPR.2017.16.
[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Represent.*, 2021.
[28] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," 2020, *arXiv:2006.03677*.
[29] J. Deng, W. Dong, R. Socher, L. -J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.
[30] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.
[31] H. D. Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. Courville, "Modulating early visual processing by language," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
[32] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015.
[33] M. Garland and P. S. Heckbert, "Simplifying surfaces with color and texture using quadric error metrics," in *Proc. Visualizat.*, Oct. 1998, pp. 263–269, doi: 10.1109/VISUAL.1998.745312.
[34] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem," *Eur. Radiol. Experim.*, vol. 4, no. 1, Aug. 2020, doi: 10.1186/s41747-020-00173-2.
[35] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imag.*, vol. 26, no. 6, pp. 1045–1057, Dec. 2013, doi: 10.1007/s10278-013-9622-7.
[36] National Lung Screening Trial Research Team, (2013), "Data from the national lung screening trial (NLST) [data set]," *The Cancer Imaging Archive*, doi: 10.7937/TCIA.HMQ8-J677.
[37] National Lung Screening Trial Research Team, D. R. Aberle, A. M. Adams, C. D. Berg, W. C. Black, J. D. Clapp, R. M. Fagerstrom, I. F. Gareen, C. Gatsonis, P. M. Marcus, and J. D. Sicks, "Reduced lung-cancer mortality with low-dose computed tomographic screening," *New England J. Med.*, vol. 365, no. 5, pp. 395–409, 2011, doi: 10.1056/NEJMoa1102873.
[38] R. Kikinis, S. D. Pieper, and K. G. Vosburgh, "3D slicer: A platform for subject-specific image analysis, visualization, and clinical support," in *Intraoperative Imaging and Image-Guided Therapy*. New York, NY, USA: Springer, 2013, pp. 277–289, doi: 10.1007/978-1-4614-7657-3_19.
[39] T. Kapur et al., "Increasing the impact of medical image computing using community-based open-access hackathons: The NA-MIC and 3D slicer experience," *Med. Image Anal.*, vol. 33, pp. 176–180, Oct. 2016, doi: 10.1016/j.media.2016.06.035.
[40] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, J. Buatti, S. Aylward, J. V. Miller, S. Pieper, and R. Kikinis, "3D slicer as an image computing platform for the quantitative imaging network," *Magn. Reson. Imag.*, vol. 30, no. 9, pp. 1323–1341, Nov. 2012, doi: 10.1016/j.mri.2012.05.001.
[41] S. Pieper, B. Lorensen, W. Schroeder, and R. Kikinis, "The NA-MIC kit: ITK, VTK, pipelines, grids and 3D slicer as an open platform for the medical image computing community," in *Proc. 3rd IEEE Int. Symp. Biomed. Imag., Macro Nano*, May 2006, pp. 698–701, doi: 10.1109/ISBI.2006.1625012.
[42] R. Bumm. *Lung CT Analyzer*. Accessed: Jul. 7, 2023. [Online]. Available: https://github.com/rbumm/SlicerLungCTAnalyzer
[43] R. L. Siddon, "Fast calculation of the exact radiological path for a three-dimensional CT array," *Med. Phys.*, vol. 12, no. 2, pp. 252–255, Mar. 1985, doi: 10.1118/1.595715.
[44] M. de Greef, J. Crezee, J. C. van Eijk, R. Pool, and A. Bel, "Accelerated ray tracing for radiotherapy dose calculations on a GPU," *Med. Phys.*, vol. 36, no. 9, pp. 4095–4102, Sep. 2009.
[45] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3D-R2N2: A unified approach for single and multi-view 3D object reconstruction," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 628–644, doi: 10.1007/978-3-319-46484-8_38.
[46] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3D object reconstruction from a single image," in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 605–613, doi: 10.1109/CVPR.2017.264.
[47] F. Bernardini, J. Mittleman, H. Rushmeier, C. Silva, and G. Taubin, "The ball-pivoting algorithm for surface reconstruction," *IEEE Trans. Vis. Comput. Graphics*, vol. 5, no. 4, pp. 349–359, Oct. 1999, doi: 10.1109/2945.817351.

GOKCE GUVEN received the B.S. degree in chemistry from Koc University, Istanbul, in 2010, where her research was focused on superconductor modeling, and the M.Sc. degree in materials science and nanoengineering from Sabanci University, Istanbul, in 2013. She is currently pursuing the Ph.D. degree with Özyeğin Üniversitesi, Istanbul, Turkey. Her Ph.D. research is dedicated to leveraging deep learning techniques for medical imaging, with the goal of enhancing diagnostic procedures through technological advancements. Her professional experience began from 2009 to 2010 as a Research Assistant with the Max Planck Institute for Chemical Physics of Solids. From 2015 to 2019, she expanded her expertise to include software development and research, working with various companies. Since 2020, she has been the Chief Technology Officer of Osteoid Health Technologies.

HASAN F. ATES (Senior Member, IEEE) received the Ph.D. degree from the Department of Electrical Engineering, Princeton University, in 2004. He was a Research Associate with Sabanci University, from 2004 to 2005. He held positions of an Assistant, an Associate, and a Full Professorship with Isik University, from 2005 to 2018. He was with Istanbul Medipol University, from 2018 to 2022. Since September 2022, he has been a Professor with the Department of Computer Science, Özyeğin Üniversitesi. He is the author/coauthor of more than 80 peer-reviewed publications in the areas of image/video processing/coding and computer vision.

H. FATIH UGURDAG (Senior Member, IEEE) received the B.S. degree in EE and physics from Boğaziçi University, Istanbul, Turkey, in 1986, and the M.S. and Ph.D. degrees in EE from Case Western Reserve University, Cleveland, OH, USA, in 1989 and 1995, respectively. He was involved full-time with industry for 13 years before joining academia, in 2004. His tenure in the industry was mostly in silicon valley and spanned companies, such as GE, GM, Lucent, Juniper, and Nvidia. He is currently a Full Professor with Özyeğin Üniversitesi, Istanbul, Turkey. His research interests include ASIC/SoC/FPGA implementation of various compute-intensive algorithms, design automation, and computer arithmetic. He is a long-time IEEE volunteer and is currently the Secretary of the IEEE Turkey Section.