

**DEEP LEARNING TECHNIQUES
FOR 3D VOLUME RECONSTRUCTION
FROM PLANAR X-RAY IMAGES**

GÖKÇE GÜVEN

ÖZYEGİN UNIVERSITY

MARCH, 2025

DEEP LEARNING TECHNIQUES FOR 3D VOLUME RECONSTRUCTION FROM PLANAR X-RAY IMAGES

by
Gökçe Güven

Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy

in
Computer Science

Advisor: Prof. Hasan F. Ateş

Graduate School of Science and Engineering
Özyeğin University
İstanbul

March, 2025

DEEP LEARNING TECHNIQUES FOR 3D VOLUME RECONSTRUCTION FROM PLANAR X-RAY IMAGES

Approved by:

Prof. Hasan F. Ateş, Advisor
Department of Artificial Intelligence and
Data Engineering
Özyeğin University

Prof. Sezer Gören Uğurdağ
Department of Electrical & Computer
Engineering
University of Massachusetts Dartmouth

Prof. H. Fatih Uğurdağ
Department of Electrical & Electronics
Engineering
Özyeğin University

Prof. Bahadır K. Güntürk
Department of Electrical & Electronics
Engineering
Istanbul Medipol University

Prof. A. Tanju Erdem
Department of Computer Science
Özyeğin University

Approval Date: January 20, 2025

To my family, friends, and mentors...



DECLARATION OF ORIGINALITY

I hereby declare that I am the sole author of this thesis and that this is the true copy of my thesis, including the final revisions, approved by my thesis committee. All data and information have been obtained, produced, and presented in accordance with the rules of research ethics and principles of academic honesty. As required by these rules, to the best of my knowledge I have acknowledged ideas, thoughts, and any copyrighted material in accordance with the standard referencing rules. I certify that any part of this thesis has not been submitted for a degree or diploma in another educational institution.

Gökçe Güven

ABSTRACT

X-ray is most often the preferred form of medical imaging due to its low cost. However, being 2D, it is not as effective as 3D imaging techniques when it comes to diagnosis. On the other hand, 3D imaging is expensive and hence not as ubiquitous as X-ray, with computed tomography posing higher radiation risks linked to tissue damage. Having said that, 3D reconstruction (3Dr) from X-ray seems like a plausible compromise. Conventional 3Dr methods, which rely on statistical templates for 2D/3D registration, often fail to account for patient-specific anatomical variations, leading to inaccuracies. This thesis introduces two novel deep learning techniques for 3Dr from a single X-ray, namely, X-ray to Volume (X2V) and X-ray to Bone (X2B). X2V, incorporating a vision transformer based encoder, reconstructs highly accurate 3D organ models, which capture individual anatomical variability. X2B applies this methodology to skeletal structures, addressing challenges such as overlapping Hounsfield Unit values in X-rays, which hinder the extraction of fine structural details. Furthermore, as extensions to X2B, X-ray to Bone Registration (X2BR) and Biplanar X-ray to Bone (X2B2) techniques integrate object detection, neural implicit modeling, and non-rigid registration to generate high-resolution 3Dr of ribs and vertebrae. We show that all four techniques are superior than the state of the art in the literature through chamfer distance, intersection over union, f-score, normal consistency metrics. By bridging the gap between traditional statistical methods and modern deep learning, this thesis enables more personalized and precise diagnostic and therapeutic solutions, with the potential of improving clinical workflows and patient outcomes.

Keywords: X-ray, 3D reconstruction, neural implicit methods

ÖZET

X-ray, düşük maliyeti nedeniyle genellikle tıbbi görüntülemede en çok tercih edilen yöntemdir. Ancak, 2 boyutlu olması sebebiyle teşhis aşamasında 3 boyutlu görüntüleme teknikleri kadar etkili değildir. Öte yandan, 3 boyutlu görüntüleme maliyetli olduğundan X-ray kadar yaygın değildir ve bilgisayarlı tomografi, doku hasarıyla ilişkili daha yüksek radyasyon riskleri taşır. Bununla birlikte, X-ray'den 3 boyutlu rekonstrüksiyon (3Dr) yapmak makul bir orta yol gibi görünmektedir. İstatistiksel şablonlar aracılığıyla 2B/3B eşleştirmeye dayanan geleneksel 3Dr yöntemleri, hastaya özgü anatomik değişkenlikleri çoğu zaman göz önünde bulunduramadıkları için doğruluk sorunlarına yol açmaktadır. Bu tez, tek bir X-ray üzerinden 3Dr için iki yeni derin öğrenme tekniği tanıtmaktadır: X-ray to Volume (X2V) ve X-ray to Bone (X2B). Görüntü dönüştürücü tabanlı bir kodlayıcı içeren X2V, bireysel anatomik farklılıkları yakalayabilen son derece hassas 3B organ modelleri üretmektedir. X2B ise bu yaklaşımı iskelet yapılarına uygulayarak, X-ray'lerdeki örtüşen Hounsfield Birimi (HU) değerleri gibi ince yapısal detayların çıkarılmasını zorlaştıran sorunları ele almaktadır. Ayrıca, X2B'nin genişletmeleri olarak sunulan X-ray to Bone Registration (X2BR) ve Biplanar X-ray to Bone (X2B2) teknikleri, kaburgalar ve omurların yüksek çözünürlüklü 3Dr'ını elde etmek amacıyla nesne tespiti, nöral örtük modelleme ve esnek eşleştirme işlemlerini içerir. Chamfer mesafesi, kesişimin birleşime oranı, f-skoru ve normal tutarlılığı metrikleri üzerinden yapılan değerlendirmeler, bu dört tekniğin literatürdeki mevcut yöntemlere kıyasla üstün olduğunu göstermektedir. Geleneksel istatistiksel yöntemlerle modern derin öğrenme arasındaki boşluğu kapatan bu tez, daha kişiselleştirilmiş ve kesin teşhis ile tedavi çözümlerine zemin hazırlayarak, klinik iş akışlarını ve hasta sonuçlarını iyileştirme potansiyeline sahiptir.

Anahtar Kelimeler: X-ray, 3B rekonstrüksiyon, nöral örtük yöntemler

ACKNOWLEDGEMENTS

The completion of this doctoral dissertation marks the culmination of several years of intensive study, research, and personal growth. It is a journey that would not have been possible without the guidance, support, and encouragement of many individuals and organizations, to whom I am profoundly indebted.

First and foremost, I express my deepest gratitude to my advisor, Prof. Hasan F. Ateş, whose exceptional mentorship and profound expertise have been the cornerstone of this research. His guidance, insightful suggestions, and unwavering encouragement have been instrumental in shaping the trajectory of this dissertation and ensuring its academic rigor. His ability to inspire critical thinking and foster intellectual curiosity has profoundly influenced my development as a researcher.

I am deeply grateful to Prof. H. Fatih Uğurdağ for his invaluable feedback, constructive criticism, and unwavering support throughout my doctoral journey. At a pivotal moment when I believed my academic path had come to an end, his belief in my potential provided me with the opportunity to restart and ultimately complete my PhD. His encouragement, thoughtful suggestions, and dedication have not only enriched this dissertation but also played a critical role in shaping its practical contributions. For his guidance and trust, I am profoundly thankful. My sincerest thanks also go to the esteemed members of my advisory committee, Prof. Sezer Gören Uğurdağ and Prof. A. Tanju Erdem, whose guidance, critical insights, and academic rigor have been vital in enhancing the quality and relevance of this research.

This research was made possible by the generous financial and technical support of Osteoid A.Ş. I am deeply appreciative of the resources, facilities, and collaboration opportunities provided by this organization, which have been instrumental in the successful

completion of this study. Their contributions not only enabled this work but also highlighted the importance of bridging academia and industry in advancing knowledge and innovation.

I would also like to extend my gratitude to the administrative staff of Özyeğin University, whose assistance and support have played a crucial role in facilitating my research activities. Their dedication and professionalism ensured a conducive environment for academic pursuit.

On a personal note, I owe an immeasurable debt of gratitude to my family. Their unwavering love, patience, and belief in my abilities have provided me with the strength and determination to overcome challenges throughout this journey. Their sacrifices and encouragement have been the bedrock of my success, and I am forever grateful for their presence in my life.

I am also deeply thankful to my dear friend, Ekin Özgönül, whose steadfast support, humor, and understanding have been invaluable during the most demanding phases of this endeavor. Her companionship and encouragement have been a source of comfort and resilience, making this journey more manageable and rewarding.

Lastly, I dedicate this dissertation to Deniz Kardeşahin, whose passion for innovation, unwavering belief in my potential, and unconditional support have been a guiding light throughout my academic and personal journey. His encouragement, insights, and love have inspired me to persevere and strive for excellence. This work stands as a testament to our shared aspirations and the strength of our partnership.

To all those who have supported me in countless ways, both big and small, I extend my deepest thanks and heartfelt appreciation. This achievement is as much yours as it is mine.

TABLE OF CONTENTS

DECLARATION OF ORIGINALITY	iv
ABSTRACT	v
ÖZET	vi
ACKNOWLEDGEMENTS	vii
LIST OF TABLES	xi
LIST OF FIGURES	xii
LIST OF ACRONYMS AND ABBREVIATIONS	xvii
1 INTRODUCTION	1
1.1 Radiography	4
1.1.1 X-ray Imaging	4
1.1.2 Computed Tomography	6
1.2 Neural Fields	7
1.2.1 Occupancy Field	9
1.2.2 Signed Distance Fields	10
1.3 Non-rigid Registration	12
1.4 Contributions	13
2 PREVIOUS WORK	17
2.1 Single Image 2D/3D Reconstruction Methods	18
2.2 Single-View Reconstruction Using Implicit Functions	19
2.3 Neural Implicit Methods in Medical Imaging	21
2.4 2D/3D Registration-Based Approaches	22
2.5 Deformation Learning Based Approaches	23
2.6 3D Bone Reconstruction from X-rays	25
2.7 Non-Rigid Registration Methods	26
3 3D ORGAN RECONSTRUCTION FROM A SINGLE X-RAY	29

3.1	X2V Model Architecture	30
3.1.1	Training	34
3.1.2	Inference	35
3.2	Dataset	36
3.2.1	Segmentation with Lung Analyzer	36
3.2.2	Calculation of Occupancy Values.....	39
3.3	Experiments	40
3.3.1	Comparative Studies	40
3.3.2	Evaluation Metrics	42
3.4	Simulation Results	45
3.4.1	Simulation Results and Discussions	45
4	3D BONE RECONSTRUCTION FROM A SINGLE X-RAY	53
4.0.1	X2B and X2BR Model Architectures	54
4.0.2	Training	62
4.0.3	Inference	63
4.0.4	Registration	64
4.1	Dataset	65
4.1.1	Segmentation with TotalSegmentator	65
4.1.2	Bone Localization and Detection	67
4.1.3	DRR Specific Template Model	68
4.1.4	Calculation of Occupancy Values.....	69
4.2	Experiments	70
4.2.1	Comparative Studies	70
4.2.2	Evaluation Metrics	71
4.3	Simulations and Results	74
4.4	Ablation Study	88
5	CONCLUSIONS	92
	REFERENCES	95
	VITA	104

LIST OF TABLES

Table 3.1: Comparison of X2V and P2M in terms of IoU, Chamfer-L1, F-score and NC metrics for mesh reconstruction accuracy.	50
Table 3.2: Comparison of X2V and PSGN in terms of IoU, Chamfer-L1, F-score and NC metrics for mesh reconstruction accuracy.....	50
Table 3.3: X2V performance evaluation at different mesh resolutions.....	51
Table 4.1: Comparison of D2IM, ED2IF2, X2V, X2B, and X2BR, in terms of maze, maxe and maye metrics in milimeter for mesh reconstruction accuracy.....	80
Table 4.2: Comparison of D2IM, ED2IF2, X2V, X2B, X2BR in terms of IoU, Chamfer-L1, F-score and NC metrics for mesh reconstruction accuracy.	81
Table 4.3: Comparison of GBCPD++, NDP, CPD in terms of IoU, Chamfer-L1, F-score and NC metrics for mesh reconstruction accuracy.	86
Table 4.4: Comparison of X2B and X2B2 results in terms of IoU, Chamfer-L1, F-score and NC metrics for mesh reconstruction accuracy.....	89
Table 4.5: Comparison of X2B and X2B2 results in terms of maze, maxe and maye metrics in milimeter for mesh reconstruction accuracy.....	89

LIST OF FIGURES

- Figure 1.1:** Schematic representation of radiographic distances. The figure illustrates the key radiographic parameters, including the Source-to-Image Distance (SID) and the Source-to-Object Distance (SOD). SID represents the distance between the X-ray tube and the detector, while SOD denotes the distance between the X-ray tube and the patient. 5
- Figure 1.2:** Visualization of the occupancy field for the Stanford Bunny. The figure illustrates the spatial representation of the Stanford Bunny using an occupancy field, where red and blue points represent sampled locations in the 3D space. The occupancy field encodes whether each sampled point lies inside or outside the surface of the mesh, enabling the implicit modeling of the object’s geometry. 10
- Figure 1.3:** SDF projections of the Stanford Bunny. The figure shows the SDF of the Stanford Bunny, visualized as three orthogonal 2D projections: XY, XZ, and YZ planes. The SDF encodes the distance of each point in the field to the nearest surface of the mesh, with the sign indicating whether the point lies inside (negative) or outside (positive) the surface. The red regions represent areas closer to the mesh surface, while blue regions are farther away. The white dashed lines delineate the contours of the bunny’s mesh in each projection. 11
- Figure 3.1:** The X2V 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 (CBN), 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 3Dr from diverse inputs..... 31

Figure 3.2:	For 3D 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.	32
Figure 3.3:	a) CT; b) DRR; c) triangular mesh representations of a patient’s thorax region.	38
Figure 3.4:	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 labeled as internal, shown by a blue dot.	40
Figure 3.5:	Real-size lung volume analysis of GT and X2V network output: (a) Single anteroposterior DRR with CLAHE of lungs as input to networks; (b) GT 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) GT 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 -0.33 mm and std 2.95 mm).	46
Figure 3.6:	Distribution of point-to-point distances for 639 test images. The mean distance is -0.5 mm, with a standard deviation of 2.95 mm. The histogram provides a visual representation of the frequency of distances, with the red dashed line indicating the mean and green dashed lines showing ± 1 standard deviation from the mean.	47
Figure 3.7:	Comparison of lung CT reconstructions. (a) GT lung mesh. (b) Lung mesh generated by the X2V network.	48

Figure 3.8:	Comparison for lung volume analysis. (a) Single antero-posterior DRR of lungs as input to networks. (b) GT 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.	49
Figure 4.1:	X2B network training pipeline. The figure illustrates the training process of the X2B network, which takes a DRR as input and uses a ConvNeXt backbone for feature extraction. The extracted features are passed through dense blocks with Conditional Batch Normalization (CBN) layers, parameterized by β_i and γ_i , to refine the latent representations.	56
Figure 4.2:	Anterior-posterior DRR is used as input for both YOLOv9 and X2B for inference. The regions are detected via YOLOv9 network and DRR specific template mesh is extracted from the template mesh model using the detected regions. DRR specific template model is registered to the course shape, which is the output of the X2BR model.	59
Figure 4.3:	DRR images for multi-view training of the X2B network. The figure displays a collection of DRR images captured at various angles around the anatomical structure. These images serve as input data for the X2B network, enabling it to learn multi-view representations for accurate 3Dr and enhanced skeletal analysis. The diverse viewing angles facilitate robust feature extraction and improve the network’s ability to generalize across different spatial perspectives.	60
Figure 4.4:	YOLOv9 region labeling for vertebral detection. The figure illustrates the application of the YOLOv9 object detection model for identifying and labeling vertebral regions in a chest X-ray. Each bounding box highlights a detected vertebra, facilitating precise localization and segmentation. This labeling process is critical for downstream tasks, such as generating person-specific templates for 3Dr.	61

Figure 4.5:	For 3D mesh inference with <i>X2B</i> 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.	64
Figure 4.6:	SKEL BSM to template mesh reconstruction. This figure illustrates the process of generating a subject-specific skeletal template model. The SKEL Body Shape Model (BSM) on the left is used as the baseline for anatomical structures, which is further processed to generate a template mesh, as shown on the right.	68
Figure 4.7:	Comparison of reconstruction results across different methods. The figure presents DRR images (leftmost column) and their corresponding GT 3D meshes alongside reconstructed outputs from various methods: D2IM, ED2IF2, X2V, X2B, and X2BR. Each row corresponds to a different DRR input, while the columns illustrate the progression of reconstruction quality across the methods.	75
Figure 4.8:	Comparison of X2B and GT. The first column displays the DRR inputs, while the second column presents the GT meshes. The subsequent columns show the X2B reconstructions (X2B-f), X2B heatmaps (X2B-h-f), GT-b, X2B-b, and X2B heatmaps in the back view (X2B-h-b).	78
Figure 4.9:	Distribution of point-to-point distances for 364 test images. The mean distance is 2.192 mm, with a standard deviation of 1.886 mm. The histogram provides a visual representation of the frequency of distances, with the red dashed line indicating the mean and the green dashed lines representing ± 1 standard deviation from the mean.	79
Figure 4.10:	Comparison of registration results for rib cage models using three methods: GBCPD++, CPD, and NDP. The figure presents front views of the rib cage models. The first column displays the X2B network predicted mesh models, while the subsequent columns illustrate the registered results from each method (GBCPD++, CPD, and NDP).	83

- Figure 4.11:** Comparison of registration results for rib cage models using three methods: GBCPD++, CPD, and NDP. The figure presents side views of the rib cage models. The first column displays the X2B network predicted mesh models. The subsequent columns illustrate the registered results obtained from each method, GBCPD++, CPD, and NDP... 85
- Figure 4.12:** Thorax reconstruction comparison across different models. The rows correspond to the X2B reconstructions, NDP, CPD, and GBCPD++ registrations, respectively. Each row contains six images, where the first, third, and fifth images display the full thorax in different perspectives (front view, side view, and cross-section), and the second, fourth, and sixth images provide zoomed-in views of specific regions to highlight finer details. 87
- Figure 4.13:** X2B2 network training pipeline. This figure demonstrates the training process of the X2B2 network, which processes bi-planar DRRs as input. Each DRR is encoded using a ConvNeXt backbone, and their features are fused using a cross-attention mechanism to extract complementary information. The fused features are refined through dense blocks with Conditional Batch Normalization (CBN) layers, parameterized by β_i and γ_i , for accurate latent representation..... 88
- Figure 4.14:** Thorax reconstruction comparison across different models. The rows correspond to the GT, X2B reconstructions, and X2B2 reconstructions, respectively. Each row contains six images, where the first, third, and fifth images display the full thorax in different perspectives (front view, side view, and cross-section), and the second, fourth, and sixth images provide zoomed-in views of specific regions to highlight finer details. 91

LIST OF ACRONYMS AND ABBREVIATIONS

3Dr	3D Reconstruction
ACM	Association for Computing Machinery
AI	Artificial Intelligence
AP	Anterior-Posterior
AR	Augmented Reality
BCPD	Bayesian Coherent Point Drift
BSM	Biomechanical Skeleton Model
CAD	Computer Aided Design
CBCT	Cone Beam Computer Tomography
CBN	Conditional Batch Normalization
CD	Chamfer Distance
CNN	Convolutional Neural Network
COCO	Common Objects in Context
CT	Computer Tomography
CVF	Computer Vision Foundation
ED2IF2-Net	Disentangled Deformed Implicit Fields and Enhanced Displacement Fields Network
D2IM-Net	Detail Disentangled Implicit Fields Network

DICOM	Digital Imaging and Communications in Medicine
DISN	Deep Implicit Surface Network
DRR	Digitally Reconstructed Radiograph
EMBS	Engineering Medicine & Biology Society
FEA	Finite Element Analysis
GAN	Generative Adversarial Networks
GBCPD	Geodesic Based Coherent Point Drift
GCN	Graph Convolutional Network
GELU	Gaussian Error Linear Unit
GMM	Gaussian Mixture Model
GPU	Graphics Processing Unit
GT	Ground Truth
GTX	Graphics Technology Extreme
HDM	Hybrid Deformation Model
HU	Hounsfield Unit
ICP	Iterative Closest Point
IEEE	Institute of Electrical and Electronics Engineers
IGCN	Image to Graph Convolutional Network
IGRT	Image Guided Radiation Therapy

IoU	Intersection over Union
ITK	Insight Toolkit
MDISN	Multiscale Deformed Implicit Surface Network
MISE	Multiscale Isosurface Extraction
ML	Machine Learning
MLP	Multi Layer Perceptron
MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
NA-MIC	National Alliance for Medical Image Computing
NBIA	National Biomedical Imaging Archive
NC	Normal Consistency
NDP	Neural Deformation Pyramids
NET	Network
NLST	National Lung Screening Trial
OBI	On-Board Imaging
P2M	Pixel to Mesh
PSGN	Point Set Generation Network
PVT	Pyramid Vision Transformer
ReLU	Rectified Linear Unit

RGB	Red Green Blue
SDF	Signed Distance Function
SID	Source to Image Distance
SKEL	From Skin to Skeleton
SNAF	Sparse-view CBCT Reconstruction with Neural Attenuation Fields
SOD	Source to Object Distance
SSM	Statistical Shape Model
SDF	Signed Distance Field
TCIA	The Cancer Imaging Archive
VGG	Visual Geometry Group
ViT	Vision Transformer
VR	Virtual Reality
VTK	Visualization Toolkit
X2B	X-ray to Bone
X2B2	Biplanar X-ray to Bone
X2BR	X-ray to Bone with Non-rigid Registration Method
X2V	X-ray to Volume

1. 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.

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 [3] and learning deformations from average organ shapes.

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 pre-operative 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.

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 3Dr 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 (Digitally 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 thesis 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. 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.

The X2V network 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 second part of this thesis introduces the X2B network for bone volume extraction, which extends the X2V framework by incorporating ConvNeXt as an image encoder. The network is trained using randomly selected DRRs from multiple views to reconstruct detailed skeletal structures, enabling applications in biomedical simulations and augmented reality. A significant advancement in this work is the integration of implicit and explicit modeling approaches, referred to as Hybrid Coordinate Networks. Specifically, the X2BR Hybrid Coordinate Network combines the occupancy network framework for coarse shape estimation with the Geodesic Bayesian Coherent Point Drift (GBCPD) algorithm [11] for non-rigid alignment of person-specific templates with predicted shapes. These templates are generated through a transfer learning network for vertebra and rib cage detection, pre-trained on the YOLOv9 [12] object detection model. Detected regions are extracted from a primary template based on the SKEL model [13], resulting in watertight meshes suitable for biomedical applications. The X2B2 network further enhances multiview image processing by integrating a ConvNeXt [14] backbone with a cross-attention mechanism to fuse features from bi-planar DRR inputs. Leveraging a pre-trained ConvNeXt for feature extraction, the network employs multi-head

cross-attention to align and combine complementary information across views, producing a cohesive latent representation. This architecture is particularly effective for tasks such as 3Dr, medical imaging, and multi-perspective scene analysis, combining advanced feature extraction with robust multiview data fusion.

1.1 Radiography

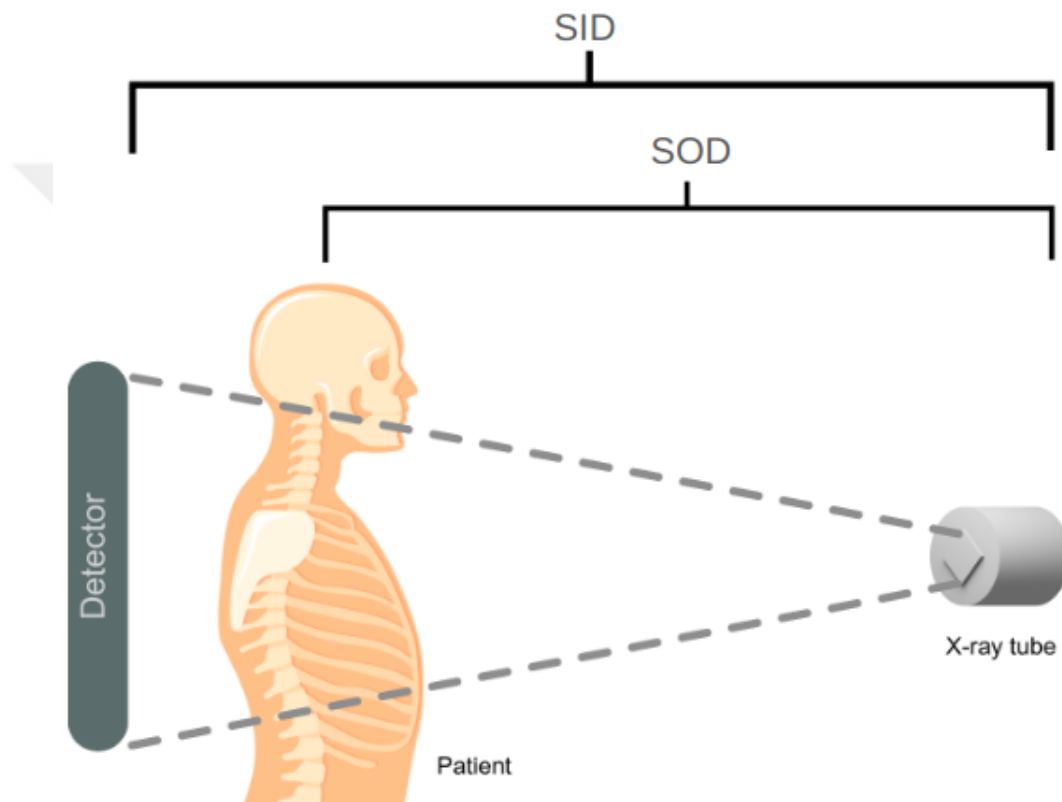
Radiography is the practice and science of using radiation to generate images of the tissues, organs, bones, and blood vessels within the human body. Radiologists, physicians specifically trained to interpret diagnostic images, analyze these visuals to make diagnoses, and effective patient treatment depends on the precise and accurate production of radiographic images. Both X-ray and CT imaging are essential diagnostic tools [15] that employ ionizing radiation to capture detailed visuals. In X-ray imaging, radiation passes through the body, with denser structures like bones absorbing more and appearing white on the images. As the earliest and most common form of medical imaging, X-rays are frequently utilized for identifying fractures, infections, arthritis, and foreign objects. In contrast, CT scans combine X-ray technology [16] and computer processing to generate cross-sectional, 3D images of organs, bones, blood vessels, and soft tissues, capturing data with a rotating X-ray tube. These detailed images are valuable for diagnosing cancers, cardiovascular diseases, and infections, as well as guiding biopsies. In both techniques, contrast agents like iodine or barium are used to improve the visibility of specific structures, aiding in precise diagnosis.

1.1.1 X-ray Imaging

X-rays, discovered [15] by Wilhelm Conrad Roentgen in 1895, are a form of invisible electromagnetic radiation with extremely short wavelengths, approximately 1/10,000th

the length of visible light waves. They share many properties with other forms of electromagnetic radiation but are distinguished by their ability to penetrate materials that typically absorb or reflect visible light, due to their short wavelength.

Figure 1.1: *Schematic representation of radiographic distances. The figure illustrates the key radiographic parameters, including the Source-to-Image Distance (SID) and the Source-to-Object Distance (SOD). SID represents the distance between the X-ray tube and the detector; while SOD denotes the distance between the X-ray tube and the patient.*



In conventional X-ray imaging, X-rays pass through the body and interact with a detection device, such as X-ray film or an image receptor, creating a two-dimensional projection of the internal tissues. This X-ray-produced “photograph” is referred to as a “radiograph”. The process is illustrated in Figure 1.1. Key spatial relationships in X-ray imaging include the Source-Image Distance (SID), which refers to the distance between the X-ray tube and the image receptor, and the Object-Image Distance (OID), which is the distance between the patient or a specific anatomical structure and the image receptor.

Another important parameter is the Source-Object Distance (SOD), which is the distance between the X-ray tube and the patient or the relevant anatomical structure.

The properties of X-ray images are typically categorized into those that define visibility and those that define accuracy. Visibility encompasses features such as brightness and contrast, while accuracy is determined by spatial resolution and distortion.

1.1.2 Computed Tomography

CT combines multiple X-ray projections taken at different angles around the body to construct high-resolution, cross-sectional images using advanced computational techniques [17] like filtered back projection (FBP) or iterative reconstruction (IR). Unlike conventional X-rays, CT scanners provide detailed volumetric representations of soft tissues, bones, and blood vessels with sub-millimeter precision, making them invaluable for diagnosing trauma, cancer, and vascular diseases, and for planning medical or surgical interventions. Modern multi-detector CT (MDCT) systems can capture multiple slices simultaneously, reducing scan time and improving image quality.

However, CT imaging involves significantly higher radiation [18] doses than standard X-rays due to the multiple exposures required to generate detailed images. While a chest X-ray has an effective dose of approximately 0.1 mSv, a chest CT scan ranges from 6 to 7 mSv—60 to 70 times higher. This increased radiation [19] exposure necessitates careful consideration, especially for vulnerable populations or patients requiring repeated scans. Techniques like low-dose CT protocols, iterative reconstruction, and automatic exposure control (AEC) help mitigate radiation risks. Despite these challenges, CT remains a cornerstone of diagnostic imaging due to its rapid acquisition, sensitivity to small lesions, and integration with contrast agents for enhanced visualization. Advances such as photon-counting detectors, deep learning reconstruction algorithms, and dual-energy CT promise to further reduce dose and enhance diagnostic capability in the future.

1.2 Neural Fields

Recent progress in machine learning has sparked considerable interest in addressing challenges in visual computing using coordinate-based neural networks, known as neural fields. These networks model the physical properties of scenes or objects across both spatial and temporal dimensions and have been effectively employed in numerous applications such as 3D shape synthesis, image synthesis, human body animation, and pose estimation. Neural fields facilitate intricate manipulation, storage, and transmission of visual information, extending their impact to fields like robotics, where they are used for planning based on 3D structures of scenes. Since 2019 [20], the adoption of these methods has grown due to their capacity for high-resolution, continuous parameterization across various dimensions, thus enhancing precision and efficiency in visual tasks. This has positioned neural fields as vital in diverse applications ranging from digital human modeling to data compression and image processing, indicating a shift towards their integration in various disciplines, including robotics and medical imaging.

In mathematical terms, a field is a quantifiable attribute at any point across space and/or time, represented through a function that maps a coordinate x to a quantity, either scalar or vectorial. Often, the method to generate these fields does not have a clear analytical expression [20], leading to the use of parameters Θ which may be crafted, optimized, or learned. These parameters define a field that outputs a quantity as shown by the equation:

$$q = \Phi(x; \Theta) \tag{1.1}$$

Neural networks, composed of numerous interconnected layers of artificial neurons [21], are designed to perform non-linear transformations from a consistent input to output. Especially, the multi-layer perceptron (MLP) is known for its ability to approximate any function, a capability underscored by [22] the universal approximation theorem.

Post-2010, there has been a surge in both interest and investment in neural network technology, driving significant improvements in hardware and software, and lowering barriers to the application of neural networks across various sectors.

The universal approximation theorem posits that a multi-layer perceptron (MLP) neural network can model any field [23], thereby defining a “neural field” as one either completely or partly represented by a neural network. Neural fields are notable for their continuity and adaptability, which allow them to use memory more effectively compared to traditional discrete parameterizations that often struggle to maintain performance at high spatio-temporal resolutions. Unlike standard continuous parameterizations, such as those using Fourier series, neural fields adjust their complexity according to the specific needs of the application, optimizing the use of network parameters. These fields are generally developed using MLPs equipped with precise gradients in their activation functions, enhancing their functionality in gradient descent optimization techniques. Such attributes, combined with their analytical differentiability and the propensity for over-parameterization, make neural fields uniquely suited for solving complex signal regression issues that are typically challenging. The term “neural field” intentionally excludes [24] neural network designs that generate spatial outputs, like voxel grids, which usually use convolutional layers to produce outputs such as RGB color grids or occupancy maps. In the realm of visual computing, “implicit neural representations,” “neural implicits,” and “coordinate-based neural networks” all refer to what is known as neural fields. In neuroscience, the term might refer to theories regarding the structure and function of the brain or to a structured array of neurons within a network.

In practical terms, the workflow for a neural fields algorithm in visual computing starts by sampling coordinates across space-time, which are input into a neural network. This network then produces field quantities relevant to the specific reconstruction area

targeted in the project. A forward mapping strategy is then employed to align these reconstructions with the sensor domain, such as an RGB image, facilitating the application of supervised learning techniques. The process concludes with the assessment of reconstruction errors or losses, which are crucial for refining the neural network's performance. This refinement is achieved by contrasting the reconstructed output from the neural network with the actual measurements from the sensor.

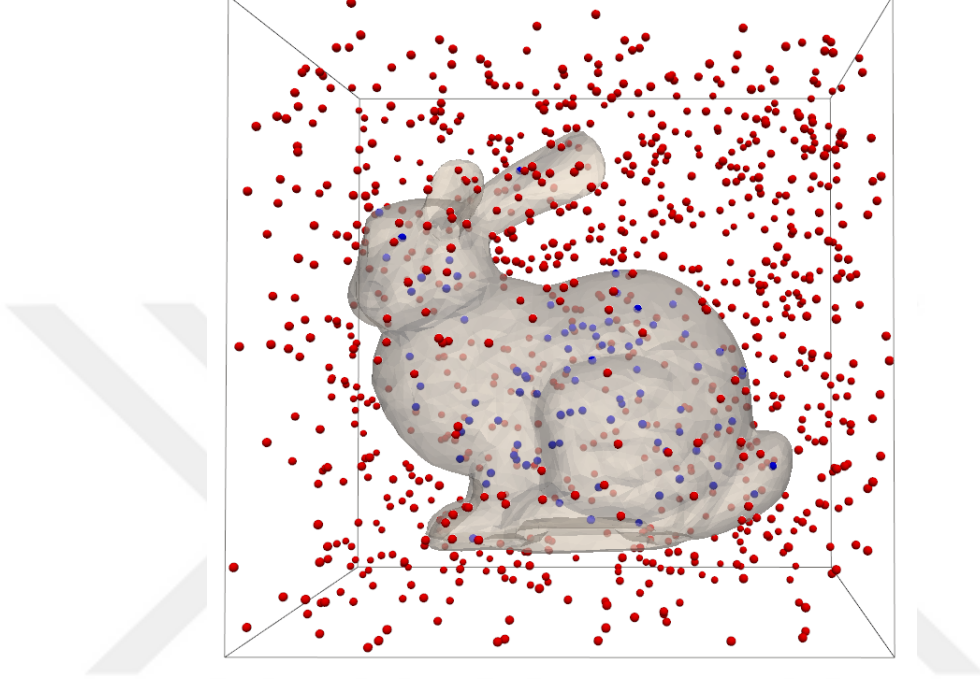
1.2.1 Occupancy Field

In the domain of 3D modeling and reconstruction, the concept of an occupancy field serves as a fundamental technique for determining the spatial occupancy of objects within a given volume. An occupancy field [10, 25] assigns a binary value to each point in a three-dimensional space, indicating whether the point is occupied by a part of an object (represented by 1) or not (represented by 0).

The visualization in Figure 1.2 illustrates an example of an occupancy field, where points inside the object (the bunny) are classified as occupied and represented in blue, while points outside the object are marked as unoccupied and represented in red. This binary classification highlights the spatial occupancy of the object within a bounding cube, forming the basis for 3D modeling and reconstruction.

Advances in neural networks, particularly through the application of coordinate-based neural networks known as neural fields, have facilitated the representation of these occupancy fields with high granularity. These networks learn to predict the occupancy status of any point in space by processing input coordinates and outputting occupancy probabilities. This approach allows for more precise and detailed 3D models from sparse or incomplete datasets, enhancing capabilities in areas such as virtual reality, where accurate and real-time environmental integration is crucial. The implementation of occupancy fields via neural networks not only improves the fidelity and resolution of 3D models but

Figure 1.2: Visualization of the occupancy field for the Stanford Bunny. The figure illustrates the spatial representation of the Stanford Bunny using an occupancy field, where red and blue points represent sampled locations in the 3D space. The occupancy field encodes whether each sampled point lies inside or outside the surface of the mesh, enabling the implicit modeling of the object’s geometry.



also introduces efficiencies in processing and storage, essential for real-time applications.

1.2.2 Signed Distance Fields

In the area of computational geometry and computer graphics, the Signed Distance Field (SDF) stands as a pivotal mathematical representation, integral to the sophisticated manipulation and rendering of complex geometric forms. An SDF assigns to each point in a three-dimensional space a scalar value representing [26] the minimum distance to a surface, where the sign of this value indicates the point’s position relative to the surface—negative inside the object, positive outside, and zero at the exact boundary. In three-dimensional geometry, the SDF [27] assigns a scalar signed distance value to each 3D coordinate $\mathbf{x}$, expressed as $f : \mathbb{R}^3 \rightarrow \mathbb{R}$. The surface $\mathcal{S}$ is defined as the set of points

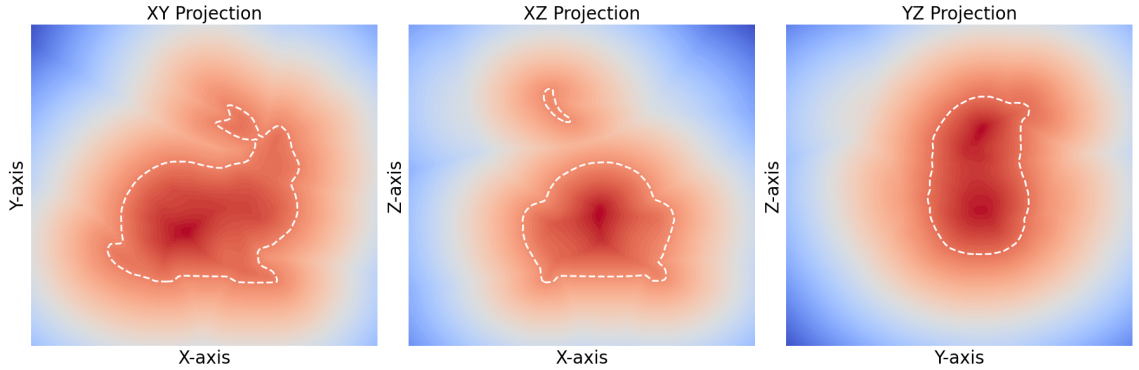
where the function equals zero, representing the zero level-set of the field.

$$\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^3 \mid f(\mathbf{x}) = 0\}. \quad (1.2)$$

This characteristic allows for a continuous, analytic description of surfaces and volumes, which is highly beneficial for numerous computational processes.

Figure 1.3 demonstrates the SDF of the Stanford bunny model, with projections onto the XY, XZ, and YZ planes. The color gradient illustrates the distance values, where red regions represent points closer to the surface, blue regions indicate points farther away, and the white dashed lines mark the zero-level set, corresponding to the exact boundary of the object.

Figure 1.3: *SDF projections of the Stanford Bunny. The figure shows the SDF of the Stanford Bunny, visualized as three orthogonal 2D projections: XY, XZ, and YZ planes. The SDF encodes the distance of each point in the field to the nearest surface of the mesh, with the sign indicating whether the point lies inside (negative) or outside (positive) the surface. The red regions represent areas closer to the mesh surface, while blue regions are farther away. The white dashed lines delineate the contours of the bunny’s mesh in each projection.*



SDFs are particularly esteemed for their utility in efficient rendering techniques such as ray marching, where the distance values facilitate skipping vast sections of empty

space, significantly enhancing computational efficiency. Additionally, the precise information about surface proximity and orientation provided by SDFs is instrumental in robust, real-time collision detection algorithms in physical simulations and interactive environments. Moreover, the continuous nature of the field supports complex shape transformations and Boolean operations, enabling dynamic modeling tasks that are foundational in virtual reality applications and digital sculpting.

1.3 Non-rigid Registration

Non-rigid registration is a fundamental technique in medical imaging, enabling the precise alignment of images or point sets that experience complex, elastic deformations. Unlike rigid registration, which only considers translations and rotations, non-rigid registration accommodates local deformations to achieve more accurate correspondence between anatomical structures. This is critical for applications such as monitoring disease progression, surgical planning, and integrating multi-modal imaging data. Advanced algorithms like Coherent Point Drift (CPD) [28] and Gaussian Mixture Models (GMMs) are frequently used. The Coherent Point Drift (CPD) algorithm is a probabilistic method for non-rigid point set registration that models one point set as centroids of a Gaussian Mixture Model (GMM) and fits it to another point set by maximizing a likelihood function, ensuring smooth motion via a motion coherence constraint. Bayesian Coherent Point Drift (BCPD) [29] extends CPD by incorporating Bayesian inference, replacing the motion coherence theory with a prior distribution over displacement vectors, which enhances robustness and provides a probabilistic framework for parameter estimation. Recent innovations include geodesic-based methods and neural network approaches like the Neural Deformation Pyramid [30], which enhance the accuracy and robustness of non-rigid registration by leveraging deep learning and geometric properties. Geodesic Bayesian Coherent Point Drift (GBCPD) [11] further advances BCPD method by integrating geodesic

distances, making it suitable for complex curved spaces or manifolds. The primary differences lie in the incorporation of Bayesian inference in BCPD and the use of geodesic distances in GBCPD, allowing for more flexible and accurate registrations in varied geometrical contexts.

These sophisticated techniques are essential for handling the variability and complexity of biological tissues, making non-rigid registration indispensable in advancing medical diagnostics and personalized treatment strategies.

1.4 Contributions

In this thesis, the development of a method to synthesize 3D bone and organ volumes from a single X-ray addresses a critical gap in medical imaging by providing a low-cost, accessible, and safe alternative to traditional 3D imaging techniques. Conventional methods, such as CT and MRI, while highly accurate, are often inaccessible in low-resource settings due to their expense, complexity, and infrastructure requirements. Additionally, CT scans expose patients to significant levels of radiation, raising concerns about their routine use, especially in vulnerable populations such as children. By contrast, the method proposed in this work leverages the ubiquity and affordability of X-ray imaging to reconstruct detailed 3D anatomical models, opening new avenues for advanced diagnostics and treatment planning.

The implications of this innovation extend across multiple domains. Clinically, the ability to generate accurate 3D models from a single X-ray offers transformative benefits in surgical planning, enabling precise interventions such as fracture repairs, joint replacements, and radiation therapy optimization. This capability is particularly impactful in emergency and pediatric care, where time-sensitive decisions and the need to minimize radiation exposure are paramount. Beyond direct patient care, the technology facilitates advancements in research and education, providing anatomically accurate 3D models for

biomechanical analysis, prosthetic design, and medical training. These contributions underscore the thesis’s potential to address critical challenges in healthcare by improving accessibility, reducing costs, and enhancing patient safety.

This work represents a significant step toward bridging the gap between the limitations of current imaging modalities and the growing demand for personalized, efficient, and equitable healthcare solutions. By enabling advanced 3D imaging through widely available X-ray technology, the outcomes of this research hold the promise of reshaping medical practice, fostering innovation, and advancing global health equity.

Addressing critical challenges in medical imaging, this work focuses on reconstructing precise, patient-specific 3D anatomical structures from traditional 2D imaging modalities, such as X-rays. Conventional methods, which predominantly rely on statistical 3D organ templates for 2D/3D registration, fail to adequately capture individual anatomical variability, leading to inaccuracies and suboptimal outcomes in clinical applications. Furthermore, reconstructing skeletal structures is particularly challenging due to overlapping Hounsfield Unit (HU) values in X-ray images, which obscure fine anatomical details, especially in complex regions like the spine. To overcome these limitations, novel methodologies are proposed in this work, leveraging neural implicit representations and deep learning to achieve accurate, high-resolution 3Drs, bridging the gap between traditional statistical approaches and modern computational techniques.

This thesis introduces X2V, a novel deep learning framework for reconstructing 3D organ volumes from single X-ray images. The key contributions of X2V 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 networks are capable of extracting the targeted volume when adapted to other organs, even when the default body has lost its topological properties due to bone fracture or tumor volume extraction.
- This study 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 3Dr accuracy, using metrics such as Chamfer-L1 distance, IoU, and F-score, when compared to recent existing approaches.

X2B extends the neural implicit modeling approach to skeletal structures, enabling the reconstruction of detailed 3D bone geometries from single X-ray images. The main contributions of X2B include:

- Introduces X2BR (X-ray to Bone Registration), which integrates object detection, neural implicit modeling, and non-rigid registration (GBCPD algorithm) to generate high-resolution rib and vertebra reconstructions.
- Proposes X2B2 (Biplanar X-ray to Bone), which incorporates a cross-attention mechanism for improved multi-view fusion, significantly enhancing 3D bone reconstructions from biplanar X-rays.
- Demonstrates superior reconstruction accuracy over traditional methods, achieving high-resolution, non-self-intersecting, watertight 3D bone meshes.

The thesis is structured into five chapters to address these challenges comprehensively. Chapter I (Introduction) provides an overview of the key concepts, including radiography (X-ray imaging and CT), neural fields (occupancy fields and SDFs), and non-rigid registration, while also outlining the primary contributions of this research. Chapter II (Previous Work) reviews existing approaches in 2D/3Dr, neural implicit techniques, registration-based methods, and deformation learning, highlighting their limitations and motivating the need for innovative solutions. Chapter III (3D Organ Reconstruction from Single X-rays) introduces the X2V network architecture, detailing its training, inference process, dataset preparation, experiments, and results, including segmentation and the calculation of occupancy values. Chapter IV (3D Bone Reconstruction from Single X-rays) presents the X2B network, describing its architecture, dataset preparation, experiments, and results, with a particular focus on reconstructing skeletal structures such as ribs, vertebrae, and the sternum. Chapter V (Conclusions) summarizes the findings, contributions, and potential directions for future research. This structure ensures a logical and detailed exploration of the novel contributions to medical imaging presented in this work.

2. PREVIOUS WORK

In this thesis, the literature pertaining to four pivotal domains within the realm of image analysis is explored: single image reconstruction networks, single view reconstruction with neural implicit methods, neural implicit methods in medical imaging, and 3D reconstruction (3Dr) from single planar X-rays. Each of these thematic areas presents unique methodologies and challenges in the context of 3D shape reconstruction from X-ray images, providing a comprehensive framework to understand and innovate upon existing paradigms in the field.

Addressing the challenges of substantial radiation exposure and computational complexity in extracting 3D topology and volumetric data of human anatomy from CT, a wealth of research has been conducted in recent years. Investigations have spanned various anatomical structures, including skeletal elements (e.g., knee, femur, spine, tibia, fibula, pelvis, lower extremity, and knee joint) [31, 32] and soft organs (e.g., lung, liver, stomach, duodenum, kidneys, and pancreas) [5, 6, 7, 8, 9], each with distinct X-ray transmittance or HU values. These structures are subject to unique topological deformations due to various diseases and temporal phases, necessitating diverse solution architectures to address phenomena like soft organ deformations and bone fractures. Single image reconstruction methods are favored as obtaining two or more corresponding X-ray images for 3D reconstruction is often clinically impractical.

Neural implicit representations are making a substantial impact in computer vision and graphics, utilizing deep learning to intricately model geometric shapes and surfaces without explicit parameterization. These methods encode objects and scenes as continuous functions, enabling tasks such as 3Dr, novel view synthesis, and shape interpolation with great flexibility and accuracy. Implicit representations, unlike explicit (or discrete) representations that directly encode features or signal values, are characterized by a generator function that maps input coordinates to their respective values within the input space.

Their ability to capture high-quality details renders them invaluable in virtual reality, animation, medical imaging, and robotics.

Single image 3Dr models powered by deep learning have significantly changed how three-dimensional insights are derived from two-dimensional images, marking a breakthrough in computer vision. These models offer detailed 3D perspectives from a single image, crucial for fields like augmented reality, robotics, and autonomous driving that require swift and precise spatial understanding. Advances in this technology enhance interaction within virtual environments and improve real-world navigation and object manipulation.

In medical imaging, neural implicit methods are increasingly important, offering advanced capabilities for depicting intricate anatomical structures with high-resolution, continuous volumetric representations. These techniques enable accurate 3Drs and segmentations from medical scans, essential for precise diagnostics, surgical preparation, and tailored treatments. The integration of neural implicit techniques in medical imaging promises improved visualizations, enhanced predictive analysis, and optimized patient care, playing a critical role in advancing medical technology.

2.1 Single Image 2D/3D Reconstruction Methods

In recent times, the field of 3Dr 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[33].

Bednarik *et al.* [34] and Patch-Net [35] 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 [36] and IsMo-GAN [37], 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) [38] and VANet [39] 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.* [40] 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.

2.2 Single-View Reconstruction Using Implicit Functions

Neural implicit representation primarily uses a MultiLayer Perceptron (MLP) to determine whether a point’s coordinates lie inside or outside a boundary [10, 41], or to compute the signed distance from that point. In deep learning frameworks focused on single-view implicit 3Ds, the two most commonly used data types are occupancy probability and signed distance functions. Occupancy networks [10] utilize a novel, implicit, coordinate-based learning approach, integrating a feature vector with a 3D point via a network of four ResNet blocks. This setup enables highly precise classification of points and detailed reconstructions of complex shapes, representing a significant leap forward in the field. Furthermore, the IM-NET model [41] advances 3D shape modeling by incorporating implicit fields within a deep learning framework. It features a 6-layer MLP decoder

that determines binary occupancy based on feature vectors and 3D coordinates, facilitating precise modeling from sparse data. These models are capable of reconstructing only the basic shape and are unable to capture the finer details of the object. Additionally, in high-resolution digital human modeling, PIFu,[42] merges pixel-level details from 2D images with implicit functions to create accurate 3D models of clothed humans. This method involves re-projecting 3D points into pixel-aligned feature spaces, enhancing the creation of complex models.

Several recent advances in CNN-based models with local and global feature vectors have been documented [43, 44, 45]. DISN [43] merges global and local features from the input image with point features from query points to produce a unified signed distance function. MDISN [44] adjusts a randomly generated signed distance function to better match the ground-truth (GT) signed distance function by responding to local feature changes at the layer level. Ray-ONet [45] combines global and local features with scaling parameters to more accurately estimate the occupancy probability of spatial query points along rays, enhancing efficiency and outperforming the previous Occupancy Network [10]. These newer models [43, 44, 45] have improved ability to capture structural details and refine shapes through the integration of local features, setting them apart from previous models [10, 41, 42]. However, they continue to face challenges in accurately reconstructing critical surface details, such as depth, essential for visual perception. D²IM-Net [46] targets these surface details, often achieving notable results in surface features, yet struggles to replicate precise topological structures. ED²IF²-Net [47], based on the Pyramid Vision Transformer (ViT), segments 3D into topological and surface detail phases, enhancing model fidelity. Utilizing multi-scale feature extraction and three specialized decoders, it achieves superior accuracy and efficiency on the ShapeNet dataset compared to existing models. Unlike traditional methods that use points, G2IFu [48] leverages

graphs to capture broader spatial structures and enhances accuracy using a graph convolution network. The model also incorporates a prior boundary loss and combines global and local image features with a self-attention module to improve performance. LIST [49] effectively utilizes a combination of global and local image features to initially predict a coarse shape of the object, which serves as a foundation for more detailed and higher-resolution reconstruction. This method enhances accuracy in capturing the intricacies of self-occluded geometries and complex topological structures without the need for camera estimation or pixel alignment, offering reliable reconstructions from the viewpoint of the input image.

2.3 Neural Implicit Methods in Medical Imaging

Neural implicit methods have emerged as transformative approaches in medical imaging, enabling continuous and high-fidelity modeling of 3D structures from sparse or partial 2D data. This section explores several state-of-the-art techniques leveraging deep implicit representations to address challenges in dental and radiological imaging.

X2Teeth [50] consists of three interconnected subnets: ExtNet, SegNet, and ReconNet. ExtNet analyzes panoramic radiographs to extract detailed representations of teeth. Leveraging these deep representations, SegNet engages in pixel-wise classification and enhances the segmentation maps for precise localization of teeth. Following this, ReconNet retrieves tooth patches from these enhanced feature maps to carry out reconstructions of individual teeth. Oral-3Dv2 [51] is designed to implicitly map 2D coordinates to voxel density values in 3D space, thereby representing the 3D oral structure. It employs a multi-head model for simultaneous prediction of multiple voxel values from a 2D axial coordinate, and utilizes a dynamic sampling strategy to enhance the detail accuracy in the density distribution of the reconstruction. ToothInpaintor [52] is designed to reconstruct a complete tooth model, including detailed roots, from a partial 3D dental model and a 2D

panoramic image. Implicit representations is employed for both 3D and 2D inputs and develop a latent space that encapsulates the complete tooth shapes.

MedNeRF [53] presents a Deep Learning model that can generate CT projections from a minimal number of X-ray images, including just one. The model uses an architecture based on neural radiance fields that allows the model to continuously represent CT scans by distinguishing the shape and depth of both the surface and internal anatomical features from 2D images. ImplicitVol [54] refines 3D positions from 2D scans and reconstructs 3D volumes using a deep implicit representation. This method reconstructs volumes from 2D ultrasound images without sensors, modeling the volume continuously rather than using traditional voxel grids. Structure-Aware X-ray Neural Radiodensity Fields [55] facilitates 3Dr from sparse-view X-rays using two main components. The first is the Line Segment-based Transformer (Lineformer), which models internal structures in 3D space by focusing on line segment dependencies in X-rays. The second component is the Masked Local-Global ray sampling strategy, designed to capture contextual and geometric information from 2D projections. SNAF [56] is a technique designed for reconstructing Cone Beam CT (CBCT) from sparse-view data. It utilizes neural attenuation fields and features an innovative view augmentation strategy to effectively manage the data limitations inherent in sparse input views.

2.4 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 [57], 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 [57].

Humbert *et al.* [58] and Chaibi *et al.* [59] presented parametric models to represent spine and femur bones from biplanar X-rays. Cresson *et al.* [60] 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. Liang *et al.* [61] 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.

2.5 Deformation Learning Based Approaches

As a foundational study in 3D reconstruction, P2M [38] has established a robust framework for generating 3D structures from 2D images in computer vision by deforming an ellipsoid template using latent image features. This work set a strong precedent for deformation learning methodologies and inspired subsequent advancements in deep learning-based 2D/3D deformable registration. Building upon this foundation, Wang *et al.*'s P2M [38] further demonstrated the capability to generate 3D meshes from 2D images through similar ellipsoid deformation techniques, solidifying its role as a pivotal approach in 3D reconstruction.

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 and Nakao *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 thesis 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.

2.6 3D Bone Reconstruction from X-rays

Recent advancements in 3D bone reconstruction from X-ray have harnessed the power of Statistical Shape Models (SSMs) and deep learning, enabling precise and efficient reconstruction of anatomical structures from limited 2D data. These methods integrate data-driven modeling with neural networks to address challenges in reconstructing complex bone geometries.

Statistical Shape Models (SSMs) stand out as a significant example of data-driven strategies, focusing on capturing both the mean shape and the diverse variations of anatomies. This is achieved through the development of atlases, which are compiled from extensive databases typically comprising healthy anatomical samples [62, 63]. Aubert et al [64], presents a new, fast, and automated approach to 3D spinal reconstruction that utilizes a statistical shape model aligned to images through convolutional neural networks (CNNs). The CNNs systematically and iteratively detect key anatomical landmarks on the spine, helping to shape and refine the spinal model. Jiang et al [65] introduces a method for 2D/3D registration that reconstructs the 3D geometry of a patient's spine using frontal-view X-ray images. X23D [66] introduces a new deep learning method for estimating the 3D shape of lumbar vertebrae during surgery using limited multi-view X-ray data. The method utilizes a sophisticated multi-view stereo machine learning approach that integrates X-ray calibration parameters directly into the neural network. BX2S-Net [67], a convolutional neural network for 3D spinal reconstruction from bi-planar X-ray images combines an encoder-decoder architecture with dimensionality enhancement and a Full-Scale Feature Attention Guidance module for improved semantic alignment and feature integration. Yang et al [68] adapted the earlier X2CT-GAN [69] model to create a new 2Dto3D-GAN for spinal reconstruction from bi-planar X-ray images.

The article [70] introduces a novel technique that uses deep learning and transfer learning to transform a single X-ray image into a precise 3D model of the femur. This method enhances the accuracy of femur shape reconstruction by leveraging specific deformation parameters from a Statistical Shape Model (SSM). It focuses on selecting an optimal pose from the X-ray that best captures the upper part of the femur, thereby maximizing the network's ability to deduce the 3D structure from the 2D image. The FracReconNet [71] is a novel deep learning framework for reconstructing 3D proximal femoral fractures from biplanar radiographs to enhance preoperative planning. The framework includes the 3Dr Network (3DReconNet), which uses a convolutional network architecture for detailed bone imaging, an Auxiliary Class (AC) to improve fracture representation, and Fractural Augmentation (FA) to expand the training dataset and enhance reconstruction accuracy.

2.7 Non-Rigid Registration Methods

Point set registration is essential for various computer vision applications, including robot navigation, image-guided surgery, motion tracking, and face recognition, supporting tasks such as object alignment, stereo matching, and image segmentation. The challenge involves identifying meaningful correspondences between point sets and determining the transformation mapping one set to the other. Points often represent features like interest points from images, line segments, and curves. Registration techniques are either rigid, preserving distances, or non-rigid, allowing flexible transformations like affine, which accommodates scaling and skewing. Advanced non-linear, non-rigid registration methods are needed for many applications, posing significant challenges.

The non-rigid point set registration problem is defined as an optimization problem [11] where a similarity transformation f_{sim} and a deformation matrix $V \in \mathbb{R}^{M \times D}$ are sought to correspond to a non-rigid deformation. This is expressed as follows:

$$(\hat{V}, \hat{f}_{\text{sim}}) = \arg \min_{(V, f_{\text{sim}})} \mathcal{L}(V, f_{\text{sim}}, X, Y), \quad (2.1)$$

where $\mathcal{L}$ is a loss function evaluating the dissimilarity between $X \in \mathbb{R}^{N \times D}$ and $T = f_{\text{sim}}(Y + V) \in \mathbb{R}^{M \times D}$. The target shape $X = (x_1, \dots, x_N)^T \in \mathbb{R}^{N \times D}$ is defined as a set of points in an N -dimensional space. The source shape $Y = (y_1, \dots, y_M)^T \in \mathbb{R}^{M \times D}$ is defined as a set of points in an M -dimensional space. The matrix $V = (v_1, \dots, v_M)^T \in \mathbb{R}^{M \times D}$ collects the deformation vectors of a source shape. D is the dimensionality of the space wherein a point set is embedded and N, M are the numbers of points in the target and source shapes, respectively.

The Coherent Point Drift (CPD) algorithm addresses the challenges of non-rigid point set registration by employing a probabilistic approach, treating the registration as a Maximum Likelihood (ML) estimation problem with a motion coherence constraint. This method models one point set (template) as centroids of a Gaussian Mixture Model (GMM) and iteratively fits it to the data points of another set (reference), ensuring smooth motion by penalizing rapid changes in velocity. The approach involves fitting a GMM to the first point set, with Gaussian centroids initialized from points in the second set. Unlike previous methods that assume a thin-plate spline transformation [72], no explicit assumption about the transformation model is made. Instead, the adaptation of Gaussian centroids from their initial to final positions is considered a temporal motion process, with a motion coherence constraint imposed over the velocity field. This constraint, based on Motion Coherence Theory [73], ensures that nearby points move coherently by penalizing derivatives of all orders in the velocity field.

The registration process involves maximizing the posterior probability, formalized by minimizing the energy function where α controls the Gaussian spread and λ regulates the smoothness. A Gaussian mixture density $p(x)$ can be defined where $x|m$ follows a normal distribution $\mathcal{N}(y_m, \sigma^2 I_D)$. The centroids $\mathbf{Y}$ are D -dimensional points, and $\mathbf{X}$

denotes the data points. To ensure smooth motion, the prior $p(\mathbf{Y}|\lambda)$ incorporates a regularization function $\phi(\mathbf{Y})$. By applying Bayes' theorem, the optimal parameters $\mathbf{Y}$ are found by minimizing the following energy function:

$$E(Y) = - \sum_{n=1}^N \log \sum_{m=1}^M \exp \left(-\frac{\|x_n - y_m\|^2}{2\sigma^2} \right) + \frac{\lambda}{2} \phi(Y) \quad (2.2)$$

The robustness of CPD algorithm stems from using a soft-matching approach for shape correspondence, avoiding the assumption of one-to-one point correspondences. Furthermore, the algorithm's sequential update of residual variance in the GMM narrows the candidate points for each source point during optimization. In the Bayesian framework of CPD (BCPD) [29], Bayesian inference replaces the Motion Coherence Theory, introducing motion coherence through a prior distribution of displacement vectors. The BCPD algorithm ensures convergence and offers intuitive motion coherence parameters. It combines rigid and non-rigid CPD variants, and its performance can be accelerated using non-Gaussian kernel functions.

Geodesic Bayesian Coherent Point Drift (GBCPD) is an advanced method for non-rigid point set registration that combines Bayesian inference with geodesic distances to handle complex curved surfaces or manifolds. Building on the Bayesian Coherent Point Drift (BCPD) framework, which introduces a probabilistic approach with a prior distribution over displacement vectors to ensure smooth transformations, GBCPD further incorporates geodesic distances to accurately measure and align points in non-Euclidean spaces. This integration allows GBCPD to handle point sets lying on curved surfaces, such as anatomical structures in medical imaging or 3D models in computer graphics, more effectively. By leveraging the shortest path along the manifold's surface rather than straight-line Euclidean distances, GBCPD provides a robust and flexible solution for point set registration in complex geometrical contexts, ensuring accurate and coherent alignments even in the presence of significant non-linear deformations.

3. 3D ORGAN RECONSTRUCTION FROM A SINGLE X-RAY

This chapter introduces a novel approach for three-dimensional (3D) organ volume reconstruction from a single planar X-ray, referred to as the X2V network. This method has significant clinical relevance, particularly in applications such as real-time image-guided radiotherapy, computer-aided surgery, and patient monitoring during follow-up sessions. Conventional techniques for reconstructing 3D volumes from X-rays typically rely on statistical 3D organ templates for 2D/3D registration. However, these approaches often fail to accurately capture inter-patient anatomical variability, leading to suboptimal results.

To address this limitation, the X2V model utilizes neural implicit representations, incorporating a ViT as the encoder network. This encoder is specifically designed to focus attention on key regions within the X-ray image, enabling enhanced feature extraction and representation. The reconstructed 3D meshes closely resemble the GT organ volumes, demonstrating the model’s capability to effectively infer 3D structure from 2D inputs.

The X2V network is capable of reconstructing lung volumes from a single X-ray image and can be extended to other organs as well. However, the development and evaluation of the network have primarily focused on lung data due to the abundance of publicly available datasets and the unique characteristics of lung anatomy. The HU values of the lungs are significantly low due to their air-filled structure, creating a strong contrast in X-ray images. This contrast enhances the visibility of lung boundaries, making them more distinct compared to other organs. These factors collectively contribute to the suitability of lungs as a primary focus for this research while preserving the potential for broader applicability to other anatomical structures.

The performance of X2V is rigorously evaluated on lung X-rays using metrics

such as volumetric Intersection over Union (IoU). The model achieves a significant improvement over the current state-of-the-art (DeepOrganNet), with IoU scores ranging from 0.892 to 0.942, compared to DeepOrganNet’s 0.815 to 0.888, representing a performance gain of approximately 7–9%. These results highlight the potential of X2V as a transformative solution for accurate and efficient 3D organ volume reconstruction in clinical practice.

3.1 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 the other hand, The ViT model has revolutionized the field of computer vision by applying transformers, traditionally used in natural language processing, to image classification tasks [74]. The X2V network applies a novel ViT based neural implicit model, conditioned on the input X-ray image, for 3D volume reconstruction.

The X2V model architecture is shown in Figure 3.1. 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 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 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.

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 3Dr. 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.

Figure 3.1: The X2V 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 (CBN), 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 3Dr from diverse inputs.

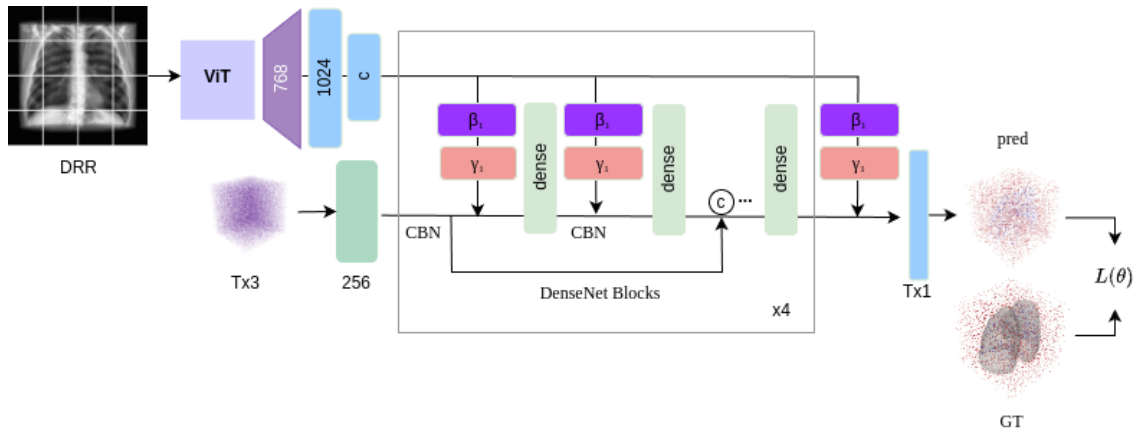
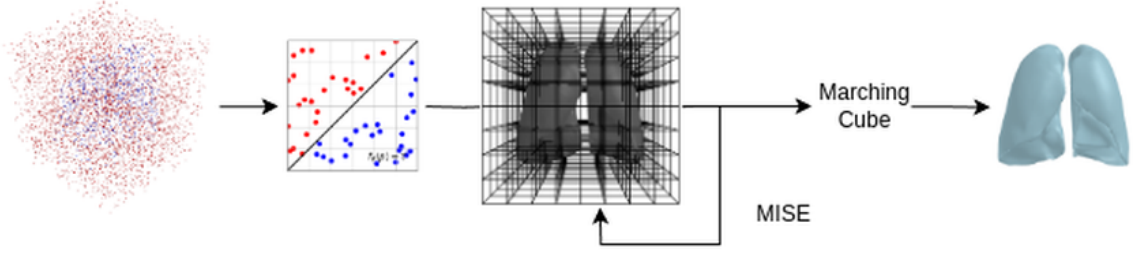


Figure 3.2: For 3D mesh inference with X2V model, a modified Multiresolution Iso-Surface Extraction (MISE) algorithm for high-resolution mesh extraction is integrated, starting with a base resolution and evaluating against the occupancy network. The occupancy threshold is set at $\tau = 0.2$ for balance 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.



The ViT (specifically, ViT-B/16) model [75, 76] 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 code-vector 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 [77] blocks that first apply conditional batch normalization (CBN) enabling dynamic adjustment of normalization parameters based on the conditional codevector c (See Figure 3.1). CBN [78] 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, 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} :

$$\text{CBN}_i(c, f_{\text{in}}) = \gamma_i(c) \left(\frac{f_{\text{in}} - \mu}{\sqrt{\sigma^2 + \epsilon}} \right) + \beta_i(c) \quad (3.1)$$

The densenet blocks integrate CBN 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 3.1, 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/3Dr performance of the proposed X2V model.

3.1.1 Training

Ideally, the 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 3Dr 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:

$$L_B(\theta) = \frac{1}{|B|} \sum_{i=1}^{|B|} L(f_\theta(p_i, x_i), o_i) \quad (3.2)$$

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 [79] 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 $\epsilon = 10^{-8}$.

3.1.2 Inference

Figure 3.2 depicts the inference of 3D 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 the network. The 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 [80], 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.

3.2 Dataset

The development of a high-quality dataset is fundamental to the success of this study, which aims to achieve accurate 3D organ reconstruction from X-ray images. In the absence of publicly available datasets containing paired X-ray images and their corresponding 3D organ meshes, we curated a custom dataset using advanced data acquisition, segmentation, and preprocessing techniques. By utilizing publicly available medical imaging repositories and state-of-the-art algorithms, the dataset was meticulously prepared to ensure its suitability for training and validating the proposed reconstruction framework. This section provides a comprehensive overview of the dataset creation pipeline, including lung segmentation using cutting-edge neural networks, the synthesis of DRRs, the calculation of occupancy values, and the pairing of X-ray images with their corresponding segmented 3D organ volumes.

3.2.1 Segmentation with Lung Analyzer

The 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) [81], 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) [82].
- Lung segmentation from CT images as watertight manifold meshes via U-net (R231) network [81].
- 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 [83, 84] is downloaded from TCIA [82] using NBIA Data Retriever Tool. A volumetric analysis and visualization in 3D Slicer is performed [85, 86, 87, 88] via the Lung CT Analyzer project with U-net(R231) network architecture [81, 89] on NLST dataset leveraging SimpleITK library. For the purpose of training and validating our proposed mesh reconstruction approach, an ideal dataset would consist of a large collection of paired X-rays and corresponding mesh reconstructions. However, such datasets are not readily available and the process of collecting them can be expensive. To overcome this challenge, the DRR technology [90] is utilized to synthesize the corresponding X-rays using a real CT volume, as depicted in Figure 3.3

In our study, approximately 2560 subjects from the publicly available National Lung Screening Trial (NLST) dataset are selected [91] that comprises the low-dose CT scans from 26,254 of the subjects in TCIA [82].

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

Figure 3.3: a) CT; b) DRR; c) triangular mesh representations of a patient's thorax region.



step involves removing low-quality segmentations following the 3D organ mesh creation phase. In the segment of the NLST dataset [91] 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.

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 ± 0.03 on the test dataset [82]. 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 [85, 86, 87, 88]. Essentially, the code resamples the DICOM to a $1.5\text{mm} \times 1.5\text{mm} \times 1.5\text{mm}$ 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.

3.2.1.1 Pairing Xray and DRR and their corresponding Segmented Volumes

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 $1.0\text{mm} \times 1.0\text{mm} \times 1.0\text{mm}$. 512×512 DRR images with 1.0×1.0 spacing are generated using a Siddon Ray Tracing algorithm on a GPU [90, 92]. In this study, Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied to DRR images to enhance the contrast and resolution of local details.

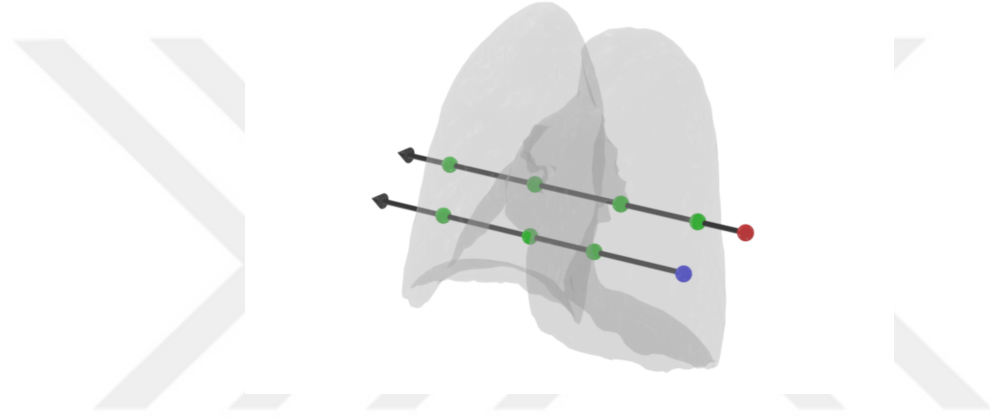
3.2.2 Calculation of Occupancy Values

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 3Drs [10].

Employing the voxelization method from Choy *et al.* [93], 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 3.4). Both the 100k point coordinates and their occupancies are saved, and a 2048-point subset is randomly selected during the training phase.

Figure 3.4: *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 labeled as internal, shown by a blue dot.*



3.3 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 3Dr performance of X2V is compared against two recent 2D/3Dr models from the literature, namely, Point Set Generation Network (PSGN) [94] and P2M Network [38].

3.3.1 Comparative Studies

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

3.3.1.1 Point Set Generation Network

The Point Set Generation Network (PSGN) [94], a 3Dr 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 [94], normalization of the 3D GT 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.

3.3.1.2 Pixel2Mesh

P2M [38] 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 [38] 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 GT, 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 [38]. 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.

3.3.2 *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 GT model. This metric is inherently symmetric and normalized, facilitating a comprehensive and robust comparison of 3D accuracy. To calculate the Chamfer L_1 distance, for each vertex p in the predicted mesh (PR), $p \in \partial M_{PR}$, the nearest vertex $q \in \partial M_{GT}$ within the 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.

$$\begin{aligned} \text{C-L}_1(M_{\text{PR}}, M_{\text{GT}}) \equiv & \frac{1}{2|\partial M_{\text{PR}}|} \int_{\partial M_{\text{PR}}} \min_{q \in \partial M_{\text{GT}}} \|p - q\| dp + \\ & \frac{1}{2|\partial M_{\text{GT}}|} \int_{\partial M_{\text{GT}}} \min_{p \in \partial M_{\text{PR}}} \|p - q\| dq. \end{aligned} \quad (3.3)$$

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 GT meshes, respectively. This metric ranges from 0 (no overlap) to 1 (perfect overlap), providing a standardized measure of spatial congruence between two 3D models.

$$\text{IoU}(P_{\text{PR}}, P_{\text{GT}}) \equiv \frac{|P_{\text{PR}} \cap P_{\text{GT}}|}{|P_{\text{PR}} \cup P_{\text{GT}}|}. \quad (3.4)$$

A higher IoU indicates greater overlap and similarity between the meshes, useful for evaluating accuracy in 3Drs. 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 [95].

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:

$$F\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall} + \epsilon} \quad (3.5)$$

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

$$\text{Precision} = \frac{\sum_{i=1}^{N_{\text{PR}}} \mathbb{I}(d_{\text{PR}}^i < t)}{N_{\text{PR}}} \quad (3.6)$$

$$\text{Recall} = \frac{\sum_{i=1}^{N_{\text{GT}}} \mathbb{I}(d_{\text{GT}}^i < t)}{N_{\text{GT}}} \quad (3.7)$$

In these equations, N_{PR} and N_{GT} denote the number of points in the predicted and GT meshes, respectively. The function $\mathbb{I}(\cdot)$ 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.

3.4 Simulation Results

The simulation results presented in this section comprehensively evaluate the performance of the proposed X2V network against state-of-the-art methods, Point-to-Mesh (P2M) and Point Set Generation Network (PSGN), in single-view 3Dr tasks. To ensure a robust assessment, both qualitative and quantitative evaluations are conducted, providing a holistic perspective on the strengths and limitations of each approach. The results are analyzed using key performance metrics such as IoU, Chamfer-L1 distance, F-score, and normal consistency (NC). These metrics are calculated on a diverse test set to compare the structural accuracy, geometric fidelity, and robustness of the reconstructed meshes. Additionally, visual comparisons are presented to highlight the differences in mesh quality, structural connectivity, and adherence to GT anatomical features. The following subsections detail the experimental findings, supported by illustrative examples and statistical analyses.

3.4.1 *Simulation Results and Discussions*

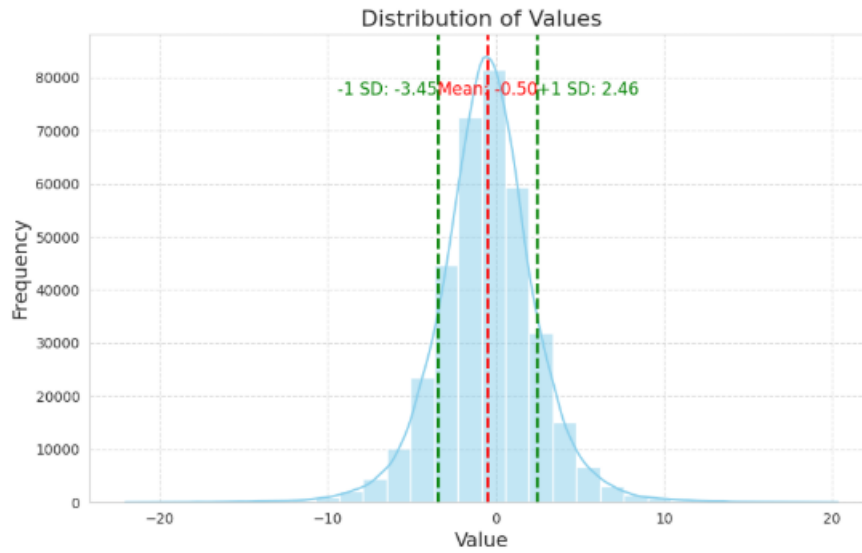
This section compares the performance three tested methods, X2V, PSGN and P2M, first visually and then using the objective evaluation metrics.

The GT and predicted meshes, along with their corresponding distance heatmaps, are presented in Figure 3.5. The X2V model outputs are displayed sequentially from left to right, starting with the DRRs. These are followed by the anterior views of the GT and the predicted 3D structures. Adjacent to these are the Euclidean distance maps, which quantify the spatial discrepancies between the GT and predicted models. Finally, the posterior views of both the GT and predicted structures are shown to provide a comprehensive visual comparison of reconstruction accuracy. The heatmaps, shown in Figure 3.5, specifically illustrate the point-to-point differences between the GT and X2V predictions for the front and back sides of the lungs, with deviations visualized using a color bar ranging from

Figure 3.5: Real-size lung volume analysis of GT and X2V network output: (a) Single anteroposterior DRR with CLAHE of lungs as input to networks; (b) GT 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) GT 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 -0.33 mm and std 2.95 mm).



Figure 3.6: Distribution of point-to-point distances for 639 test images. The mean distance is -0.5 mm, with a standard deviation of 2.95 mm. The histogram provides a visual representation of the frequency of distances, with the red dashed line indicating the mean and green dashed lines showing ± 1 standard deviation from the mean.

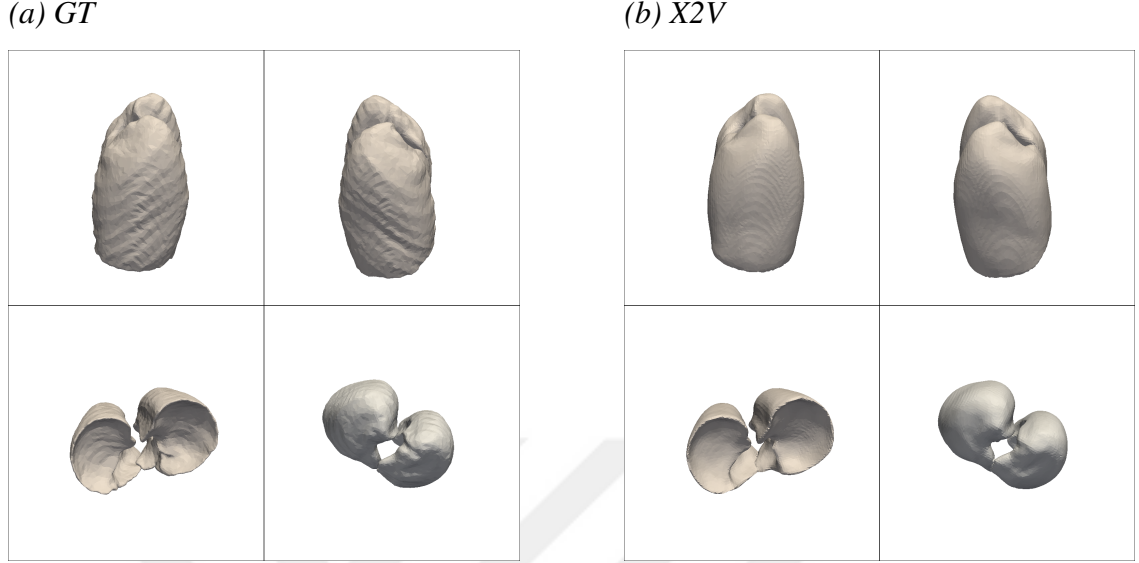


-22.5 mm to +22.5 mm.

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 3.5. Close examination of X2V-derived meshes reveals that corners of the lung meshes and detailed topological features align closely with the GT. 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.

As illustrated in Figure 3.6, the X2V model achieves high prediction accuracy, with a mean point-to-point error of -0.5 mm and standard deviation of 2.95 mm across 639 test images. The histogram in Figure 3.6 provides a statistical overview of these discrepancies, highlighting the overall consistency of X2V predictions relative to the GT meshes. Notably, the X2V model demonstrates the ability to predict left and right lungs with precise 3D spatial alignment, achieving a real-size mesh accuracy of 0.5 mm. Close

Figure 3.7: Comparison of lung CT reconstructions. (a) GT lung mesh. (b) Lung mesh generated by the X2V network.



examination of the X2V-derived meshes reveals that corners and detailed topological features align closely with the GT, with no significant distortions. Together, Figures 3.5 and 3.6 emphasize the spatial and statistical reliability of the X2V model in accurately reconstructing lung structures.

Building on the visual comparison provided in Figure 3.5, Figure 3.7 further highlights the structural differences between the GT and the X2V predicted meshes. In Figure 3.7, the GT 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.

Evaluating the effectiveness of different reconstruction methods is essential for accurately assessing their capability to replicate 3D geometries from 2D projections. Figure 3.8 shows that each tested method effectively replicates the 3D geometry from an antero-posterior DRR. Note that, the results depicted in Figure 3.8 are normalized within a unit cube. The PSGN method demonstrates the ability to produce relatively high-quality

Figure 3.8: Comparison for lung volume analysis. (a) Single antero-posterior DRR of lungs as input to networks. (b) GT 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.



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 3.8, which displays non-manifold surface properties and holes throughout the surface of the meshes.

Tables 3.1 and 3.2 present quantitative comparison of our X2V approach against standard benchmarks in single image 3Dr, utilizing our test set. The values in Tables 3.1 and 3.2 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 GT meshes.

Table 3.1: *Comparison of X2V and P2M in terms of IoU, Chamfer-L1, F-score and NC metrics for mesh reconstruction accuracy.*

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

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.

Table 3.2: *Comparison of X2V and PSGN in terms of IoU, Chamfer-L1, F-score and NC metrics for mesh reconstruction accuracy.*

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

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 3.2 and 3.1 respectively, matching the resolutions of PSGN and P2M predictions.

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.

Table 3.3: *X2V performance evaluation at different mesh resolutions.*

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

In Table 3.3, 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.

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.

4. 3D BONE RECONSTRUCTION

FROM A SINGLE X-RAY

In the previous chapter, we explored the extraction of lung volume from DRR images using the X2V network structure. This chapter shifts focus to the extraction of ribs, costal cartilages, the sternum, and cervical, thoracic, and lumbar vertebrae from DRR images using an object detection, a neural implicit model and a non-rigid registration algorithm. The development of three distinct neural network models, named X2B, X2BR, and X2B2, will be presented and discussed in detail.

Reconstructing accurate three-dimensional (3D) bone structures from single planar X-rays presents a significant challenge in medical imaging, particularly for anatomically complex regions such as the ribs, vertebrae, and sternum. Conventional methods, which typically rely on statistical 3D templates or explicit registration techniques, often fail to accommodate patient-specific anatomical variations and intricate bone geometries. To overcome these limitations, this study proposes a neural implicit framework, termed X2B, which encodes continuous volumetric representations directly from X-ray input coordinates. The framework incorporates a ConvNext-based encoder, designed to extract hierarchical spatial features and enhance the model’s ability to focus on critical regions within the X-ray images, effectively mitigating challenges posed by overlapping HU values. Extensive experiments conducted on clinical datasets demonstrate that X2B significantly outperforms existing state-of-the-art methods, achieving notable improvements in key metrics such as volumetric IoU and Chamfer Distance. This framework enables precise, patient-specific 3D bone reconstructions, making it particularly suited for applications in surgical planning, biomechanical analysis, and personalized treatment strategies. By integrating the strengths of ConvNext and neural implicit representations, X2B provides a robust and innovative solution to persistent challenges in X-ray-based medical

imaging, advancing the accuracy and reliability of 3D bone reconstruction methods.

4.0.1 X2B and X2BR Model Architectures

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 ConvNeXt model has advanced the field of computer vision by employing convolutional neural network architectures optimized for modern hardware, achieving high performance on image classification tasks [14]. The network applies a novel ConvNeXt-based neural implicit model, conditioned on the input X-ray image, for 3D mesh reconstruction.

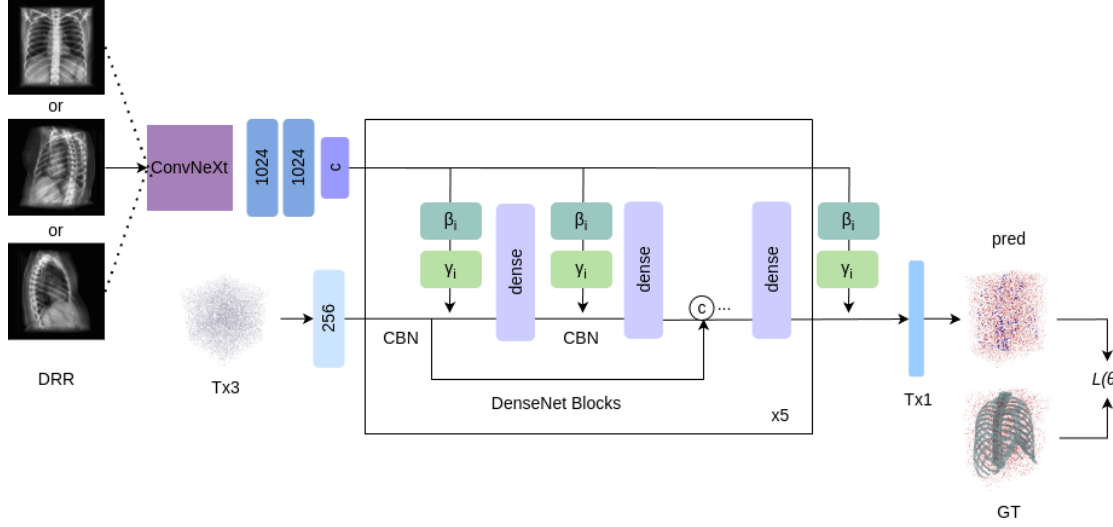
The X2B model architecture is shown in Figure 4.1. The network accepts as input a DRR image of size 224×224 and a set of 3D-coordinates of T random points in space. The network predicts the occupancy probability (i.e. the probability that a point in space is occupied) of each point at its output. To facilitate the reconstruction of a three-dimensional surface from a DRR image, the occupancy function is conditioned on the DRR image. This image is captured from 24 distinct angles during the training phase of the model, ensuring comprehensive multi-angle representation. The occupancy function [10], central to our methodology, implicitly represents the bone surface as a decision boundary. This approach generates non-self-intersecting closed surfaces without relying on input template meshes.

In the X2B network, a ConvNeXt model is employed as a feature extractor to condition the occupancy probability of 3D points in space on the input DRR images. ConvNeXt is a modern convolutional neural network (ConvNet) designed to compete with ViTs while retaining the simplicity and efficiency of standard ConvNets. ConvNeXt modernizes the ConvNet architecture by incorporating advanced training techniques such as the AdamW optimizer, Mixup, Cutmix, and Stochastic Depth, significantly improving

performance. Inspired by Swin Transformers, ConvNeXt adopts a multi-stage design in which the computation is distributed across multiple hierarchical stages, each processing increasingly abstract representations of the input data. The "stage compute ratios" are adjusted to allocate computational resources more effectively across stages, ensuring a balance between computational cost and feature extraction at different levels. Additionally, ConvNeXt employs a simplified "patchify stem," which replaces the complex overlapping convolutional operations typically used in traditional ConvNets with a 4x4 non-overlapping convolution. This patchify stem efficiently divides the input image into discrete patches or tokens while reducing computational overhead, providing a streamlined and computationally efficient mechanism for initializing the feature extraction pipeline. Key architectural modifications include the use of depthwise convolutions to reduce computational complexity, inverted bottlenecks to improve efficiency, and the exploration of large kernel sizes (up to 7x7) to capture larger receptive fields. Micro design changes involve replacing ReLU with GELU for smoother activation, reducing the number of activation functions and normalization layers per block, and using Layer Normalization instead of Batch Normalization for better performance and stability. Separate downsampling layers between stages stabilize training and improve performance. Empirical results show that ConvNeXt variants (T/S/B/L) achieve competitive performance with ViTs and Swin Transformers on ImageNet classification, object detection, and segmentation tasks (COCO, ADE20K). ConvNeXt combines the best of ConvNet and Transformer design principles, resulting in a simple yet powerful architecture that achieves state-of-the-art performance across various vision benchmarks, challenging the dominance of ViTs.

Within the X2B network, a ConvNeXt [14] model serves as the feature extractor, conditioning the occupancy probability of 3D points on the input DRR image. ConvNeXt is an advanced deep learning architecture designed to enhance convolutional neural network (CNN) performance by integrating features from transformer-based models while

Figure 4.1: X2B network training pipeline. The figure illustrates the training process of the X2B network, which takes a DRR as input and uses a ConvNeXt backbone for feature extraction. The extracted features are passed through dense blocks with Conditional Batch Normalization (CBN) layers, parameterized by β_i and γ_i , to refine the latent representations.



retaining the efficiency and simplicity of traditional CNNs. It improves upon conventional CNN designs by replacing batch normalization with layer normalization, adopting a streamlined block design, and utilizing GELU activations instead of ReLU. By merging the advantages of convolutional and transformer models, ConvNeXt provides a powerful and scalable solution for various computer vision tasks, including image classification, object detection, and feature extraction.

The ConvNeXt 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 3Dr. The ConvNeXt base architecture is initialized with pre-trained weights, and its classifier layer is removed. Following this, a fully connected layer (fc) is added to project the features to the desired embedding dimension which is 1024.

The occupancy network model takes the output of ConvNeXt (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 processed through four DenseNet blocks [77], which utilize Conditional Batch Normalization (CBN) [78] to dynamically adjust normalization parameters based on the conditional codevector c (see Figure 4.1). CBN modifies traditional batch normalization by incorporating external information via c , with linear layers $\gamma(c)$ and $\beta(c)$ generating scale and shift factors. This approach fine-tunes normalization based on contextual information, enhancing the model's ability to handle varied data efficiently.

Initially, the conditioning vector c from the encoder network, representing external information such as a 2D image, passes through two fully-connected layers, producing 256-dimensional vectors $\beta(c)$ and $\gamma(c)$. The 256-dimensional input feature vector f_{in} for each CBN layer is then normalized using its mean and variance, scaled by $\gamma(c)$, and shifted by $\beta(c)$. The normalization applied by the CBN layer to f_{in} is given by:

$$\text{CBN}_i(c, f_{\text{in}}) = \gamma_i(c) \left(\frac{f_{\text{in}} - \mu}{\sqrt{\sigma^2 + \epsilon}} \right) + \beta_i(c) \quad (4.1)$$

DenseNet blocks integrate CBN with convolutional layers and ReLU activation functions to transform the input tensor, initially normalizing and transforming it, followed by concatenation with the original input for further transformation to the desired output dimension.

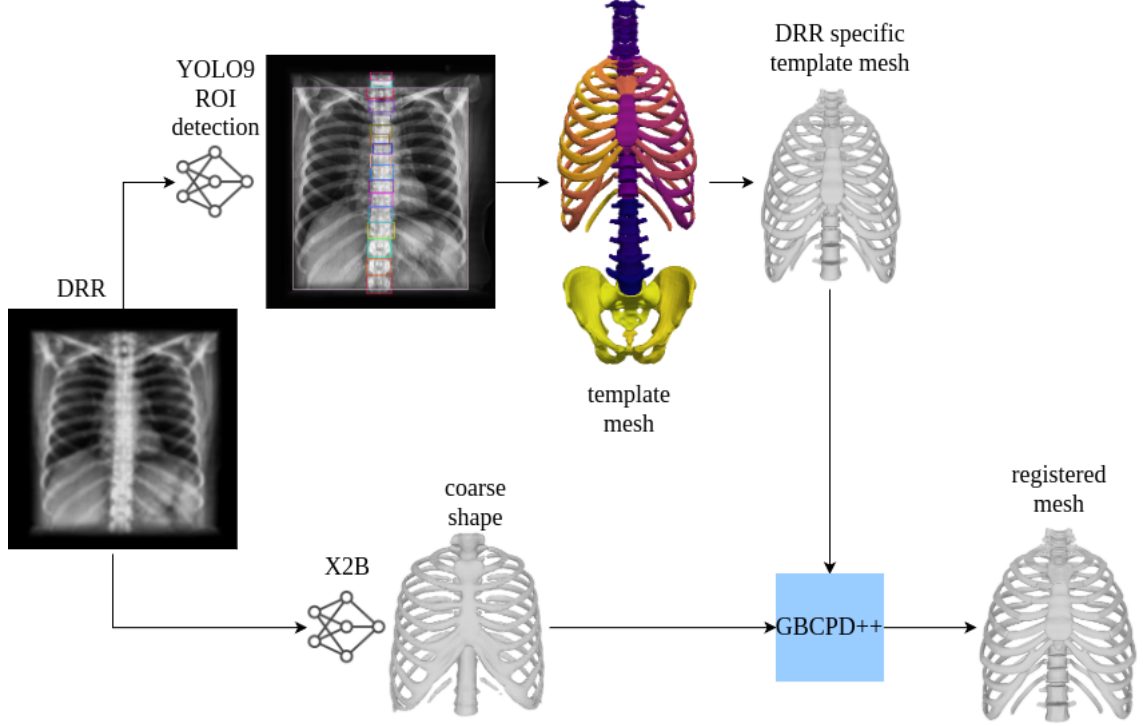
As illustrated in Figure 4.1, each DenseNet block employs skip connections to transfer features from shallow to deeper layers, facilitating better gradient flow during back-propagation. This dense connectivity ensures comprehensive feature integration, with each layer receiving inputs from all previous layers, enhancing feature fusion's depth

and subtlety. CBN enables dynamic adjustment of normalization parameters, tailored to the input DRR image’s specific attributes. The combined use of ConvNeXt-based encoding, CBN layers, and DenseNet architecture significantly enhances the 2D/3Dr performance of the proposed X2B model.

X-ray imaging introduces substantial challenges due to the overlapping HU values of different tissues, which hinder the extraction of high-resolution three-dimensional details from anatomically complex structures such as the spine. The X2B methodology focuses on obtaining a coarse skeletal representation from a single planar X-ray, while its advanced variant, X2BR, handles the difficulties in reconstructing detailed vertebral geometries. To address these challenges, a non-rigid registration approach has been employed, enabling the generation of an approximate yet detailed three-dimensional skeletal structure from DRRs. This method is specifically designed to manage complex deformations and facilitate precise alignment of anatomical points or features that cannot be resolved through rigid transformations. The limitations of X-ray diffraction in capturing topologically intricate structures, including vertebrae, ribs, and costal cartilages, are mitigated through the application of the Geodesic Bayesian Coherent Point Drift (GBCPD) algorithm [11]. This technique refines a coarse skeletal shape derived from a modified occupancy network and integrates it with a mesh template specifically adapted for DRRs. The DRR-specific template is constructed using a pretrained YOLOv9 model for the detection of vertebrae and rib-cage structures, in conjunction with the OpenSim skeletal framework. Finally, the output of the YOLOv9 model is combined with the SKEL template mesh [13] to produce a DRR-specific template model, offering an accurate and adaptable solution for the reconstruction of intricate skeletal structures.

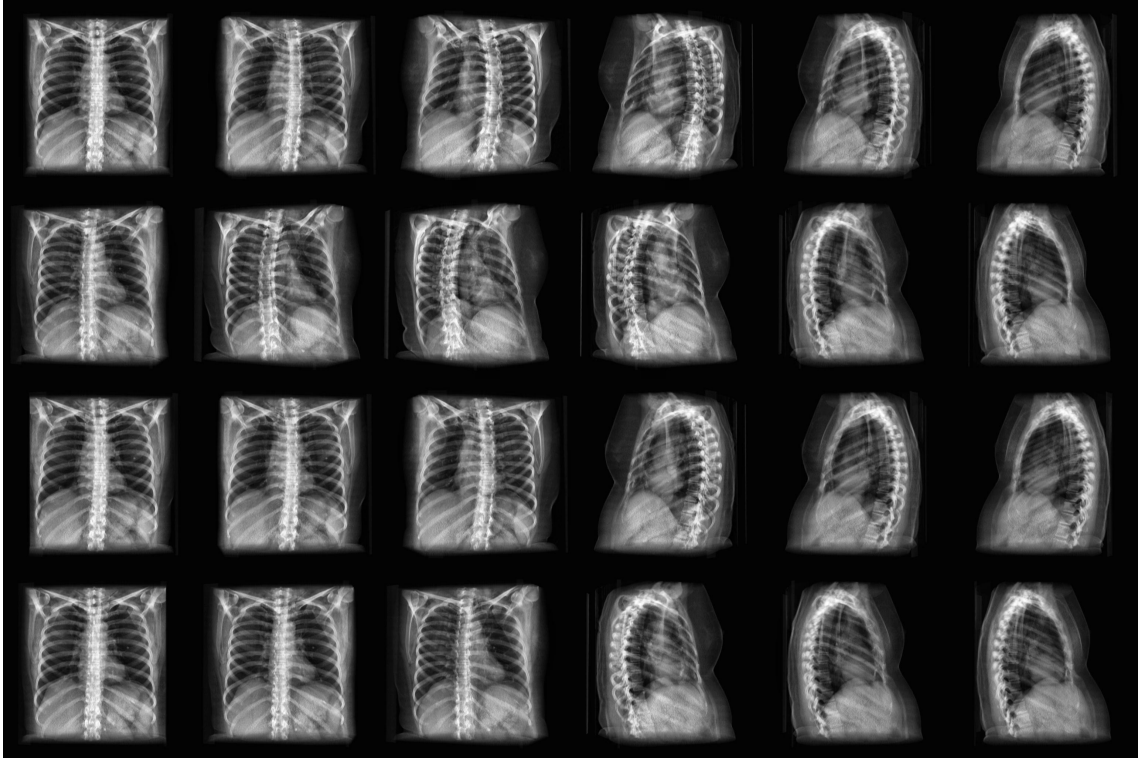
The architecture of the X2BR model is illustrated in Figure 4.2. This network takes as input a DRR image of size 224×224 along with a set of 3D coordinates for T random points in space. The X2B network’s output, which is called as the coarse shape,

Figure 4.2: Anterior-posterior DRR is used as input for both YOLOv9 and X2B for inference. The regions are detected via YOLOv9 network and DRR specific template mesh is extracted from the template mesh model using the detected regions. DRR specific template model is registered to the coarse shape, which is the output of the X2BR model.



is the occupancy probability for each point, indicating the likelihood that a given point in space is occupied. To reconstruct a 3D coarse surface, the occupancy function is conditioned on a single DRR image. During training of the model, the input DRR image is randomly selected from a set of images captured at different angles. This process leverages the variability in viewpoints as shown in Figure 4.3 to provide a more comprehensive reconstruction. This occupancy function [10], which is central to our approach, implicitly defines the bone surface as a decision boundary. These varied perspectives of the chest and spine are used to enhance the network’s ability to accurately reconstruct 3D organ and skeletal volumes from a single planar X-ray image.

Figure 4.3: DRR images for multi-view training of the X2B network. The figure displays a collection of DRR images captured at various angles around the anatomical structure. These images serve as input data for the X2B network, enabling it to learn multi-view representations for accurate 3Dr and enhanced skeletal analysis. The diverse viewing angles facilitate robust feature extraction and improve the network’s ability to generalize across different spatial perspectives.

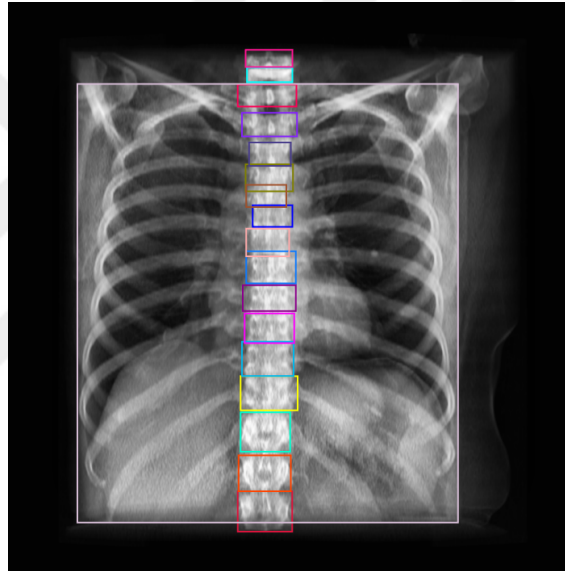


The fine-tuned YOLOv9 model is employed in this work due to its favorable balance between average precision (AP) and inference latency, making it well-suited for efficiently extracting regions of interest in skeletal imaging. Specifically, the model is used to identify and localize vertebrae, ribs, costal cartilages, and the sternum from X-ray images. Vertebrae are further classified into cervical, thoracic, and lumbar regions, with each vertebra accurately numbered. The ribs, costal cartilages, and sternum are detected collectively as a single region, as illustrated in Figure 4.4, providing a comprehensive segmentation of the skeletal system.

To establish a template model for the human skeleton, the SKEL Biomechanical

Skeleton Model (BSM) [13] is utilized. This custom skeleton model was developed by the SKEL team using the OpenSim framework [96], a widely recognized platform for biomechanical modeling and simulation. The BSM serves as a precise and anatomically accurate mesh template, forming the foundation for subsequent analysis and reconstruction tasks in this study.

Figure 4.4: *YOLOv9 region labeling for vertebral detection. The figure illustrates the application of the YOLOv9 object detection model for identifying and labeling vertebral regions in a chest X-ray. Each bounding box highlights a detected vertebra, facilitating precise localization and segmentation. This labeling process is critical for downstream tasks, such as generating person-specific templates for 3Dr.*



To summarize the data flow in the X2BR network, the input DRR is first processed by the fine-tuned YOLOv9 model to identify and classify the vertebrae (including their types and numbers) and the thoracic region. Simultaneously, the same DRR is provided as input to the X2B network, which, along with 3D location information, generates a coarse 3D skeletal shape. A patient-specific template mesh is constructed by extracting the relevant bone structures (classified by the YOLOv9 model) from the pre-defined template model. Finally, the patient-specific template model and the coarse shape generated by the X2B network are used as inputs to the Geodesic Bayesian Coherent Point Drift (GBCPD)

algorithm to achieve a refined and accurate reconstruction.

4.0.2 Training

The X2B network is trained to reconstruct 3D skeletal structures from DRRs. As depicted in Figure 4.1 during training of the model randomly selected input DRR image that is captured at different angles is passed through a ConvNeXt backbone, which extracts multi-dimensional feature representations. These features are then processed through a series of DenseNet blocks that include Conditional Batch Normalization (CBN) layers, parameterized by β_i and γ_i , to refine the latent representations. The network iteratively processes these features through multiple dense layers, producing a final prediction that reconstructs the 3D skeletal structure. The loss function $L(\theta)$ is computed by comparing the predicted 3D output to the GT, enabling the network to optimize its parameters for accurate reconstruction.

The objective is to estimate the occupancy at every point $p \in \mathbb{R}^3$ to derive the coarse shape from the X2B network’s output. The occupancy function

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

defines whether a 3D point is occupied. This function can be approximated by a neural network, which assigns each point p an occupancy probability between 0 and 1, using a binary classification network’s decision boundary to implicitly represent the object’s surface.

The network, represented as $f : \mathbb{R}^3 \times X \rightarrow \mathbb{R}$, maps the pair (p, x) to a real number indicating the probability of occupancy, where $x \in X$ is the observation conditioning the 3Dr task.

Training the binary-classification network involves computing the mini-batch loss $L_B(\theta)$ at randomly sampled points within the object’s 3D bounding volume. This is

expressed as:

$$L_B(\theta) = \frac{1}{|B|} \sum_{i=1}^{|B|} L(f_\theta(p_i, x_i), o_i) \quad (4.2)$$

where x_i is the i^{th} observation in batch B , $o_i \equiv o(p_i)$ represents the true occupancy at point p_i , and $L(\cdot, \cdot)$ is the binary cross-entropy classification loss. The method’s effectiveness depends on the sampling strategy for locations p_i during training, with optimal results achieved through uniform sampling within the object’s bounding box [10].

For all experiments, the Adam optimizer [79] was employed with a learning rate (η) of 10^{-4} . The default hyperparameters provided by PyTorch were used: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$.

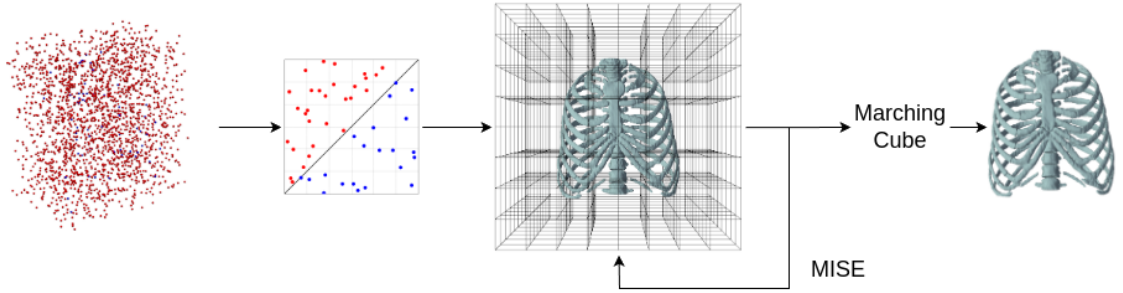
4.0.3 Inference

Figure 4.5 demonstrates the inference of 3D meshes from occupancy predictions at the output of the X2B network. In the X2B implementation, a modified Multiresolution IsoSurface Extraction (MISE) algorithm [10] is employed, as detailed in the occupancy network inference approach, to enable efficient high-resolution mesh extraction.

The process starts with the discretization of the volumetric space and the evaluation of occupancy values at this initial resolution. Voxelization involves thresholding to classify each voxel as inside or outside the object based on the model’s predictions. This critical step converts probabilistic outputs into a discrete 3D representation, facilitating accurate reconstruction from a single image viewpoint. Grid points exceeding an occupancy threshold, set to $\tau = 0.2$ for the X2B method, determine the surface thickness, balancing accuracy and completeness as per ONet recommendations [10].

Active voxels are identified and subdivided iteratively until the desired resolution is reached. The Marching Cubes algorithm then generates the initial mesh, which is refined using the Fast-Quadric-Mesh-Simplification algorithm [80]. This algorithm applies

Figure 4.5: For 3D mesh inference with X2B model, a modified Multiresolution Iso-Surface 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.



iterative edge contraction and quadric error metrics, followed by gradient-based optimization, ensuring mesh quality enhancement and simplification.

4.0.4 Registration

The Geodesic-Based Bayesian Coherent Point Drift algorithm [11] introduces a non-rigid registration method addressing a significant drawback of the traditional Coherent Point Drift (CPD) algorithm. CPD, while prevalent for shape matching and deformation, often unnaturally deforms shapes with neighboring parts, such as human legs. This issue arises from the proximity-based deformation constraint known as motion coherence. The GBCPD++ method redefines motion coherence using geodesic distances, the shortest paths on a shape’s surface, to mitigate inappropriate deformations. Numerical experiments demonstrate the efficacy of the geodesic-based approach in avoiding CPD’s shortcomings and its scalability to handle shapes with millions of points. Key contributions include utilizing geodesic and Gaussian kernels for improved registration, providing a theoretical basis for converting indefinite geodesic kernels into positive-semidefinite

ones, and employing the Nyström method and parallel computations to accelerate the process. The GBPCD++ method, integrated into the X2BR model, refines the reconstruction of 3D skeletal structures by enabling anatomically consistent registration of detailed vertebral features. By leveraging geodesic distances and scalable computations, it addresses complex deformations and enhances the accuracy of skeletal alignment.

4.1 Dataset

The dataset curated for this study forms the cornerstone for advancing accurate and reliable 3D bone reconstruction from X-ray images. The scarcity of publicly available paired datasets comprising X-ray images and corresponding 3D bone meshes necessitated the creation of a custom, high-quality dataset. By leveraging advanced segmentation techniques, state-of-the-art DRR technology, and extensive preprocessing pipelines, the dataset was meticulously designed to address the specific requirements of the proposed X2B network. The dataset encompasses a comprehensive collection of CT scans, DRR images, and associated 3D bone meshes, which were systematically processed to ensure uniformity, precision, and suitability for training and validation purposes. This section provides a detailed overview of the data acquisition process, segmentation methodologies, generation of DRRs, and the computation of occupancy values to enable efficient and accurate mesh reconstruction.

4.1.1 Segmentation with TotalSegmentator

The X2B network is designed to isolate various organs from X-ray images using a single network, specifically targeting the extraction of ribs, costal cartilages, and vertebrae. This emphasis is due to the abundant availability of thorax region data and the higher HU contrast of mineral-rich bones compared to other tissues. Early experiments with bone segmentation models showed the superior accuracy of TotalSegmentator [97], leading us to focus on bone segmentation. The scapula bones were excluded due to the

high incidence of segmentation errors, such as holes in the bone structures. As a result, the selection of body segments and the architectural approach was influenced by data availability and the precision of the segmented mesh from CT scans. The steps for dataset generation are summarized below.

- Data collection through The Cancer Imaging Archive (TCIA) [82].
- Bone segmentation from CT images as watertight manifold meshes via TotalSegmentator [97]
- Occupancy value calculation for the points in the bone meshes.
- Collecting corresponding DRR images of the CT scans from multi views.
- Applying CLAHE contrast enhancement to obtain enhanced DRR images.

A portion of the NLST dataset [83, 84] was downloaded from TCIA [82] using the NBIA Data Retriever Tool. For training and validating our proposed mesh reconstruction approach, an ideal dataset would include a large collection of paired X-rays and corresponding mesh reconstructions. However, such datasets are scarce and expensive to obtain. To address this, we used DRR technology [90] to synthesize corresponding X-rays from real CT volumes.

In the X2B network, approximately 964 subjects from the publicly available National Lung Screening Trial (NLST) dataset were selected [91], which contains low-dose CT scans from 26,254 subjects in TCIA [82]. CT scans that did not cover the entire thorax region and cases with flawed CT images were excluded. Suitable rotation matrices were applied to align CT scans that were not oriented similarly, ensuring uniform alignment with respect to the anterior, posterior, left, and right anatomical planes of the patients. Additionally, low-quality segmentations were removed following the creation of 3D organ meshes.

From the compiled NLST dataset [91], one random CT scan per patient from various years was selected when multiple scans were available. In total, 3,640 3D-CT scans were collected from the 964 selected subjects. The 3D organ meshes were segmented using TotalSegmentator, which achieved a Dice Similarity Coefficient (DSC) of 0.943 ± 0.04 on the test dataset [97]. A Python script was developed [85, 86, 87, 88] to streamline the segmentation process. The script resampled the DICOM images to a 1.5mm x 1.5mm x 1.5mm spacing using B-spline interpolation, segmented the bone regions, exported the closed surface representation of multiple segments to files, and merged all segments into a single mesh.

4.1.1.1 Mapping DRR Images to Corresponding Segmented Volumes

A total of 3,640 3D-CT scans were collected from 964 subjects and resampled to a resolution of 512 x 512 x 512 pixels with a voxel size of 1.0mm x 1.0mm x 1.0mm. Using a Siddon Ray Tracing algorithm on a GPU [90, 92], 512 x 512 DRR images with a spacing of 1.0 x 1.0 were generated. To enhance the contrast and resolution of local details, Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied to the DRR images.

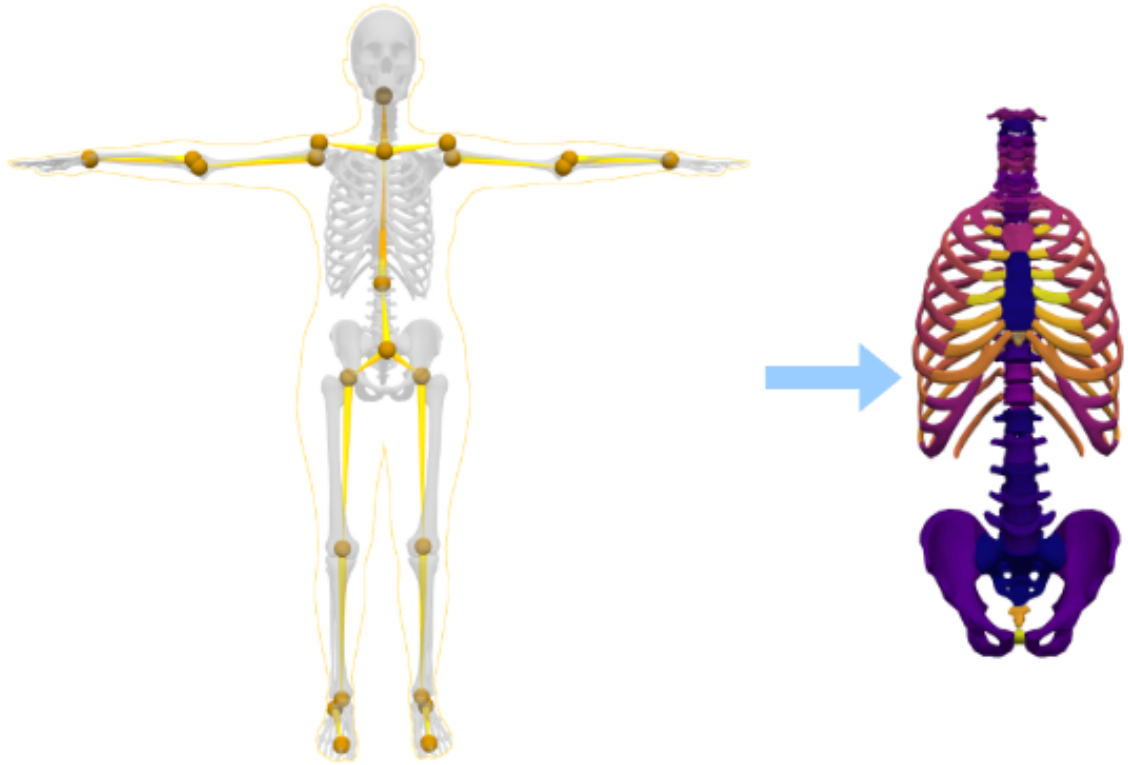
4.1.2 Bone Localization and Detection

Anterior-Posterior (AP) DRR images with a resolution of 512x512 pixels are acquired and subsequently annotated using the Roboflow software suite [98]. Initial annotations are performed manually to establish a baseline of high-quality labels. To efficiently annotate the remaining dataset, the Roboflow autolabeler is utilized, leveraging its automated labeling capabilities to generate preliminary annotations. These automatically generated labels undergo a thorough review process, during which inaccuracies are identified and corrected to ensure the highest level of labeling precision. The finalized dataset, comprising both manually and automatically refined annotations, is then employed to

train the YOLOv9 model, ensuring robust and accurate detection performance.

4.1.3 DRR Specific Template Model

Figure 4.6: *SKEL BSM to template mesh reconstruction.* This figure illustrates the process of generating a subject-specific skeletal template model. The SKEL Body Shape Model (BSM) on the left is used as the baseline for anatomical structures, which is further processed to generate a template mesh, as shown on the right.



The watertight meshes of the thorax, pelvis, and spine sections from the SKEL Body Shape Model (BSM) serve as the foundation for constructing a DRR-specific template model, as illustrated in Figure 4.6. The SKEL BSM provides anatomically accurate and geometrically consistent representations of the skeletal structures, ensuring a robust starting point for template generation. These meshes undergo a series of processing steps, including refinement and adaptation, to produce a subject-specific skeletal template mesh tailored for the non-rigid registration component. The resulting template mesh, depicted

on the right in Figure 4.6, is designed to align seamlessly with the input DRR data. This alignment is critical for enabling precise deformation and accurate registration, which are essential for generating detailed and anatomically consistent reconstructions in downstream tasks. The integration of the SKEL BSM into this workflow ensures a high degree of anatomical fidelity and computational efficiency, enhancing the overall performance of the registration process.

4.1.4 Calculation of Occupancy Values

The trimesh Python library is used to ensure the watertightness of the meshes, which is crucial for determining point occupancy within a mesh (e.g., for calculating IoU) [10]. All meshes are centered and rescaled to fit a 3D bounding box centered at (0,0,0) with the longest edge equal to 1, aligning them with voxelization formats. The bounding box's maximum dimension among the x, y, and z coordinates of the 3D lung volumes is used to normalize all meshes uniformly. As in Occupancy Networks (ONet) [10], voxel representations facilitate efficient querying of occupancy probabilities in 3D space, enabling detailed and accurate 3Ds [10].

Using the voxelization method from Choy et al. [93], 32^3 voxel grids are generated. The dataset is split into training, validation, and test subsets in a 7:2:1 ratio. All voxels intersecting the mesh surface are labeled as occupied. Evaluation involves selecting a random point within each voxel to determine its position relative to the mesh—inside or outside. If the point is inside, the voxel is labeled as occupied. For detailed analysis, 100,000 points are generated within a padded unit cube, and their positions relative to the watertight mesh are determined by counting intersections with a ray parallel to the z-axis; an even count indicates an external point, while an odd count signifies an internal point. Both the 100,000 point coordinates and their occupancies are saved, and during training, a subset of 2,048 points is randomly selected.

4.2 Experiments

This section presents the simulation results and comparative analysis of the X2V, X2B, and X2BR networks. It provides a comprehensive explanation of the evaluation metrics used and discusses the results in detail. The 3Dr performance of X2BR is compared against two recent single image 2D/3Dr models from the literature: D2IM-Net [46], and ED2IF2-Net [47]. The section includes an in-depth discussion of the evaluation metrics and a thorough analysis of the findings.

4.2.1 Comparative Studies

This section explains the details of D2IM-Net and ED2IF2-Net models, which are used for performance comparison.

4.2.1.1 D2IM-Net

D2IM-Net [46] is a single-view 3Dr network that excels in recovering both topological structures and fine surface details by disentangling 3D representations into a coarse implicit field and displacement maps for the front and back surfaces. Its pipeline leverages global and local features via two decoders and fuses the results, achieving high-quality reconstructions. Unlike IM-Net [41] and DISN [43], which struggle with fine detail recovery, D2IM-Net introduces a Laplacian loss that enhances surface detail representation, yielding superior performance in Chamfer Distance (CD), IoU, and Edge Chamfer Distance metrics. Applications include detail transfer between objects, reconstruction from “in-the-wild” images, and embedding features like logos into reconstructed shapes. While it is highly effective, limitations include its assumption of flat surfaces for surface detail representation, challenges with overhanging or highly curved surfaces, and some degradation in accuracy when viewed from perspectives significantly different from the input. Overall, D2IM-Net represents a significant advancement in detail-preserving 3Dr and sets

a new standard for single-view reconstruction tasks.

4.2.1.2 ED2IF2-Net

ED2IF2-Net [47] is a novel single-view 3Dr network that leverages the Pyramid ViT (PVT) to reconstruct high-fidelity 3D shapes from RGB images. The network disentangles the implicit field into two components: a deformed implicit field capturing topological structures and an enhanced displacement field for surface details. The architecture comprises three decoders: a coarse shape decoder for base shapes, a deformation decoder using implicit field deformation blocks (IFDBs) for structural refinement, and a surface detail decoder equipped with hybrid attention modules (HAMs) for detail enhancement. A novel loss function integrates coarse shape, deformation, overall shape, and Laplacian terms to ensure comprehensive learning.

ED2IF2-Net achieves state-of-the-art performance on the ShapeNet dataset, surpassing methods like D2IM-Net and DISN in metrics such as IoU and Chamfer Distance (CD). It demonstrates robustness in applications like surface detail transfer and logo embedding. While it excels in both reconstruction quality and inference speed, limitations include suboptimal handling of redundant local features and a lack of real-time reconstruction capabilities. Future work aims to optimize the algorithm for embedded platforms and explore advanced transformers for further enhancement. ED2IF2-Net is a benchmark solution in 3Dr, advancing both theoretical and practical applications in computer vision and robotics.

4.2.2 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, mean absolute deviation along the z-axis, and normal consistency score metrics.

The Chamfer L_1 distance [10], denoted as $C-L_1$, is a crucial metric for assessing

the similarity between two 3D shapes. It calculates the average nearest-point Euclidean distance between the surfaces of the predicted 3D model and the GT model. This metric is symmetric and normalized, enabling a robust comparison of 3Dr accuracy. The Chamfer L_1 distance is computed by identifying the nearest vertex in the GT mesh for each vertex in the predicted mesh and vice versa, averaging these minimum distances to provide an overall measure of spatial deviation between the two meshes.

$$\begin{aligned} \text{C-L}_1(M_{\text{PR}}, M_{\text{GT}}) \equiv & \frac{1}{2|\partial M_{\text{PR}}|} \int_{\partial M_{\text{PR}}} \min_{q \in \partial M_{\text{GT}}} \|p - q\| dp + \\ & \frac{1}{2|\partial M_{\text{GT}}|} \int_{\partial M_{\text{GT}}} \min_{p \in \partial M_{\text{PR}}} \|p - q\| dq. \end{aligned} \quad (4.3)$$

The IoU for 3D meshes is the ratio of their intersecting volume to the volume of their union. For evaluating 3D point cloud similarity, the voxelized IoU is calculated by 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).

$$\text{IoU}(P_{\text{PR}}, P_{\text{GT}}) \equiv \frac{|P_{\text{PR}} \cap P_{\text{GT}}|}{|P_{\text{PR}} \cup P_{\text{GT}}|}. \quad (4.4)$$

A higher IoU indicates greater overlap and similarity between the meshes, which is useful for evaluating accuracy in 3Dr. The meshes and point clouds are voxelized with a voxel size of 0.5 for IoU calculations, due to the limitations of the D2IM-Net and ED2IF2-Net methods, which do not generate watertight meshes.

Normal consistency (NC) is assessed by calculating the absolute dot product of surface normals from one mesh with the nearest neighbors in the other mesh, reflecting the alignment accuracy of corresponding surface elements.

The F-score is a widely recognized metric for evaluating 3D mesh reconstructions, combining precision and recall. It is calculated as the harmonic mean of the proportion of

correctly identified points in the predicted mesh (precision) and the proportion of actual mesh points that are correctly identified (recall).

$$F\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall} + \epsilon} \quad (4.5)$$

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

$$\text{Precision} = \frac{\sum_{i=1}^{N_{\text{PR}}} \mathbb{I}(d_{\text{PR}}^i < t)}{N_{\text{PR}}} \quad (4.6)$$

$$\text{Recall} = \frac{\sum_{i=1}^{N_{\text{GT}}} \mathbb{I}(d_{\text{GT}}^i < t)}{N_{\text{GT}}} \quad (4.7)$$

In these equations, N_{PR} and N_{GT} denote the number of points in the predicted and GT meshes, respectively. The function $\mathbb{I}(\cdot)$ is an indicator function that returns 1 if the condition 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. The threshold t is set to 0.02 in the experiments, with a small constant $\epsilon = 10^{-8}$ added to avoid division by zero in the F-score calculation.

To calculate the mean absolute z -axis error (mzse) between corresponding meshes, the following procedure was employed. The meshes were aligned rigidly using the Iterative Closest Point (ICP) algorithm, which computes an optimal transformation matrix comprising rotation and translation to minimize the distances between the vertices of the two meshes. After alignment, surface points were sampled uniformly from each mesh to ensure a consistent number of comparison points. For each sampled point on the first mesh, its nearest neighbor on the second mesh was identified using a k-d tree search, and the difference in z -axis coordinates was computed. The absolute values of these z -axis

differences were then aggregated across all points and all mesh pairs. Finally, the mean and standard deviation of the absolute z -axis differences were calculated to quantify the overall magnitude and variability of the error. This methodology ensures robust comparison by accounting for differences in mesh resolution and alignment.

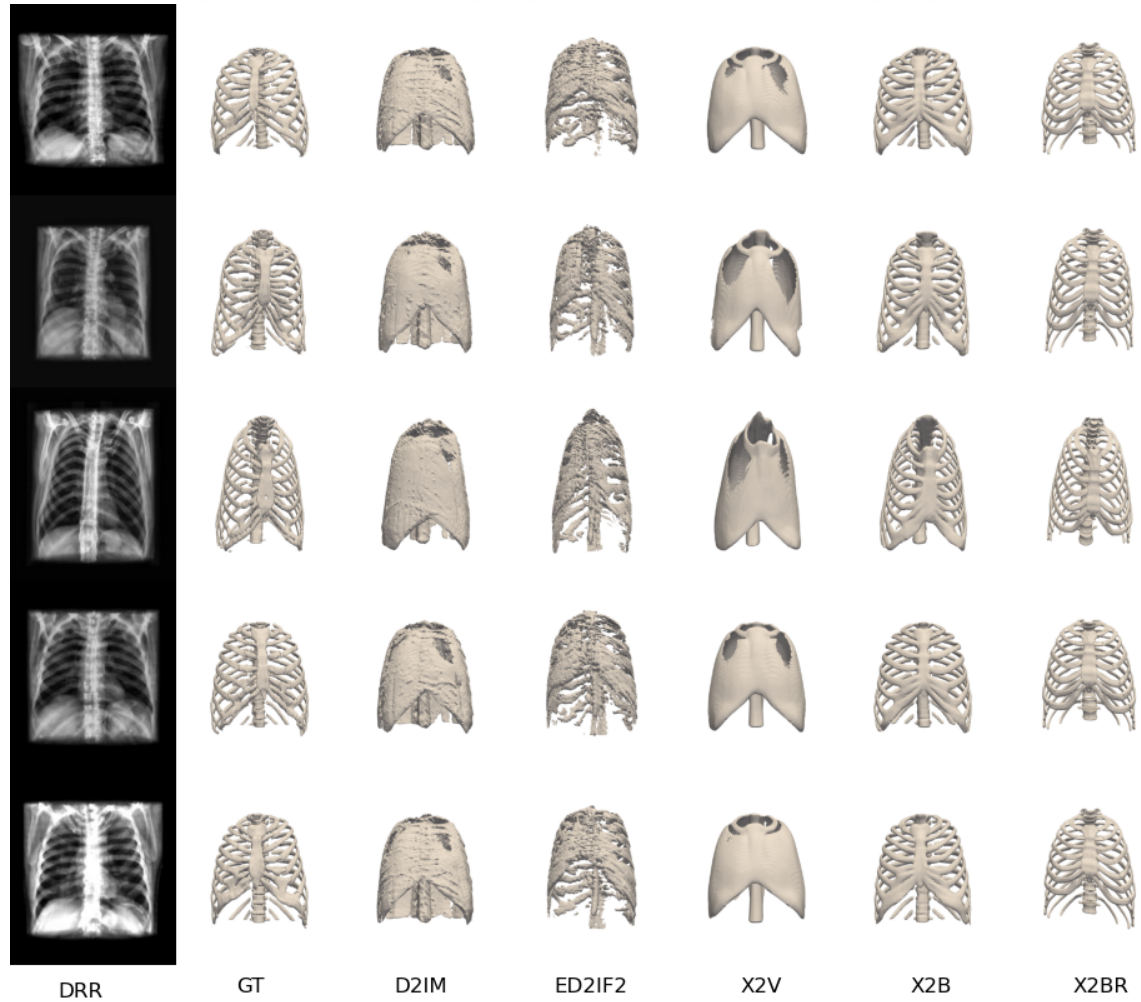
4.3 Simulations and Results

Figure 4.7 provides a comparative analysis of 3Dr results from different methods, including D2IM, ED2IF2, X2V, X2B, and X2BR, using DRRs as inputs. The first column displays the DRR images, followed by their corresponding GT 3D meshes. Each row represents results for a different DRR input, while the subsequent columns show the reconstructed outputs produced by each method. A progression in reconstruction quality is observed, with X2B and X2BR achieving significantly improved skeletal detail and accuracy compared to earlier methods like D2IM and ED2IF2. Notably, X2BR excels in preserving fine-grained anatomical features, such as rib curvature and spacing, which are critical for biomedical applications. These results highlight the advancements made by the X2BR model in capturing complex structural details, underscoring its robustness in 3Dr tasks involving medical imaging.

The inferior performance of D2IM and ED2IF2 in reconstructing skeletal structures, as shown in Figure 4.7, can be attributed to their reliance on displacement maps, which, while retaining background information, focus primarily on surface normal geometry. This representation may inadequately capture the depth and intensity variations inherent in X-ray images, which are crucial for reconstructing intricate skeletal structures such as ribs and spine. Furthermore, the overlapping anatomical features and background complexity in X-rays pose challenges for these methods, as their architectures may lack the capacity to effectively leverage the global context. The limited ability of D2IM and

ED2IF2 to infer fine details and resolve occlusions further hinders their reconstruction accuracy. In contrast, models like X2B and X2BR are explicitly designed to process X-ray inputs, utilizing advanced architectures to integrate contextual information and resolve background complexities, resulting in more anatomically precise and detailed reconstructions.

Figure 4.7: Comparison of reconstruction results across different methods. The figure presents DRR images (leftmost column) and their corresponding GT 3D meshes alongside reconstructed outputs from various methods: D2IM, ED2IF2, X2V, X2B, and X2BR. Each row corresponds to a different DRR input, while the columns illustrate the progression of reconstruction quality across the methods.



As demonstrated in Figure 4.7, the X2V network is proficient in reconstructing the

overall coarse shape but lacks the ability to capture fine local details. Conversely, the X2B network excels in reconstructing the thorax and spine regions with significantly enhanced local detail. The coarse shape extracted by X2V often fails to accurately reconstruct the rib cage.

Figure 4.8 provides a detailed comparison between X2B reconstructions and GT meshes for DRR inputs. The first column displays the DRR images, followed by the second column, which contains the front (GT-f) and back (GT-b) views of the GT meshes. Subsequent columns present the X2B reconstructions (X2B-f and X2B-b) alongside corresponding heatmaps (X2B-h-f and X2B-h-b), which visualize the reconstruction errors. The heatmaps highlight the discrepancies between the reconstructed meshes and the GT, with the color scale ranging from -20 mm (blue) to +20 mm (red). These visualizations emphasize regions of alignment as well as areas with reconstruction deviations. The front and back views provide a comprehensive understanding of the model's performance across different perspectives, demonstrating the capability of X2B in accurately reconstructing skeletal structures while identifying areas for improvement.

Figure 4.9 provides a statistical representation of the heatmap distance values shown in Figure 4.8. The histogram aggregates the point-to-point distances across 364 test images, capturing the overall alignment accuracy. The mean distance is 2.192 mm, with a standard deviation of -0.305 mm on the lower side and +4.078 mm on the upper side. The red dashed line in the graph indicates the mean distance, while the green dashed lines represent ± 1 standard deviation from the mean. Together, Figures 4.8 and 4.9 offer both spatial and statistical insights into the X2B model's reconstruction performance, emphasizing its capability to align closely with the GT while identifying regions with higher deviations.

The consistent positive error observed in the heatmaps of the X2B model is primarily a consequence of the inherent characteristics and limitations of the occupancy

network architecture used for 3Dr. Occupancy networks represent surfaces implicitly by predicting a continuous probability field that determines whether a point in 3D space lies inside or outside the object. While this approach ensures smooth and differentiable surface reconstructions, it introduces a bias toward over-smoothing and over-estimating surface boundaries, particularly in regions with high-frequency geometric details, such as the rib cage and vertebrae. This overestimation manifests as systematic positive errors in the reconstructed surfaces when compared to the GT.

The inability of the network to perfectly capture sharp transitions and intricate anatomical features is further compounded by the resolution of the training data. During training, the 3D points used to condition the occupancy function are sampled at a finite density, which may not provide sufficient coverage for capturing fine skeletal structures. Moreover, the interpolation mechanisms inherent in occupancy networks smooth transitions between sampled points, further contributing to boundary inflation. Regularization techniques, which are employed to stabilize the training process and enforce smoothness in the reconstructed geometry, exacerbate this effect by discouraging sharp or abrupt changes in the predicted surface.

Additionally, the presence of overlapping structures in X-ray images adds another layer of complexity. The network may misinterpret these overlaps as part of the skeletal structure, resulting in slight outward shifts of the predicted surface. This issue is particularly prominent in thin or elongated anatomical regions, such as ribs, where the continuous representation of the occupancy network may "thicken" these features, further amplifying the positive error.

The GT meshes used for evaluation are typically highly detailed and precise, providing an idealized representation of the anatomical structure. Comparatively, the X2B model's predictions, constrained by the smoothness and generalization properties of the

Figure 4.8: Comparison of X2B and GT. The first column displays the DRR inputs, while the second column presents the GT meshes. The subsequent columns show the X2B reconstructions (X2B-f), X2B heatmaps (X2B-h-f), GT-b, X2B-b, and X2B heatmaps in the back view (X2B-h-b).

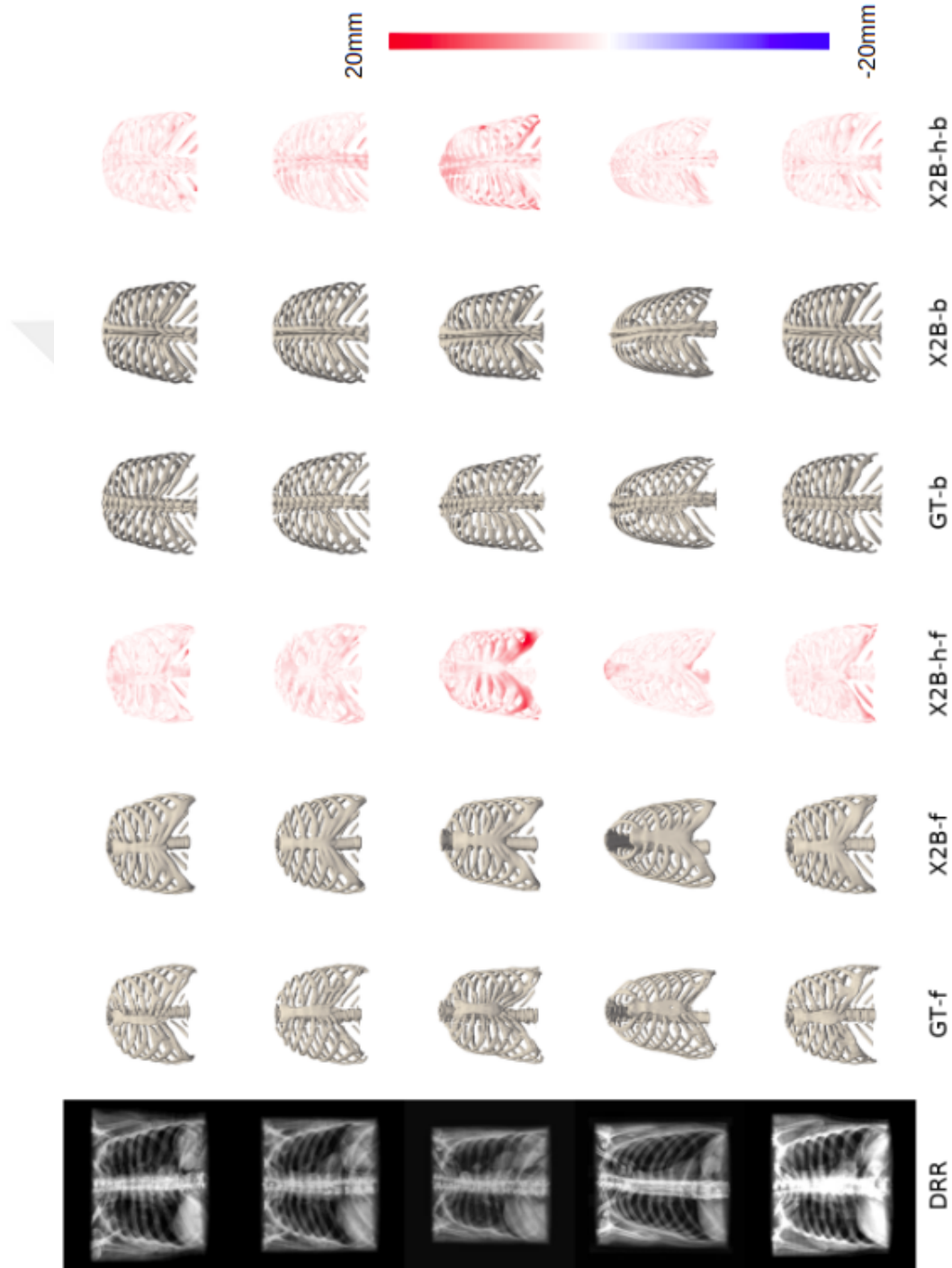
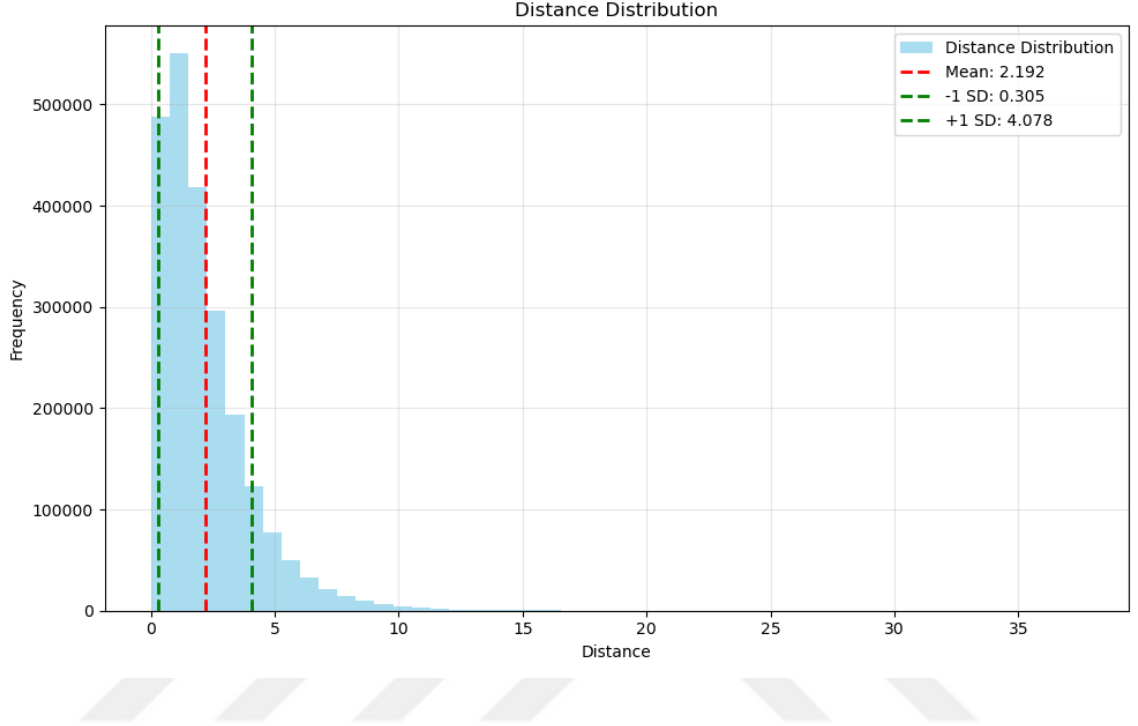


Figure 4.9: *Distribution of point-to-point distances for 364 test images. The mean distance is 2.192 mm, with a standard deviation of 1.886 mm. The histogram provides a visual representation of the frequency of distances, with the red dashed line indicating the mean and the green dashed lines representing ± 1 standard deviation from the mean.*



occupancy network, fall short in achieving the same level of detail and precision. As a result, even minor deviations, such as rounding of edges or slight expansions of the surface, become apparent as positive errors when compared against the GT.

In summary, the consistent positive error observed in the X2B heatmaps arises from the interplay of over-smoothing in occupancy network predictions, limitations in training data resolution and sampling, regularization effects, and challenges in interpreting complex overlapping structures in X-ray images. These factors collectively lead to a systematic bias in the reconstructed surfaces, resulting in anatomically consistent but slightly overestimated skeletal geometries.

Table 4.1 provides a comparative analysis of five reconstruction methods—D2IM, ED2IF2, X2V, X2B, and X2BR—in terms of their performance on three key metrics:

Table 4.1: Comparison of D2IM, ED2IF2, X2V, X2B, and X2BR, in terms of maze, maxe and maye metrics in milimeter for mesh reconstruction accuracy.

Method	maxe	maye	maze
D2IM	3.468 ± 3.157	2.772 ± 2.474	3.304 ± 3.095
ED2IF2	11.871 ± 10.267	6.254 ± 6.655	7.114 ± 6.794
X2V	4.545 ± 5.996	4.706 ± 9.352	4.442 ± 6.698
X2B	2.821 ± 2.717	2.515 ± 2.320	2.590 ± 2.443
X2BR	3.562 ± 3.909	3.667 ± 4.203	3.003 ± 2.793

mean absolute x -axis error (maxe), mean absolute y -axis error (maye), and mean absolute z -axis error (maze). These metrics, expressed in millimeters, are reported as “mean $\pm$ standard deviation” to reflect both the central tendency and variability of the reconstruction accuracy. Lower values across these metrics indicate higher precision in 3D mesh reconstruction. The standard deviation of the z -axis values for the GT meshes, calculated across 364 test samples, is found to be 84.358 mm, reflecting the variability in the vertical dimension of the dataset.

The slightly worse numerical performance of the X2BR model compared to X2B, as observed in Table 4.1, can be attributed to the additional architectural complexity and processing steps introduced in X2BR. While X2B employs a streamlined pipeline using ConvNeXt and CBN layers to directly generate a coarse 3Dr, X2BR enhances this process by integrating a DRR-specific template mesh and performing non-rigid registration through GBCPD++. Although these enhancements aim to improve fine-grained anatomical detail, they introduce potential sources of error, such as inaccuracies in region detection by YOLOv9, template alignment, and the registration process itself. This multi-step approach, while effective in refining local details, can result in minor inconsistencies in global alignment. In contrast, X2B’s simpler architecture avoids these challenges, resulting in more consistent numerical performance across reconstruction metrics. The trade-off between detail refinement and global accuracy highlights the complexity of balancing architectural improvements in medical imaging tasks.

ED2IF2 showed the highest errors ($maxe: 11.871 \pm 10.267$, $maye: 6.254 \pm 6.655$, $maze: 7.114 \pm 6.794$), indicating significant challenges in accurately reconstructing the 3D meshes. Among all methods, X2B achieved the best overall performance, demonstrating the lowest errors across all metrics ($maxe: 2.821 \pm 2.717$, $maye: 2.515 \pm 2.320$, $maze: 2.590 \pm 2.443$). This suggests that X2B provides the most precise 3D reconstructions, particularly in capturing fine structural details with minimal deviation. X2BR also showed competitive results with a $maze$ of 3.003 ± 2.793 , indicating its robustness in capturing vertical details. Meanwhile, X2V exhibited intermediate performance, with slightly higher errors than X2B and X2BR, but still outperforming ED2IF2.

These results highlight the effectiveness of X2B in achieving high reconstruction accuracy across all axes, making it a promising approach for 3D reconstruction tasks. The variability observed in the performance of other methods underscores the importance of evaluating multi-axis errors to comprehensively assess reconstruction accuracy.

Table 4.2: Comparison of D2IM, ED2IF2, X2V, X2B, X2BR 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
<i>D2IM</i>	0.903 ± 0.028	0.009 ± 0.001	0.912 ± 0.027	0.499 ± 0.006
<i>ED2IF2</i>	0.573 ± 0.041	0.019 ± 0.002	0.567 ± 0.048	0.495 ± 0.003
<i>X2V</i>	0.859 ± 0.057	0.009 ± 0.001	0.904 ± 0.050	0.496 ± 0.005
<i>X2B</i>	0.952 ± 0.038	0.005 ± 0.001	0.974 ± 0.038	0.505 ± 0.003
<i>X2BR</i>	0.875 ± 0.036	0.009 ± 0.001	0.913 ± 0.039	0.504 ± 0.003

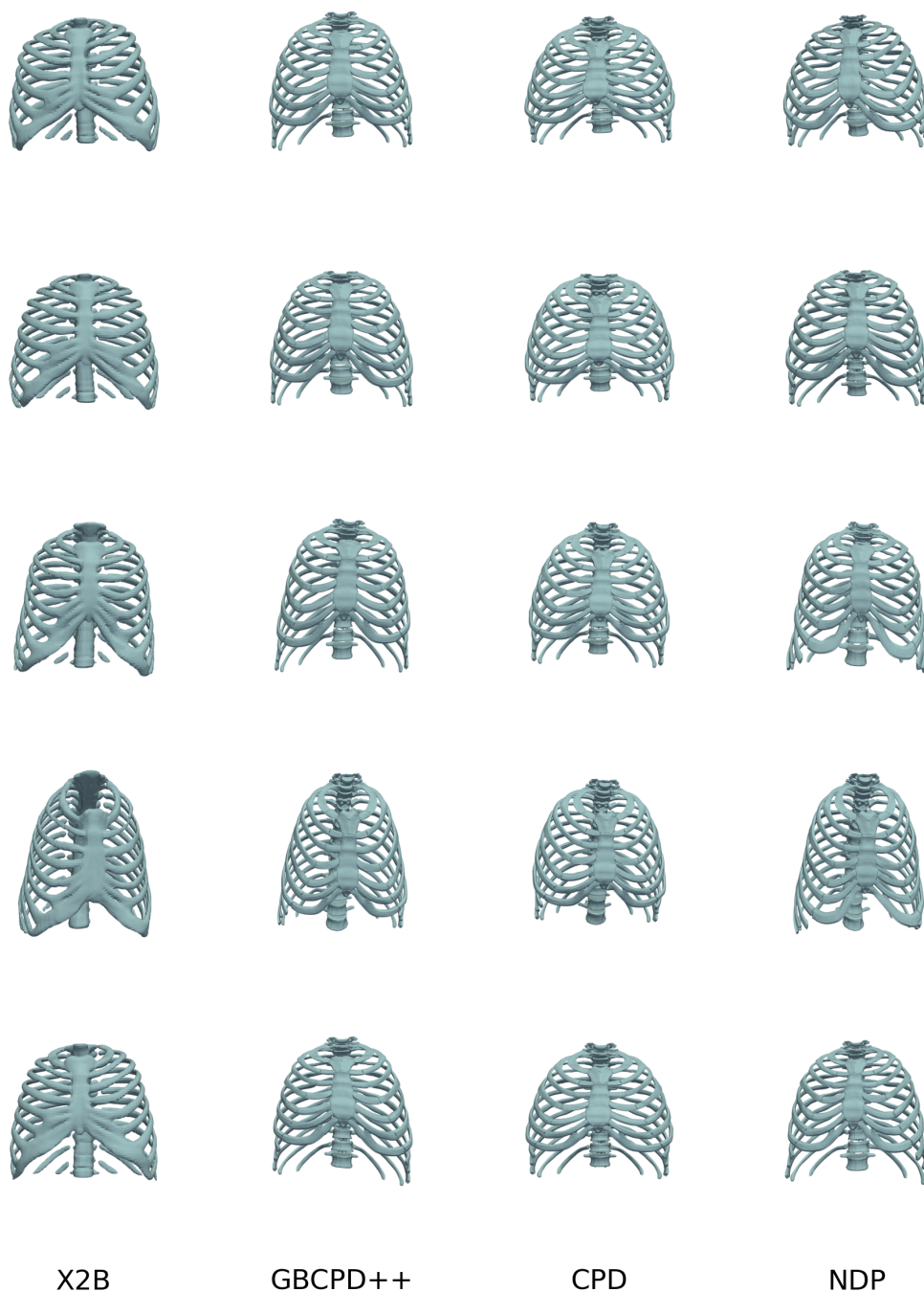
Table 4.2 provides a comparative evaluation of five methods—D2IM, ED2IF2, X2V, X2B, and X2BR—based on key metrics for mesh reconstruction accuracy: IoU, Chamfer-L1 distance, F-score (at $t = 0.02$), and Normal Consistency (NC). Among the methods, X2B achieves the best overall performance, with the highest IoU score (0.952 ± 0.038), the lowest Chamfer-L1 distance (0.005 ± 0.001), and the highest F-score (0.974 ± 0.038). These results indicate its strong capability to reconstruct accurate and consistent 3D meshes.

Conversely, ED2IF2 demonstrates the lowest performance across all metrics, with a significantly lower IoU (0.573 ± 0.041), the highest Chamfer-L1 distance (0.019 ± 0.002), and the lowest F-score (0.567 ± 0.048). The intermediate methods, such as X2V and X2BR, perform competitively, with X2BR showing notable improvement in NC (0.504 ± 0.003) and robust performance in other metrics. While D2IM performs relatively well, its NC score (0.499 ± 0.006) is lower compared to X2B and X2BR. This analysis highlights the superior reconstruction accuracy and consistency of the X2B model, demonstrating its advancements over earlier methods like ED2IF2 and its ability to strike a balance between geometric precision and consistency.

The analysis of reconstruction accuracy is extended in Figure 4.10, which evaluates the performance of three non-rigid registration methods—GBCPD++, NDP, and CPD—in refining anatomical alignment for rib cage models. The X2BR model employs the GBCPD++ algorithm, which demonstrates notable improvements in registration accuracy compared to the alternatives. While the NDP algorithm successfully predicts the global shape, it struggles with local deformations, often leading to distortions. In contrast, the CPD algorithm performs poorly in addressing both global and local deformations, highlighting its limitations in achieving precise anatomical reconstructions.

Figure 4.10 provides a comprehensive comparison of rib cage registration results using three methods: GBCPD++, CPD, and NDP. The first column presents the X2B network predicted meshes, which serve as the initial models before registration. The subsequent columns illustrate the registered outputs from each method. Each row corresponds to a different rib cage instance, allowing for a detailed evaluation of how well each registration method aligns the initial X2B-generated meshes. The comparison highlights structural variations introduced by different registration techniques, particularly in terms of curvature preservation and rib spacing consistency. The visual results suggest that GBCPD++ outperforms CPD and NDP in maintaining anatomical integrity, especially in

Figure 4.10: Comparison of registration results for rib cage models using three methods: GBCPD++, CPD, and NDP. The figure presents front views of the rib cage models. The first column displays the X2B network predicted mesh models, while the subsequent columns illustrate the registered results from each method (GBCPD++, CPD, and NDP)



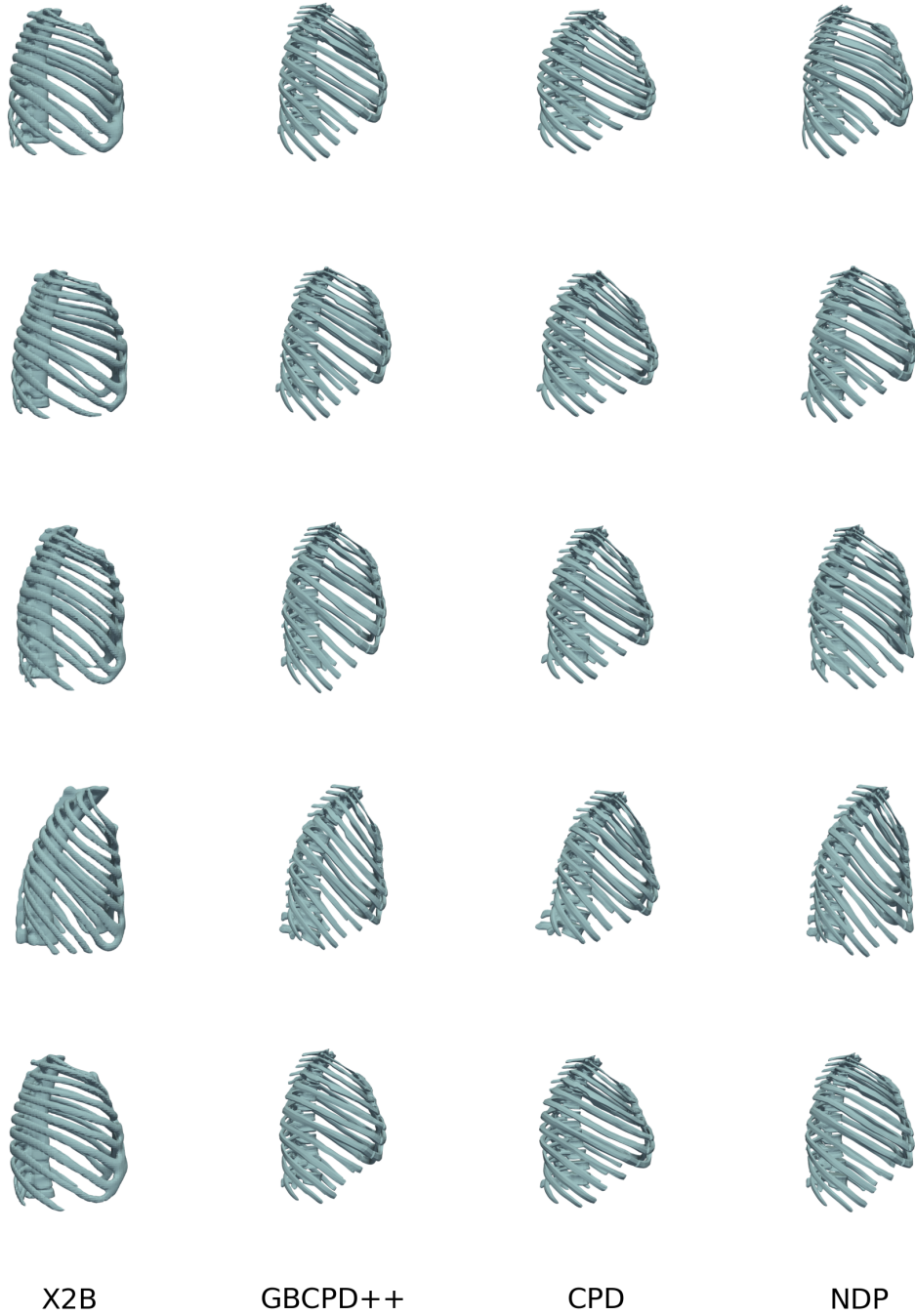
complex rib regions. This is further validated by the quantitative results in Table 4.3, where GBCPD++ achieves the highest IoU score of 0.900 and the lowest Chamfer-L1 distance of 0.010, confirming its superior registration accuracy. These findings highlight the effectiveness of GBCPD++ in ensuring precise and reliable non-rigid registration, reinforcing its robustness in maintaining geometric fidelity and anatomical consistency in medical imaging applications.

The CPD method, shown in the third column, produces results with moderate accuracy, evident from slight deviations in the alignment of rib structures compared to the GT meshes. This aligns with its Chamfer-L1 distance of 0.014 and IoU of 0.744, which indicate a reasonable performance but fall short of GBCPD++'s precision. On the other hand, the NDP results in the fourth column exhibit significant misalignment, particularly noticeable in the irregularities and distortions in the rib shapes. The lower IoU of 0.664 and higher Chamfer-L1 distance of 0.008 further confirm the limitations of NDP in handling complex deformations. These visual comparisons validate GBCPD++ as the most effective method for accurate non-rigid registration.

Figure 4.11 illustrates the side views of the rib cage models generated by the three non-rigid registration methods, providing additional insight into their ability to preserve shape consistency and alignment with the X2B-predicted GT meshes. Similar to Figure 4.11, GBCPD++ exhibits the most accurate registration, as evidenced by its ability to maintain the curvature and structural integrity of the ribs from the side perspective. This is consistent with its F-score of 0.931 and NC score of 0.500, indicating its robustness in preserving local geometry and achieving high alignment accuracy.

In comparison, CPD, shown in the third column, demonstrates moderate success in aligning the side views, with some observable deviations in the rib curvature. The NC score of 0.498 and F-score of 0.748 suggest that CPD can handle non-rigid deformations to a reasonable extent but lacks the precision of GBCPD++. The NDP results in the

Figure 4.11: Comparison of registration results for rib cage models using three methods: GBCPD++, CPD, and NDP. The figure presents side views of the rib cage models. The first column displays the X2B network predicted mesh models. The subsequent columns illustrate the registered results obtained from each method, GBCPD++, CPD, and NDP.



fourth column, however, show notable distortions and misalignments in the side views, particularly in the rib spacing and overall structure. This highlights the method’s limited ability to capture complex deformations, as reflected by its lowest NC score of 0.498 and F-score of 0.637. These observations reinforce the superior performance of GBCPD++ in achieving accurate and consistent mesh registration from different perspectives.

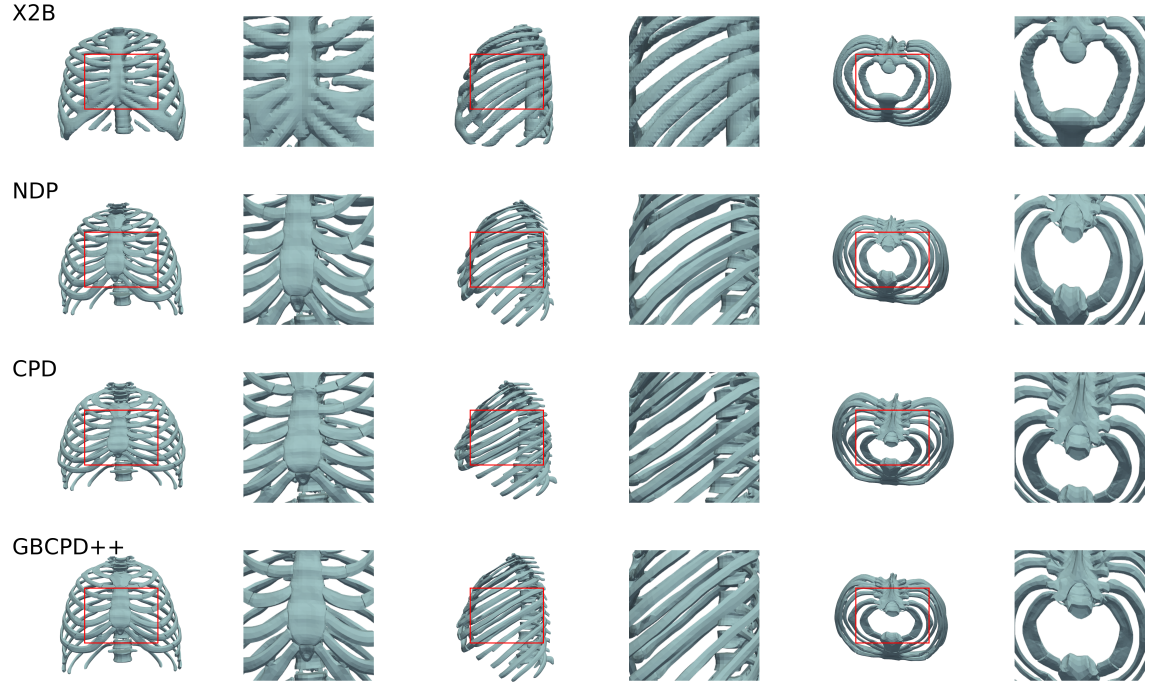
Table 4.3: Comparison of GBCPD++, NDP, CPD in terms of IoU, Chamfer-L1, F-score and NC metrics for mesh reconstruction accuracy.

Method	<i>IoU</i>	<i>Chamfer-L1</i>	<i>F-score(t=0.02)</i>	<i>NC</i>
<i>NDP</i>	0.664 ± 0.111	0.008 ± 0.001	0.637 ± 0.124	0.498 ± 0.003
<i>CPD</i>	0.744 ± 0.049	0.014 ± 0.001	0.748 ± 0.051	0.498 ± 0.003
<i>GBCPD++</i>	0.900 ± 0.019	0.010 ± 0.001	0.931 ± 0.019	0.500 ± 0.004

The results demonstrate that GBCPD++ outperforms the other two methods across all evaluation criteria, achieving the highest IoU of 0.900 ± 0.019 , the lowest Chamfer-L1 distance of 0.010 ± 0.001 , and the best F-score (0.931 ± 0.019) and NC (0.500 ± 0.004). This highlights its superior capability in accurately capturing fine details and ensuring high-quality mesh alignment. CPD shows moderate performance with an IoU of 0.744 ± 0.049 and a Chamfer-L1 distance of 0.014 ± 0.001 , indicating that it can handle non-rigid deformations to a reasonable extent but falls short of the precision exhibited by GBCPD++. NDP demonstrates the lowest performance with an IoU of 0.664 ± 0.111 and a Chamfer-L1 distance of 0.008 ± 0.001 , suggesting its limited ability to effectively handle complex deformations.

Figure 4.12 compares thorax reconstructions across different registration methods: GBCPD++, CPD, and NDP, with X2B results included as a baseline. Each row represents the outputs of a specific method, displaying six images per row: the first, third, and fifth images show the full thorax in front, side, and cross-sectional views, while the second, fourth, and sixth images provide zoomed-in views to emphasize finer details. Among the methods, GBCPD++ demonstrates the highest registration accuracy, closely aligning

Figure 4.12: Thorax reconstruction comparison across different models. The rows correspond to the X2B reconstructions, NDP, CPD, and GBCPD++ registrations, respectively. Each row contains six images, where the first, third, and fifth images display the full thorax in different perspectives (front view, side view, and cross-section), and the second, fourth, and sixth images provide zoomed-in views of specific regions to highlight finer details.



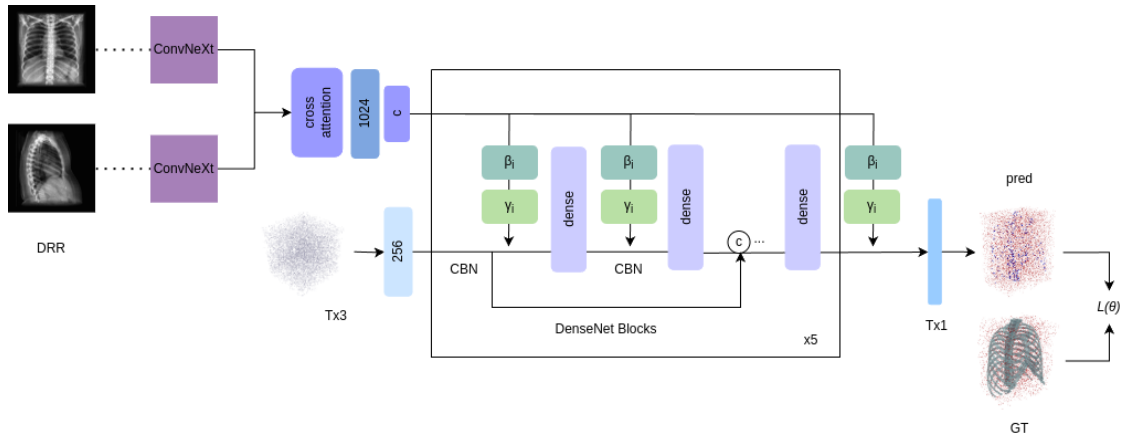
with the GT and preserving fine anatomical features. CPD performs moderately well, maintaining reasonable global alignment but with less precision in local regions compared to GBCPD++. NDP ranks third, showing noticeable discrepancies in both global structure and finer details, highlighting its limitations relative to the other methods. This figure underscores the superior performance of GBCPD++ in thorax reconstruction tasks.

The superior performance of GBCPD++ indicates its robustness and reliability in the challenging task of aligning predicted and template meshes with significant non-rigid deformations. These results validate the choice of GBCPD++ as the most effective method for this task and highlight the importance of advanced algorithms in achieving high-quality mesh reconstruction for person-specific applications.

4.4 Ablation Study

The X2B2 network extends the X2B architecture by integrating bi-planar DRR inputs to enhance reconstruction accuracy. This section systematically evaluates the impact of incorporating multi-view data and cross-attention mechanisms in the X2B2 framework. By comparing X2B with X2B2, we aim to quantify the advantages of multi-view fusion in refining 3D reconstructions. The study examines key performance metrics, including IoU, Chamfer-L1, F-score, and Normal Consistency (NC), alongside error measurements such as maxe, maye, and maze. The results highlight the incremental yet meaningful improvements achieved by X2B2, particularly in capturing finer anatomical details and maintaining geometric consistency. Through both quantitative analysis and visual comparisons, this section provides a comprehensive assessment of how architectural refinements in X2B2 contribute to enhanced 3D reconstruction performance.

Figure 4.13: X2B2 network training pipeline. This figure demonstrates the training process of the X2B2 network, which processes bi-planar DRRs as input. Each DRR is encoded using a ConvNeXt backbone, and their features are fused using a cross-attention mechanism to extract complementary information. The fused features are refined through dense blocks with Conditional Batch Normalization (CBN) layers, parameterized by β_i and γ_i , for accurate latent representation.



The X2B2 network, as illustrated in Fig. 4.13, builds upon the X2B architecture by incorporating bi-planar DRR inputs to enhance reconstruction accuracy. While X2B

processes a single DRR image, X2B2 utilizes two orthogonal DRRs, allowing it to leverage additional spatial information through a multi-head cross-attention mechanism. This mechanism aligns and fuses complementary features from the bi-planar inputs, resulting in a cohesive latent representation. Although the overall improvement in performance between X2B and X2B2 is modest, the incorporation of multi-view data fusion enables X2B2 to achieve greater numerical accuracy and slightly more detailed 3Drs.

Table 4.4: Comparison of X2B and X2B2 results in terms of IoU, Chamfer-L1, F-score and NC metrics for mesh reconstruction accuracy.

Method	<i>IoU</i>	<i>Chamfer-L1</i>	<i>F-score(t=0.02)</i>	<i>NC</i>
<i>X2B</i>	0.952 ± 0.038	0.005 ± 0.001	0.974 ± 0.038	0.505 ± 0.003
<i>X2B2</i>	0.964 ± 0.020	0.004 ± 0.001	0.984 ± 0.017	0.505 ± 0.003

A pre-trained ConvNeXt backbone is employed in X2B2 for feature extraction, with its classification head removed to focus solely on generating rich feature representations. These extracted features are refined and fused through the cross-attention mechanism, which facilitates the interaction between views to capture shared and unique details. An optional fully connected layer projects the fused features into a predefined latent space dimensionality, optimizing the network for reconstruction tasks.

Table 4.5: Comparison of X2B and X2B2 results in terms of maze, maxe and maye metrics in milimeter for mesh reconstruction accuracy.

Method	<i>maxe</i>	<i>maye</i>	<i>maze</i>
<i>X2B</i>	2.821 ± 2.717	2.515 ± 2.320	2.590 ± 2.443
<i>X2B2</i>	2.473 ± 2.143	2.431 ± 2.212	2.441 ± 2.247

The performance comparison between X2B and X2B2, as detailed in Tables 4.4 and 4.5, highlights the numerical advantages of the X2B2 network in 3Dr tasks. Table 4.4 compares the results based on IoU, Chamfer-L1, F-score, and Normal Consistency (NC) metrics. While both models perform exceptionally well, X2B2 demonstrates marginal improvements over X2B, achieving the highest IoU (0.964 ± 0.020) and the lowest Chamfer-L1 distance (0.004 ± 0.001), indicating its ability to produce more geometrically accurate

reconstructions. Furthermore, X2B2 attains the highest F-score (0.984 ± 0.017), reflecting its robustness in reconstructing fine structural details, with both models maintaining identical NC scores (0.505 ± 0.003).

Although the differences in performance between X2B and X2B2 are not drastic, X2B2 demonstrates a consistent numerical advantage, particularly in preserving finer anatomical details in 3Dr. Its ability to leverage multi-view data makes it especially suited for applications such as medical imaging analysis, three-dimensional reconstruction, and other multi-perspective tasks. These architectural refinements underscore the network’s capability to provide more robust and accurate reconstructions while maintaining the simplicity and efficiency of the X2B framework.

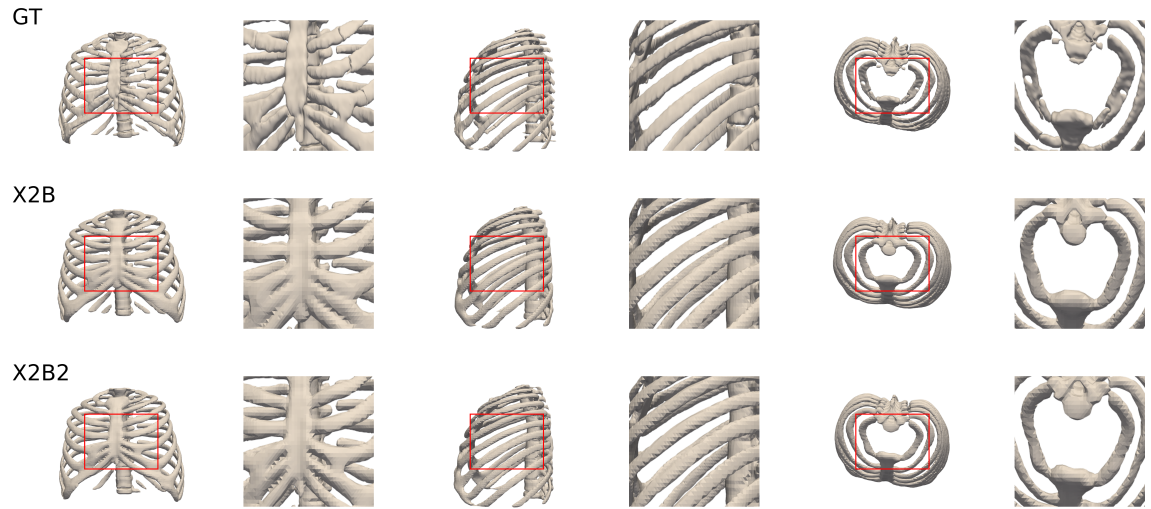
Table 4.5 further illustrates the improvement of X2B2 over X2B in terms of maxe, maye, and maze metrics, which assess reconstruction accuracy across different axes. X2B2 achieves the lowest mean errors across all metrics, with maxe (2.473 ± 2.143), maye (2.431 ± 2.212), and maze (2.441 ± 2.247), outperforming X2B in reducing positional errors. These results demonstrate X2B2’s consistent ability to refine reconstruction accuracy by leveraging its bi-planar input processing and multi-view data fusion capabilities.

The incremental improvements of X2B2, while modest, underscore its enhanced proficiency in producing accurate and detailed 3D meshes. By incorporating bi-planar DRR inputs and a robust cross-attention mechanism, X2B2 effectively captures complementary spatial information, offering improved consistency and precision in medical imaging applications. These numerical advancements further establish X2B2 as a reliable framework for high-fidelity 3Dr tasks.

Figure 4.14 provides a detailed comparison of thorax reconstructions for the GT, X2B, and X2B2 models. The zoomed-in regions (second, fourth, and sixth columns) highlight the differences in reconstruction quality, where X2B2 shows slightly more detailed

and accurate reconstructions compared to X2B, particularly in capturing finer anatomical features. This visual arrangement emphasizes the incremental improvements achieved by X2B2 over X2B in reconstructing complex thorax structures.

Figure 4.14: *Thorax reconstruction comparison across different models. The rows correspond to the GT, X2B reconstructions, and X2B2 reconstructions, respectively. Each row contains six images, where the first, third, and fifth images display the full thorax in different perspectives (front view, side view, and cross-section), and the second, fourth, and sixth images provide zoomed-in views of specific regions to highlight finer details.*



5. CONCLUSIONS

This thesis has demonstrated the efficacy of the X2V and X2B networks in the field of 3Dr from single X-ray images, particularly emphasizing their application to medical imaging. The X2V network has shown proficiency in capturing overall coarse shapes but has limitations in reconstructing finer local details, which is critical for accurate anatomical modeling. The X2B network, with its enhanced capabilities, excels in reconstructing complex structures such as the thorax and spine, offering significantly improved local detail.

Our X2V model has distinctly established itself as a superior tool in the domain of single-image 3Dr, 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.

The research extends to the X2BR network, which builds upon the X2B model by integrating advanced techniques such as non-rigid registration. This enhancement further refines the reconstruction accuracy, particularly in the thoracic region, ensuring more detailed and precise modeling of anatomical structures. The comparative analysis between the X2B and X2BR networks, alongside D2IM-Net and ED2IF2-Net, as well as recent 3D non-rigid registration models like Geodesic Bayesian Coherent Point Drift (GBCPD) and Neural Deformation Pyramid (NDP), underscores the superior performance of our proposed models across various evaluation metrics. In particular, X2B and X2BR consistently outperform these models in reconstructing complex anatomical regions, achieving higher accuracy in capturing fine details and reducing errors in 3D mesh reconstructions.

Future work on the X2BR network will focus on addressing its current limitations and enhancing its non-rigid registration capabilities to improve the accuracy and anatomical fidelity of 3Drs. A significant improvement will involve replacing the current low-vertex-count template with more detailed statistical models of the human thorax region, enabling higher resolution and better representation of intricate anatomical features. Additionally, leveraging advanced techniques such as the geodesic-based maximal clique search algorithm, which utilizes geodesic distances and compatibility graphs to ensure robust point correspondences even under large deformations, could further refine the registration process. Validating these advancements with real-world clinical datasets

and collaborating with medical professionals will ensure the model's applicability in diverse medical imaging scenarios, positioning the X2BR network as a powerful tool for diagnostics, surgical planning, and healthcare applications.

Furthermore, these networks hold significant potential for applications beyond traditional medical imaging. They can be integrated into augmented reality (AR), mixed reality (MR), and virtual reality (VR) systems, providing immersive and interactive experiences for medical professionals and students. In biomedical simulations, the detailed 3Drs can enhance the realism and accuracy of simulations, facilitating better training and research outcomes. Additionally, the networks can be utilized in finite element analysis (FEA) to create precise models for simulating biomechanical properties and behaviors, thus contributing to the advancement of both diagnostic and therapeutic technologies.

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,” *Medical Physics*, vol. 33, pp. 1573–1582, 2006.
- [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,” *Medical Physics*, vol. 35, pp. 480–486, 2008.
- [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 in Robotics and AI*, vol. 8, p. 716007, 2021.
- [4] P. Markelj, D. Tomaževič, B. Likar, and F. Pernuš, “A review of 3D/2D registration methods for image-guided interventions,” *Medical Image Analysis*, vol. 16, pp. 642–661, 2012.
- [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 *IEEE EMBS International Conference on Biomedical & Health Informatics*, pp. 1–4, 2019.
- [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 Transactions on Visualization and Computer Graphics*, vol. 26, pp. 960–970, 2020.
- [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 *Proceedings of Annual International Conference of IEEE Engineering in Medicine and Biology Society*, pp. 1608–1611, 2020.
- [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 *Proceedings of Medical Image Computing and Computer Assisted Intervention*, vol. 12904, pp. 259–268, 2021.
- [9] M. Nakao, M. Nakamura, and T. Matsuda, “Image-to-graph convolutional network for 2D/3D deformable model registration of low-contrast organs,” *IEEE Transactions on Medical Imaging*, vol. 41, pp. 3747–3761, 2022.

- [10] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, “Occupancy networks: Learning 3D reconstruction in function space,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4455–4465, 2019.
- [11] O. Hirose, “Geodesic-based bayesian coherent point drift,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 5816–5832, 2022.
- [12] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, “YOLOv9: Learning what you want to learn using programmable gradient information,” 2024.
- [13] M. Keller, K. Werling, S. Shin, S. Delp, S. Pujades, C. K. Liu, and M. J. Black, “From skin to skeleton: Towards biomechanically accurate 3d digital humans,” *ACM Transactions on Graphics*, vol. 42, pp. 1–12, 2023.
- [14] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A convnet for the 2020s,” in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.
- [15] J. T. Bushberg and J. M. Boone, *The Essential Physics of Medical Imaging*. Lippincott Williams & Wilkins, 2011.
- [16] W. R. Hendee and E. R. Ritenour, *Medical Imaging Physics*. John Wiley & Sons, 4th ed., 2003.
- [17] F. E. Boas, D. Fleischmann, *et al.*, “CT artifacts: causes and reduction techniques,” *Imaging in medicine*, vol. 4, pp. 229–240, 2012.
- [18] C. H. McCollough, S. Leng, L. Yu, D. D. Cody, J. M. Boone, and M. F. McNitt-Gray, “CT dose index and patient dose: They are not the same thing,” *Radiology*, vol. 259, pp. 311–316, 2011.
- [19] D. J. Brenner and E. J. Hall, “Computed tomography—an increasing source of radiation exposure,” *New England Journal of Medicine*, vol. 357, pp. 2277–2284, 2007.
- [20] Y. Xie, T. Takikawa, S. Saito, O. Litany, S. Yan, N. Khan, F. Tombari, J. Tompkin, V. Sitzmann, and S. Sridhar, “Neural fields in visual computing and beyond,” *Computer Graphics Forum*, vol. 41, pp. 1–36, 2022.
- [21] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, 2015.
- [22] T. Kim and T. Adali, “Approximation by fully complex multilayer perceptrons,” *Neural Computation*, vol. 15, pp. 1641–1666, 2003.
- [23] M. Stinchcomb, “Multilayer feedforward networks are universal approximators,” *Neural Networks*, vol. 2, pp. 359–366, 1989.

- [24] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “NeRF: Representing scenes as neural radiance fields for view synthesis,” *Communications of the ACM*, vol. 65, pp. 405–421, 2020.
- [25] S. Peng, M. Niemeyer, L. Mescheder, M. Pollefeys, and A. Geiger, “Convolutional occupancy networks,” in *Proceedings of the European Conference on Computer Vision*, vol. 12348, pp. 523–540, Springer, 2020.
- [26] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, “DeepSDF: Learning continuous signed distance functions for shape representation,” in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 165–174, 2019.
- [27] J. Ortiz, A. Clegg, J. Dong, E. Sucar, D. Novotny, M. Zollhöfer, and M. Mukadam, “iSDF: Real-time neural signed distance fields for robot perception,” in *Proceedings of Robotics: Science and Systems*, pp. 1–16, 2022.
- [28] A. Myronenko, X. Song, and M. Carreira-Perpinan, “Non-rigid point set registration: Coherent point drift,” in *Advances in Neural Information Processing Systems*, vol. 19, pp. 1–8, 2006.
- [29] O. Hirose, “A bayesian formulation of coherent point drift,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, pp. 2269–2286, 2020.
- [30] Y. Li and T. Harada, “Non-rigid point cloud registration with neural deformation pyramid,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 27757–27768, 2022.
- [31] P. Maken and A. Gupta, “2D-to-3D: A review for computational 3D image reconstruction from X-ray images,” *Archives of Computational Methods in Engineering*, vol. 30, pp. 85–114, 2023.
- [32] B. Goswami and S. K. Misra, “3D modeling of X-ray images: A review,” *International Journal of Computer Applications*, vol. 132, pp. 40–46, 2015.
- [33] 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,” *Journal of Imaging*, vol. 8, 2022.
- [34] J. Bednarik, P. Fua, and M. Salzmann, “Learning to reconstruct texture-less deformable surfaces from a single view,” in *International Conference on 3D Vision (3DV)*, pp. 606–615, 2018.
- [35] A. Tsoli and A. Argyros, “Patch-based reconstruction of a textureless deformable 3D surface from a single rgb image,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 4034–4043, 2019.

- [36] V. Golyanik, S. Shimada, K. Varanasi, and D. Stricker, “HDM-Net: Monocular non-rigid 3D reconstruction with learned deformation model,” in *Proceedings of Virtual Reality and Augmented Reality: EuroVR International Conference*, pp. 51–72, 2018.
- [37] S. Shimada, V. Golyanik, C. Theobalt, and D. Stricker, “IsMo-GAN: Adversarial learning for monocular non-rigid 3D reconstruction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–10, Jun 2019.
- [38] 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. 52–71, Sep 2018.
- [39] Y. Yuan, J. Tang, and Z. Zou, “Vanet: A view attention guided network for 3D reconstruction from single and multi-view images,” in *IEEE International Conference on Multimedia and Expo*, pp. 1–6, Jul 2021.
- [40] A. Salvi, N. Gavenski, E. Pooch, F. Tasoniero, and R. Barros, “Attention-based 3D object reconstruction from a single image,” in *International Joint Conference on Neural Networks*, pp. 1–8, Jul 2020.
- [41] Z. Chen and H. Zhang, “Learning implicit fields for generative shape modeling,” in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5939–5948, 2019.
- [42] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, “PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization,” in *Proceedings of IEEE/CVF International Conference on Computer Vision*, pp. 2304–2314, 2019.
- [43] Q. Xu, W. Wang, D. Ceylan, R. Mech, and U. Neumann, *DISN: Deep implicit surface network for high-quality single-view 3D reconstruction*, vol. 32, pp. 1–11. Curran Associates, Inc., 2019.
- [44] Y. Wang, Y. Zhuang, Y. Liu, and B. Chen, “MDISN: Learning multiscale deformed implicit fields from single images,” *Visual Informatics*, vol. 6, pp. 41–49, 2022.
- [45] W. Bian, Z. Wang, K. Li, and V. A. Prisacariu, “Ray-ONet: Efficient 3D reconstruction from a single RGB image,” in *Proceedings of British Machine Vision Conference*, pp. 1–13, British Machine Vision Association, 2021.
- [46] M. Li and H. Zhang, “D2IM-Net: Learning detail disentangled implicit fields from single images,” in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10246–10255, 2021.

- [47] X. Zhu, X. Yao, J. Zhang, M. Zhu, L. You, X. Yang, J. Zhang, H. Zhao, and D. Zeng, “ED²IF²-Net: Learning disentangled deformed implicit fields and enhanced displacement fields from single images using pyramid vision transformer,” *Applied Sciences*, vol. 13, p. 7577, 2023.
- [48] R. Chen, Y. Yang, and C. Tong, “G2IFu: Graph-based implicit function for single-view 3D reconstruction,” *Engineering Applications of Artificial Intelligence*, vol. 124, p. 106493, 2023.
- [49] M. S. Arshad and W. J. Beksi, “LIST: Learning implicitly from spatial transformers for single-view 3D reconstruction,” in *Proceedings of IEEE/CVF International Conference on Computer Vision*, pp. 9321–9330, 2023.
- [50] Y. Liang, W. Song, J. Yang, L. Qiu, K. Wang, and L. He, “X2Teeth: 3D teeth reconstruction from a single panoramic radiograph,” in *Medical Image Computing and Computer Assisted Intervention–MICCAI: 23rd International Conference, Proceedings, Part II 23*, pp. 400–409, 2020.
- [51] W. Song, H. Zheng, D. Tu, C. Liang, and L. He, “Oral-3Dv2: 3D oral reconstruction from panoramic x-ray imaging with implicit neural representation,” *arXiv*, 2023.
- [52] Y. Yang, Z. Cui, C. Li, and W. Wang, “ToothInpaintor: Tooth inpainting from partial 3D dental model and 2D panoramic image,” *arXiv*, 2022.
- [53] A. Corona-Figueroa, J. Frawley, S. Bond-Taylor, S. Bethapudi, H. P. Shum, and C. G. Willcocks, “MedNeRF: Medical neural radiance fields for reconstructing 3D-aware CT-projections from a single X-ray,” in *Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, pp. 3843–3848, 2022.
- [54] P.-H. Yeung, L. Hesse, M. Alias, M. Haak, W. Xie, A. I. Namburete, *et al.*, “Implicitvol: Sensorless 3D ultrasound reconstruction with deep implicit representation,” *arXiv*, pp. 1–11, 2021.
- [55] Y. Cai, J. Wang, A. Yuille, Z. Zhou, and A. Wang, “Structure-aware sparse-view X-ray 3D reconstruction,” in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11174–11183, 2024.
- [56] Y. Fang, L. Mei, C. Li, Y. Liu, W. Wang, Z. Cui, and D. Shen, “SNAF: Sparse-view CBCT reconstruction with neural attenuation fields,” *arXiv*, 2022.
- [57] X. Chen, A. Diaz-Pinto, N. Ravikumar, and A. F. Frangi, “Deep learning in medical image registration,” *Progress in Biomedical Engineering*, vol. 3, 2021.
- [58] 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,” *Medical Engineering and Physics*, vol. 31, pp. 681–687, 2009.

- [59] Y. Chaibi, T. Cresson, B. Aubert, J. Hausselle, P. Neyret, O. Hauger, J. 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,” *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 15, pp. 457–466, 2011.
- [60] 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,” in *Medical Imaging 2010: Image Processing*, vol. 7623, pp. 1–9, 2010.
- [61] J. Zhang, L. Lv, X. Shi, Y. Wang, F. Guo, Y. Zhang, and H. Li, “3-D reconstruction of the spine from biplanar radiography based on contour matching using the hough transform,” *IEEE Transactions on Biomedical Engineering*, vol. 60, pp. 1954–1964, 2013.
- [62] J. F. Hollenbeck, C. M. Cain, J. A. Fattor, P. J. Rullkoetter, and P. J. Laz, “Statistical shape modeling characterizes three-dimensional shape and alignment variability in the lumbar spine,” *Journal of biomechanics*, vol. 69, pp. 146–155, 2018.
- [63] P. R. Furrer, S. Caprara, F. Wanivenhaus, M. D. Burkhard, M. Senteler, and M. Farshad, “Patient-specific statistical shape modeling for optimal spinal sagittal alignment in lumbar spinal fusion,” *European Spine Journal*, vol. 30, pp. 2333–2341, 2021.
- [64] B. Aubert, C. Vazquez, T. Cresson, S. Parent, and J. A. de Guise, “Toward automated 3D spine reconstruction from biplanar radiographs using CNN for statistical spine model fitting,” *IEEE transactions on medical imaging*, vol. 38, pp. 2796–2806, 2019.
- [65] Y. Jiang, H. Zhai, K. Li, L. Li, H. Tao, M. Shao, X. Cheng, and H. Wang, “2D/3D shape model registration with X-ray images for patient-specific spine geometry reconstruction,” in *International Conference on Intelligent Robotics and Applications*, pp. 558–570, 2023.
- [66] S. Jecklin, C. Jancik, M. Farshad, P. Frnstahl, and H. Esfandiari, “X23D—intraoperative 3D lumbar spine shape reconstruction based on sparse multi-view X-ray data,” *Journal of Imaging*, vol. 8, p. 271, 2022.
- [67] Z. Chen, L. Guo, R. Zhang, Z. Fang, X. He, and J. Wang, “BX2S-net: Learning to reconstruct 3D spinal structures from bi-planar X-ray images,” *Computers in Biology and Medicine*, vol. 154, p. 106615, 2023.
- [68] C.-J. Yang, C.-L. Lin, C.-K. Wang, J.-Y. Wang, C.-C. Chen, F.-C. Su, Y.-J. Lee, C.-C. Lui, L.-R. Yeh, and Y.-H. D. Fang, “Generative adversarial network (GAN) for automatic reconstruction of the 3D spine structure by using simulated bi-planar X-ray images,” *Diagnostics*, vol. 12, p. 1121, 2022.

- [69] X. Ying, H. Guo, K. Ma, J. Wu, Z. Weng, and Y. Zheng, “X2CT-GAN: reconstructing CT from biplanar X-rays with generative adversarial networks,” in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10619–10628, 2019.
- [70] H.-G. Ha, J. Lee, G.-H. Jung, J. Hong, and H. Lee, “2D-3D reconstruction of a femur by single x-ray image based on deep transfer learning network,” *Innovation and Research in BioMedical Engineering*, vol. 45, no. 1, p. 100822, 2024.
- [71] D. Buttongkum, P. Tangpornprasert, C. Virulsri, N. Numkarunaruote, C. Amarase, T. Kobchaisawat, and T. Chalidabhongse, “3D reconstruction of proximal femoral fracture from biplanar radiographs with fractural representative learning,” *Scientific Reports*, vol. 13, p. 455, 2023.
- [72] H. Chui and A. Rangarajan, “A new algorithm for non-rigid point matching,” in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 44–51, 2000.
- [73] A. L. Yuille and N. M. Grzywacz, “The motion coherence theory,” in *Second International Conference on Computer Vision*, pp. 344–345, IEEE Computer Society, 1988.
- [74] A. Dosovitskiy, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv*, 2020.
- [75] 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.
- [76] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.
- [77] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.
- [78] H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville, “Modulating early visual processing by language,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [79] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference on Learning Representations*, pp. 1–15, 2015.
- [80] M. Garland and P. S. Heckbert, “Simplifying surfaces with color and texture using quadric error metrics,” in *Proceedings of IEEE Visualization Conference*, pp. 263–269, 1998.

- [81] 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,” *European Radiology Experimental*, vol. 4, pp. 1–13, 2020.
- [82] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, *et al.*, “The cancer imaging archive (TCIA): Maintaining and operating a public information repository,” *Journal of Digital Imaging*, vol. 26, pp. 1045–1057, 2013.
- [83] 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, J. Sicks, *et al.*, “Data from the national lung screening trial (NLST) [data set],” 2013.
- [84] 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, J. Sicks, *et al.*, “Reduced lung-cancer mortality with low-dose computed tomographic screening,” *New England Journal of Medicine*, vol. 365, pp. 395–409, 2011.
- [85] R. Kikinis, S. D. Pieper, and K. G. Vosburgh, *3D Slicer: a platform for subject-specific image analysis, visualization, and clinical support*, pp. 277–289. New York: Springer, 2013.
- [86] T. Kapur, S. Pieper, A. Fedorov, J.-C. Fillion-Robin, M. Halle, L. O’Donnell, A. Lasso, T. Ungi, C. Pinter, J. Finet, *et al.*, “Increasing the impact of medical image computing using community-based open-access hackathons: The NA-MIC and 3D slicer experience,” *Medical Image Analysis*, vol. 33, pp. 176–180, 2016.
- [87] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, *et al.*, “3D slicer as an image computing platform for the quantitative imaging network,” *Magnetic Resonance Imaging*, vol. 30, pp. 1323–1341, 2012.
- [88] 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 *3rd IEEE International Symposium on Biomedical Imaging: From Nano to Macro - Proceedings*, pp. 698–701, May 2006.
- [89] “SlicerLungCTAnalyzer.” <https://github.com/rbumm/SlicerLungCTAnalyzer/>, 2020.
- [90] R. L. Siddon, “Fast calculation of the exact radiological path for a three-dimensional CT array,” *Medical Physics*, vol. 12, pp. 252–255, 1985.
- [91] 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, J. Sicks, *et al.*, “Reduced lung-cancer mortality with low-dose computed tomographic screening,” *New England Journal of Medicine*, vol. 365, pp. 395–409, 2011.

- [92] M. De Greef, J. Crezee, J. Van Eijk, R. Pool, and A. Bel, “Accelerated ray tracing for radiotherapy dose calculations on a GPU,” *Medical Physics*, vol. 36, pp. 4095–4102, 2009.
- [93] 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 *Proceedings of the European Conference on Computer Vision*, vol. 9912, pp. 628–644, 2016.
- [94] H. Fan, H. Su, and L. J. Guibas, “A point set generation network for 3D object reconstruction from a single image,” in *Proceedings of IEEE Conference on Computer Vision and pattern recognition*, pp. 605–613, 2017.
- [95] 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*, vol. 5, pp. 349–359, 1999.
- [96] S. L. Delp, F. C. Anderson, A. S. Arnold, P. Loan, A. Habib, C. T. John, E. Guendelman, and D. G. Thelen, “OpenSim: open-source software to create and analyze dynamic simulations of movement,” *IEEE transactions on biomedical engineering*, vol. 54, no. 11, pp. 1940–1950, 2007.
- [97] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang, *et al.*, “TotalSegmentator: robust segmentation of 104 anatomic structures in CT images,” *Radiology: Artificial Intelligence*, vol. 5, pp. 1–14, 2023.
- [98] B. Dwyer, J. Nelson, T. Hansen, and et al., “Roboflow.” Available from, 2024. Software.

VITA

Name-Surname:

GÖKÇE GÜVEN

Education:

- Ph.D. in Computer Science, Özyeğin University, Istanbul, Turkey (2015 – 2025)
- M.Sc. in Materials Science and Engineering, Sabancı University, Istanbul, Turkey (2011 – 2013)
- B.Sc. in Chemistry, Koç University, Istanbul, Turkey (2005 – 2010)

Work Experience:

- CTO, Osteoid A.Ş., (2020 – Present)
- Project Manager & Research Scientist, TUBITAK 1507 SME R&D, (2022 – 2024)
- Project Manager & Research Scientist, TUBITAK 1507 SME R&D, (2020 – 2022)
- Research Scientist, TUBITAK Project, Özyeğin University, (2017 – 2020)
- Scientific Researcher, Gelecekhane, (2016 – 2017)
- Project Engineer & Data Scientist, BE1 Consultancy, (2015 – 2016)
- Project & Research Engineer, Project C5, Free University of Berlin, (2014)
- Research Scientist, TUBITAK Project, Sabancı University, (2011 – 2014)

Awards and Scholarships:

- Full Ph.D. Scholarship, Özyeğin University
- Tuition Waiver Scholarship, Sabancı University
- Full Merit Scholarship & Stipend, Koç University

Scientific Publications:

- Güven, G., Ates, H. F., & Ugurdag, H. F. (2024). X2V: 3D Organ Volume Reconstruction From a Planar X-Ray Image With Neural Implicit Methods. *IEEE Access*.
- Güven, G., Topcu, A., Somer, M., Afyon, S., Yilmaz, A., & Bolukbasi, O. (2022). Vibrational spectra, force constants and quantum chemical calculations of μ -1, 3-azide bridged triphenylphosphine complexes of copper (I) and silver (I). *Vibrational Spectroscopy*, 119, 103354.
- Guven, G., Atilgan, A. R., & Atilgan, C. (2014). Protonation states of remote residues affect binding–release dynamics of the ligand but not the conformation of apo ferric binding protein. *The Journal of Physical Chemistry B*, 118(40), 11677-11687.
- Abdizadeh, H., Guven, G., Atilgan, A. R., & Atilgan, C. (2015). Perturbation response scanning specifies key regions in subtilisin serine protease for both function and stability. *Journal of enzyme inhibition and medicinal chemistry*, 30(6), 867-873.

Scientific Presentations:

- Atilgan, C., Güven, G., & Atilgan, A. R. (2013, September). Control of open-closed conformers of ferric binding proteins by local and environmental perturbations. *246th ACS National Meeting & Exposition, Indianapolis, IN, USA*.

- Narin, Z. G., & Güven, G. (2013, June). Chemistry synthesis and characterization of ReB_2 and Re_7B_3 as superhard materials. *14th European Conference on Solid State Chemistry, Bordeaux, France.*

